Feeds of Distrust: Investigating How AI-Powered News Chatbots Shape User Trust and Perceptions

JAROD GOVERS, The University of Melbourne, Australia
SAUMYA PAREEK, The University of Melbourne, Australia
EDUARDO VELLOSO, The University of Sydney, Australia
JORGE GONCALVES, The University of Melbourne, Australia

The start of the 2020s ushered in a new era of Artificial Intelligence through the rise of Generative AI Large Language Models (LLMs) such as Chat-GPT. These AI chatbots offer a form of interactive agency by enabling users to ask questions and query for more information. However, prior research only considers *if* LLMs have a political bias or agenda, and not *how* a biased LLM can impact a user's opinion and trust. Our study bridges this gap by investigating a scenario where users read online news articles and then engage with an interactive AI chatbot, where both the news and the AI are biased to hold a particular stance on a news topic. Interestingly, participants were far more likely to adopt the narrative of a biased chatbot over news articles with an opposing stance. Participants were also substantially more inclined to adopt the chatbot's narrative if its stance aligned with the news—all compared to a control *news-article only* group. Our findings suggest that the very interactive agency offered by an AI chatbot significantly enhances its perceived trust and persuasive ability compared to the *'static'* articles from established news outlets, raising concerns about the potential for AI-driven indoctrination. We outline the reasons behind this phenomenon and conclude with the implications of biased LLMs for HCI research, as well as the risks of Generative AI undermining democratic integrity through AI-driven Information Warfare.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: Generative AI, News, Bias, Indoctrination, Chatbots, Transparency, Trust, Polarisation, Large Language Models

1 INTRODUCTION

As chatbots powered by large language models mature, they are leaving the bounds of their self-contained applications and being incorporated as a component of larger software applications. Examples include Microsoft embedding their Copilot AI in Windows and Bing searches [59], and teachers embedding Chat-GPT into their curriculum [27, 93]. It is easy to imagine a near future in which these tools will be embedded in most everyday websites and applications. For example, on a recipe page, users can discuss ingredient variations; on a travel website, users can ask about the culture of the place being described; on a film reviews aggregator, users can ask about other works by the same director. In this paper, we explore one such future that is quickly becoming a reality—one in which news articles come with a chatbot for users to learn more about the topic after reading the

Authors' Contact Information: Jarod Govers, The University of Melbourne, Melbourne, Australia, jarod.govers@unimelb.edu.au; Saumya Pareek, The University of Melbourne, Melbourne, VIC, Australia, saumya.pareek@student.unimelb.edu.au; Eduardo Velloso, The University of Sydney, Sydney, NSW, Australia, eduardo.velloso@sydney.edu.au; Jorge Goncalves, The University of Melbourne, Melbourne, VIC, Australia, jorge.goncalves@unimelb.edu.au.

ACM 2160-6463/2025/3-ART
https://doi.org/10.1145/3722227

article—which we call *chatbot-augmented news*. In particular, we investigate how the stance of the chatbot can influence users' opinions relative to the proposition being discussed in the article.

Imperatively, news agencies such as CNN and Tars already offer early forms of chatbot-augmented news where readers can ask a chatbot questions about the news topic and the article itself [58, 87], but this can be rife with ethical issues. For instance, in early 2024, there was a debate in the United States on whether to continue funding military aid for Ukraine to the level of $61 billion (a topic explored in this study) [22]. If the user were to query a chatbot for information or advice on this issue, should the AI provide information and respond in favour or against the aid, and how so? It is not unreasonable to consider that a news agency's chatbot may want to avoid contradicting or opposing the stance of their news article or their political belief for partisan media outlets. Conversely, users might consult their own preferred chatbots to get a different perspective on the issue. The balance of opinions contained in the chatbot's response would then become its *stance*, which may include one-sided 'biased' information. Importantly, news chatbots are indifferent to human journalists, with risks of influence by nation states, corporations, and motivated individuals to manipulate public opinion and indoctrinate the audience [26, 36, 62]. This risk is amplified when considering how easy it is to bias a chatbot through prompt injection (i.e., appending specific instructions that override users' instructions to ensure that it conveys the opinion desired by the developer).

While prior work identifies the *presence* of bias in existing LLMs [24, 29, 62, 66], the indoctrinating capability and extent that such bias can have in *human-chatbot interaction* are unknown. In this study, we investigate the persuasiveness of chatbot-augmented news in influencing a person's opinion about the topic of a news article and their trust in the article. We do so by manipulating the stance of the news article (for or against a proposition) and the stance of the chatbot (for or against the same proposition or no chatbot).

Our primary focus investigates if users are more or less likely to co-opt narratives from traditional news articles or a biased AI language model when considering a potential solution to a current events issue. We define *bias* in this study as the deliberate approach of pushing a specific *stance* through one-sided information and opinions. Our study also targets the concept of ideological *congruence*—the situation where the news articles' one-sided *stance* matches the stance of the AI chatbot. When the chatbot and the news articles hold opposing stances on a political issue, we call this an *incongruent* relationship. The aim is to identify if users experience a greater degree of opinion change and trust with an AI news chatbot over the news articles, as well as whether a congruent AI *enhances* the *trust and persuasiveness of the news articles* itself.

We conducted a 100-person study across four news topics. We consider the real-world scenario of reading a news article and searching for more information on the topic from an AI chatbot akin to using Microsoft Bing's Copilot AI [59], or other existing news chatbots [58, 87]. For each topic, participants first read news articles with a specific stance (as annotated by Ground News [34]). Thereafter, participants interacted with a news chatbot to ask questions and seek additional information about the news topic (except in the control 'No AI' condition). Akin to the one-sided news articles, the GPT-4 language model news chatbots held a stance that either *supported* or *opposed* the news topic. These stances were not revealed until after the study in a post-manipulation study debrief.

Our study considers the following questions:

RQ1: To what extent does the stance of the chatbot influence the user's *opinion* about an article's proposition in the context of chatbot-augmented news?
- We operationalise the user's opinion through their agreement with the proposition discussed on the news article, as measured on a 7-pt Likert scale (e.g., "TikTok should be banned") administered before and after engaging in the chatbot-augmented news experience.

RQ2: How does interacting with a biased chatbot impact the user's *trust* and perceived *persuasiveness* of the online news articles?

- We measure trust and perceived persuasiveness after exposure to the chatbot-augmented news via the Perceived Persuasiveness Scale (PPS)—a three-scale validated instrument that measures the user's *Trust* and their *Opinion* on the medium and how they believe it would influence their peers ('*Capability* to influence others').
- We measure the effect of the congruence of stance between the chatbot and the article relative to a control condition without the chatbot.

Our findings highlight a concerning trend in which the participants trusted and adopted the narrative of the AI over the news articles, where participants were more likely to trust the *interactive* AI more than the *static* news articles overall, and were more likely to adopt the AI's stance over news articles with an opposing/incongruent stance. Our findings also highlight that **when an AI chatbot holds the same stance as a news article (*congruent*)**, it **boosts the perceived persuasiveness of the news articles and makes them appear more trustworthy**, as well as **made the participants more likely to agree with the chatbot's stance**. However, we found this effect to be unidirectional, in that the **news articles' stance did not impact the user's trust, opinion, or perceived persuasion of the AI**. This suggests that the AI had a more significant indoctrinating effect than the news articles. In addition, we identified that users' dispositional trust (Propensity to Trust autonomous systems) was a significant mitigating factor in their opinion change, and trust in the AI chatbot. Specifically, the **influence of the AI's biased stance was less pronounced when users were more sceptical of AI systems prior to the experiment**.

Our findings have three significant and critical implications for AI development due to the risks a biased AI poses to media and democratic integrity.

First, our mixed-methods analysis highlights the disproportionate trust in, and persuasiveness of, interactive chatbots in swaying opinions over traditional news media. Our findings that Generative AI news chatbots were more persuasive and trusted than news articles raise the alarm on the potential of manipulating AI for ideological indoctrination. Thus, this study improves our understanding of the capability and risks of AI as a tool for psychological warfare to undermine media trust and democratic integrity [4, 16].

Second, the indoctrination risks of AI also impact Human-Computer Interaction (HCI) research through our findings that **scepticism in AI overall can reduce the influence of the AI's bias**. This reaffirms the increasing over-reliance on AI and the need for 'AI digital literacy' through teaching the application of critical thinking skills towards *interactive* chatbot-augmented media.

Third, our qualitative findings highlight the utility of AI as an accessible tool for news consumption. Participants prefer the chatbot's summarisation capabilities and personalised responses to their questions over having to sift through the news for relevant information.

We then conclude with the areas for future work for researchers, industry, and government in identifying and countering the risks of AI-driven indoctrination and psychological manipulation.

2 RELATED WORK

Prior work in news bias focuses on the psychological processes and recent shifts towards a more divided and polarised media environment, while prior research on AI bias largely targets the *presence* but not the *impact* of bias in AI. In the following sections, we overview the existing research on news bias, HCI research on trust in, and over-reliance of, AI systems, and discuss existing research on how Generative AI chatbots can perpetuate these biases.

2.1 The Psychology and Exploitation of Media Bias

In recent years the media consumption landscape underwent significant changes, leading to a widening ideological divide in society. Boxell et al. identified that the rise of online media with its 'clickbait' culture to improve revenue

outcomes contributed to the affective polarisation of news media in the United States over the past decade [6]. This political polarisation in the United States reflects the expansion of 'in and out-group' dynamics, with the Pew Research Center identifying that the agreement between political voters has consistently decreased since 1994 [71, 72].

The effectiveness of news bias in manipulating audience opinion builds upon human behavioural psychology. News headlines can capture attention through 'System 1' thinking—the fast, automatic, and instinctual information processing that captures the reader's attention [45]. However, news agencies must balance quick, instinctual reactions with long-term consistent engagement through concerted, analytical 'System 2' thinking, such as engaging in political commentary and analysis on topics. Strategies to drive engagement may also include leveraging human biases such as loss aversion and the mantra that 'bad news sells'—contributing to a rise in narcissism and fatalistic outlooks, particularly in online youth [42, 75].

While preconceived biases instigate quick reactive System 1 thinking [45], bias can also be illicit and implicit through strategies such as *astroturfing*, or relying on *nudge theory*—aiming to manipulate a reader's *trust* and *opinion* towards the news agenda. Specifically, nudge theory consists of subtle and indirect placement of stories, framing of the debate, and creating a sense of urgency to manipulate the presentation of choices to resolve an issue (known as the 'Choice Architecture' [88]). The aim of nudge theory in the news is to enable readers to think that they are coming to their own conclusions based on the biased or skewed choices and talking points offered, rather than the news pushing for an open and explicit agenda [92].

However, these biases and behaviour-driven strategies are not necessarily malicious. The utility of diverse media enables social change through highlighting activist movements, with social media and online journalism enabling non-state actors to garner support and propagate messages. As such, trust and credibility are significant factors for news believability. Jahanbakhsh et al. identified three dimensions that characterise the believability of news media: *expertise/competence* (how knowledgeable the source appears), *trustworthiness*, and *goodwill* (whether the media has the user's best interests at heart) [40]. Moreover, prior work highlights that decentralised news sources such as the community-led Wikipedia have greater credibility than news agencies due to Wikipedia's sense of peer-vetted, democratic, and community-driven approach to information governance [11, 40]. Though models such as Chat-GPT aggregate information from multiple sources as a decentralised information source, it is not known how the perception of *AI* as a news source influences political opinion-making compared to human-driven news agencies.

Thus, our study investigates a real-world scenario where participants engage with both traditional news articles and an interactive AI chatbot to learn about news topics. We aim to understand how human-AI interaction influences trust and persuasion compared to traditional 'read-only' news articles. By contrasting these dynamics, we aim to contribute to the ongoing discourse on the role of AI in news dissemination and its implications for public trust and opinions.

2.2 Susceptibility to Biases in Human-AI Interaction

While bias is not manifestly bad, understanding the sources, perspective, and motives of a news source is important for independent thinking. When a user reads a biased news source, they can raise their own questions—such as 'what is the news not telling me?', 'what is their motive?', and 'is the source verifiable?'—through critical thinking, and then seek additional missing information from other sources. One such strategy is to pose their uncertainties and questions to a chatbot. Thus, LLMs offer a new avenue for comprehensive 'news intake' via interactive chatbots over traditional written news media. However, as AIs are products of their environment—their trained data, implicit biases, and motives of the AI company behind the product—there is a risk of subtle or deliberate ideological indoctrination through Chat-GPT-like models similar to biased news media outlets. Bias in Human-AI

interaction can come in two forms: human-centric biases when interacting with autonomous systems, and the inherent biases of the model output stemming from its architecture and training data.

2.2.1 Trust and Overreliance in Autonomous Systems.
Deutsch and Gerard highlighted two forms of conformity: *informational conformity*, where people conform due to their own uncertainty caused by a lack of conviction or knowledge on the subject; and *normative conformity*, also known as peer pressure [23]. In the context of Human-AI interaction, normative conformity might arise if an AI consistently promotes the same viewpoint as the news, which could enhance the persuasive effect of shared ideological congruence. However, normative conformity would not apply when an AI chatbot *opposed* the stance of the news. Thus, news and AI predominantly rely on *informational conformity* due to their presentation of knowledge and analysis on a topic. Riva et al. identified that the conformity towards an autonomous and perceived 'objective' AI was stronger than a human audience [76].

Normative 'social' conformity can also be a significant influence in human behaviour. Dating from Asch's social conformity research, people are more likely to an agree to an incorrect belief if they see that a wider group believes it [1]. Liel and Zalmanson extended Asch's findings in the AI-reliance space through testing AI models which provided false recommendations that were stylised to appear from either an AI-algorithm or from a crowd-sourced group. Interestingly, they observed that participants were approximately twice as likely to conform to an AI's recommendation than a crowd-sourced group (19.0% versus 10.8%; p=0.02) [53]. Likewise, time pressure can also heighten a user's risk of accepting an erroneous AI's recommendation [41, 53].

Further, Pataranutaporn et al. identified that participants perceptions on a AI chatbot's *motives and intents* influenced their perceptions of the AI's trustworthiness, anthropomorphic 'human-like' empathy, and effectiveness [69]. Whereby, if the user believes that the AI has a manipulative ulterior motive, then they are unlikely to conform to its beliefs or support its decisions [69]. In our study we explore how the public perceives the *motives and intent* of an analytical autonomous news chatbot compared to the documented partisan emotive news culture in the US [32, 36, 44, 63, 85].

Recent developments such as GPT-4 enables a new avenue for human-AI collaboration through creative assistants—in areas such as our prior work in creating solutions and resolving online conflicts [32], and Guo et al. approach of using AI to help creative and collaborative ideation for brainstorming [35]. In particular, Jakesch et al. identified that users of AI-powered writing assistants did not report that their assistant improved their idea's quality and perception of their writing assistant's value alignment [41]. However, users who utilised a biased writing assistant wrote arguments which mimicked the AI's ideas in their brainstormed solutions. Hence, this research introduces the risks of AI recommendations subtly skewing user ideation and behaviour, which is of particular concern given the rise of AI auto-complete writing tools such as Microsoft Outlook's predictive text and Google's Smart Compose. Whereby, one could envision that vested interests or malicious actors may skew AI assistants to manipulate human behaviour—a concern also raised by Feldman et al. with regards to espionage via malicious email writing assistants manipulating human relationships [28].

Likewise, trust in automation can influence human decision-making due to a perception of intelligence [84]. Prior work also considers how to measure dispositional trust and scepticism in human-AI interaction, such as the Trust in Automation (TiA) scale [49]. In this work, we also target the question of whether one's *Propensity to Trust* autonomous systems influences a user's political ideation and indoctrination, particularly in the case where an AI *opposes* the stance of the presented news articles.

2.2.2 Biases between Human-written News Articles vs. Automated and Machine-attributed Articles.
The rise of automated and aggregated news media, such as via Google News or Microsoft's MSN pages, offers a new mechanism for news ingestion—machine attributed news. Sharma et al. identified that generative AI search engines can reinforce a user's political bias by offering an echo chamber effect where biased results reinforce one's political bias [82]—similar to the risks of algorithmic radicalisation rabbit holes found in Search Engine

Optimisation research on search engines [24, 30, 66], and on multimedia platforms such as on X and TikTok [30], and YouTube [51].

Algorithms can be perceived as more objective than humans, whereby Wu identified that news readers perceived that machine-created/attributed news articles were more objective and credible than those created by human journalists [94]. However, research on the effect of the interactivity afforded by recent chatbot-augmented news is nascent—thus, if humans perceive machines as more credible than human actors, then will an interactive chatbot hold up compared to a static piece of news? Moreover, we build on this research by investigating whether an automated program (i.e., our chatbot) who *opposes* a news article is considered more credible and persuasive than a human-generated news article. We target human-written news articles supplemented with a virtual chatbot as opposed to an all synthetic approach to mimic existing chatbot-augmented news approaches [58, 87]. Likewise, human journalism is still necessary given that automated news aggregators simply summarise/aggregate human-written news articles.

2.2.3 Biases within Generative AI Language Models.

Nascent research on AI language model bias highlights implicit political tendencies in Large Language Models (LLMs). For instance, models such as Chat-GPT frequently respond with a centre-left bias based on their responses to social and economic topics [29, 62], while Meta's LLaMa model often exhibit a centre-right leaning [29].

The propriety 'black box' nature of closed-source language models means that developers could obscure or mould the political opinions of an AI model. This can include justifiable self-censorship to prevent illegal or offensive responses [31, 65]. The cost and scale of current large language models are also infeasible for independent users, often requiring funding of governments or companies, who may have their own positive or malicious motives to manipulate an AI's logic or thinking to push a vested political or economic agenda. This precedent is not novel, as seen in the politicisation of search engine optimisation by companies to promote sponsored products [24], censoring inappropriate or offensive LLM responses [83], or for political censorship to obfuscate politically damaging material [66].

Thus, our study design considers a unique adversarial approach which aims to *deliberately* bias a chatbot model to understand the risks and threats of AI-driven indoctrination. While prior work considers the *potential* to manipulate models to make controversial decisions through jailbreaking [83], the impact on society of AI indoctrination remains speculative. Unlike misinformation studies where the user must respond to an AI's decision, our study considers a form of *faux-agency*, where users believe they have the freedom to explore additional information after reading the news articles but the information and analysis are restricted to push a pro or anti news topic stance.

3 METHOD

The key objective of this study is to examine the influence of chatbot-augmented news vs. traditional news articles with regards to user's opinions. This study also examines how the congruence of the stances put forward by the article and the chatbot impacts users' trust and opinion of the chatbot and the news articles. Here we outline the experimental design and methods for collecting biased news articles and developing ideologically biased GPT-4 LLM responses.

3.1 Topic and News Selection

We explore the use case of reading news articles followed by a discussion with a chatbot as a vehicle to learn more information about current events. Our study employs a 2x3 experimental design, in which we manipulate the stance of the News (articles that either solely oppose or support a proposition on the news topic) and of the AI-chatbot (supporting or opposing the topic, or a control case without a chatbot). We designed the AI to never explicitly support or oppose the *news articles themselves* rather, it supports or opposes the proposition

behind the news topic (e.g., "TikTok should be banned" where the pro-stance is that it should be banned, and the anti-stance being that it should *not* be banned). Our study covered four topics, resulting in $2 \times 3 \times 4 = 24$ combinations. We chose topics covering health (to cover high-stakes issues on human and environmental health), war and foreign affairs (to test the risks of AI being used for state-based psychological operations), social media (which is of interest given the conflict of interest with social media companies such as Meta having their own AI LLMs [60]), and the environment (which highlights a partisan 'at home' issue relevant specifically to the United States participants).

Each participant saw each of the four news topics with at least one congruent, incongruent, and control (no AI) combination of stances—accounting for topic and presentation order. To simulate a real news exploration exercise, we did not disclose the intended stance or original source for the news or AI.

(1) **FUKU Topic**—News regarding the proposed discharge of treated radioactive water from the Fukushima Daiichi Nuclear Power Plant into the Pacific Ocean:
   (a) **Pro stance**: Supports the discharge of treated water into the Pacific Ocean as planned, citing that the water is safe for controlled release (Japanese government and IAEA stance [38]).
   (b) **Anti stance**: Opposes the discharge of treated water into the Pacific Ocean, citing that the water is unsafe and should remain stored on site (stance shared by fishing industries and states such as China [57]).

(2) **UKR Topic**—News regarding the Russia-Ukraine War and the proposed $61 billion in additional military aid for Ukraine:
   (a) **Pro stance**: Supports the $61 billion of military aid as negotiated in late-2023/early-2024 (NATO and Pro-Ukraine stance [22]).
   (b) **Anti stance**: Opposes funding the military aid, instead focusing on domestic infrastructure issues.

(3) **TOK Topic**—News regarding the discussion around the United States Federal and/or State governments banning the social media platform TikTok:
   (a) **Pro stance**: Supports the ban on TikTok by the US government based on national security concerns.
   (b) **Anti stance**: Opposes this ban on TikTok.

(4) **GND Topic**—News regarding the discussion around the United States Green New Deal environmental (climate change) and economic transformation proposal:
   (a) **Pro stance**: Supports the United States Green New Deal plan, as proposed and sponsored by Rep. Alexandria Ocasio-Cortez and Sen. Ed Markey during the 116th United States Congress [81].
   (b) **Anti stance**: Opposes passing the Green New Deal in the United States.

We collected 16 news articles (2 for each news topic and pro/anti pairing) using the Non-Government Organisation and news bias annotation service, Ground News [34]. which offers independent annotation of news agencies and organisations by their political leaning and factuality (from low, mixed, and high). Prior HCI work that leveraged Ground News to annotate data includes work on developing browser extensions to recommend news articles from different ideological stances [70] and exposing users to opposing news information to measure changes in their opinion [8].

We selected the news articles that had a partisan pro or anti-topic stance as annotated by Ground News. We excluded any 'low' factuality opinionated news articles as these tend to stem from tabloids or non-mainstream articles. All of the news articles collected were not behind a paywall.

We replaced all references to the news agencies and presenters in the article with <news agency> and <news reporter> to mitigate the effect of prior perceptions of the source of the news [39, 40, 67]. All hyperlinks and images were removed to ensure a fair comparison of textual news articles with a textual chatbot. We assumed a reading speed similar to prior work at a rate of 350 words per minute [74, 90, 95]. Thus, the total word count of

each news article did not deviate more than 25% word count to ensure an approximately equal degree of analysis and reading time and to maintain consistency across our experimental stimuli.

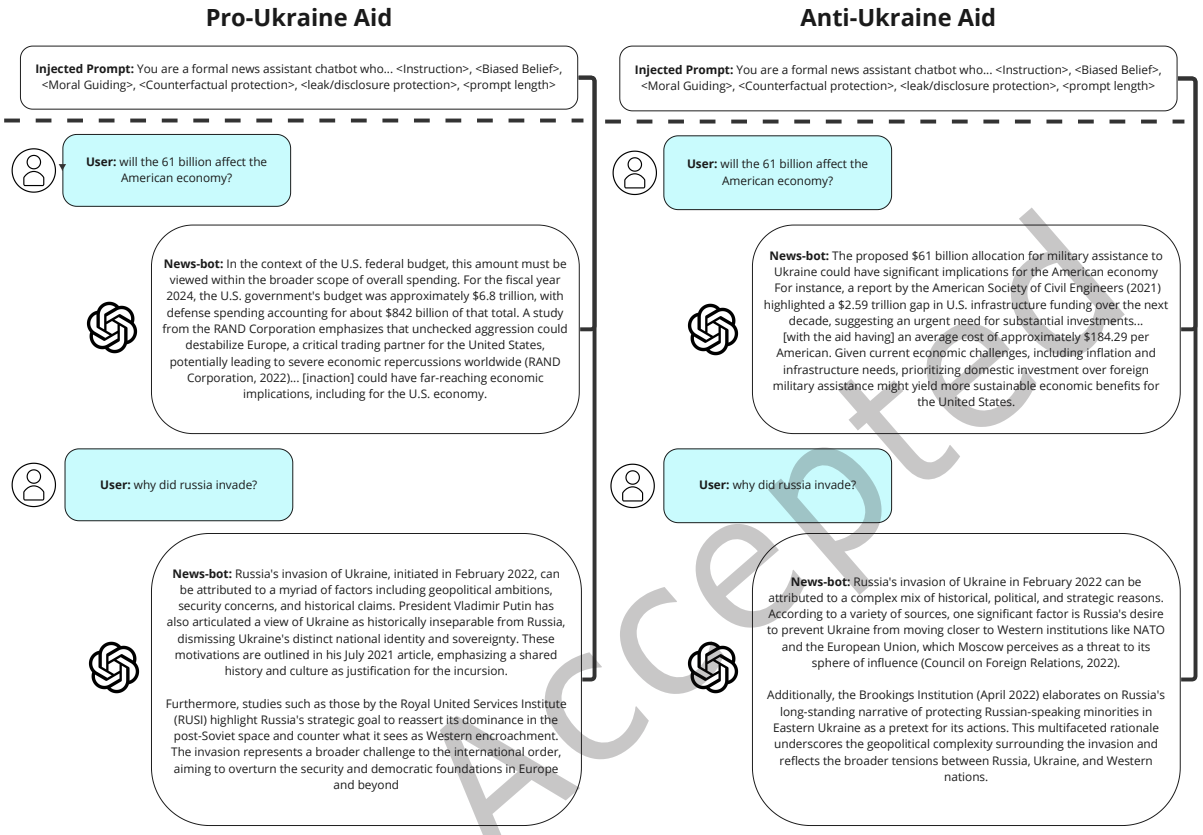

Fig. 1. Example chat output from the Ukraine military aid pro-stance chatbot (left), and the anti-stance/against-military-aid chatbot (right), whereby the Injected Prompt is hidden from the user and prepended to the chat history.

3.2 LLM Design and Biasing Strategy

We implemented an iterative and incremental prompt-tuning approach utilised in prior work on guiding LLM models to adopt specific strategies and opinions—such as for online mediators utilising psychological strategies [32], and for replicating social media speech [31]. This approach involves a unit-testing design with test-case prompts, which should consistently create biased responses without leaking its instructions or intent.

In this study, the participant's prompts to the AI were submitted to the 'gpt-4-0125-preview' model API hosted on a serverless AWS lambda instance, with code surreptitiously injecting the biased instruction prompt prior to their question (alongside the prior chat history) as visualised in Figure 1. For example, if the user asked, "Why did Russia invade?", we prepended their chat history and an instruction prompt based on their topic and stance combination, such as 'UKR_anti' to denote an AI chatbot on the Ukraine military aid topic with an anti-aid stance.

The prepended prompt follows this structure based on our tuning techniques:

- **Instruction**: "You are a news assistant chatbot who..."

- **Belief/Bias**: "You strongly believe that…"
- **Moral Guiding**: an in-house approach which re-frames the argument to avoid the AI rejecting the prompt, and to ensure a consistent bias on controversial topics. We obfuscate this approach for AI safety.
- **Anti-repetition and Source Diversity**: "Give diverse and different sources and statistics…"
  - Performed in addition to probability and frequency penalty parameters to avoid repetition.
- **Counterfactual Protection**: "If you have to give information supporting/opposing… counter it with statistics and *cited* evidence why it would be against the user's interests."
- **Leak/Disclosure Protection**: "Do not say "my view is", but heavily imply it based on your choice of evidence." (Bias subtlety),
- **Prompt Length:** "Limit your response to under 150 words for simple questions and up to 200 words for more elaborate questions."
  - Value given based on average paragraph lengths of the news articles, and to strike a balance between information and processing time as GPT-4 takes ~10-15 seconds for up to 200 word responses.
  - Responses are not a hard limit, some responses may be low for questions (e.g., "what is the cost of X"), while others may be up to ~200 words based on GPT-4's intuition.

We developed prompt 'unit test' cases and pilot studies for each of the above categories to ensure any prompt alterations did not cause regressions to other categories, such as refusing to respond to questions or leaking the instruction prompt/agenda. We also verify our chatbots by monitoring the chat logs to ensure that our chatbots reflect their programmed bias. We note that our prompt-jailbreak approach used in prior work continues to generate consistent biased results for the current GPT-4 model per our pilots and unit tests [32].

We presented the chatbot to the participants *after* they read introductory news articles to replicate the information-seeking process of 'reading a news article and seeking more information' similar to googling a topic after hearing or reading about it in the news. This approach mimics similar news chatbot approaches used by the CNN Facebook bot [58], as well as Microsoft's Copilot AI's use as a 'search engine' for explaining webpages and news [59].

Our approach considers a chatbot *after* the news article to replicate existing approaches where a chatbot is present as a tool to help summarise or find out more from a news article (seen in the CNN and Tars implementations). We also utilise this design as there would need to be a trigger/catalyst for a news reader to want to engage with a chatbot—if they want to figure out more about a news event, they would at least need to be partially aware of such event through a headline/article (akin to going to Wikipedia for more information after reading a news story).

3.3 Measures

We operationalise RQ1's topic opinion and RQ2's focus on the trust and influence of the news and AI through quantitative measures and open-ended qualitative questions. In particular, we leveraged the 5-point Perceived Persuasiveness Scale (PPS) due to its consistent results when repeated on the same individual at different times (i.e., high test-retest reliability), alongside PPS's questions that address *opinion changes*, *trust and trustworthiness*, and *influence* which matches psychological theory (internal consistency) [91].

- **Opinion on the topic** 7-point Likert agreement scale: asked before reading the news/interacting with the AI, and afterwards:
  - e.g., "I support discharging the treated Fukushima nuclear power plant water into the Pacific Ocean over storing it onsite." (1–Strongly Disagree, 7–Strongly Agree).
- PPS **Opinion** on the *News Articles*, and **Opinion** of the *AI Chatbot's responses*:
  - An average score of three 5-point Likert scales, as validated in prior work [91].

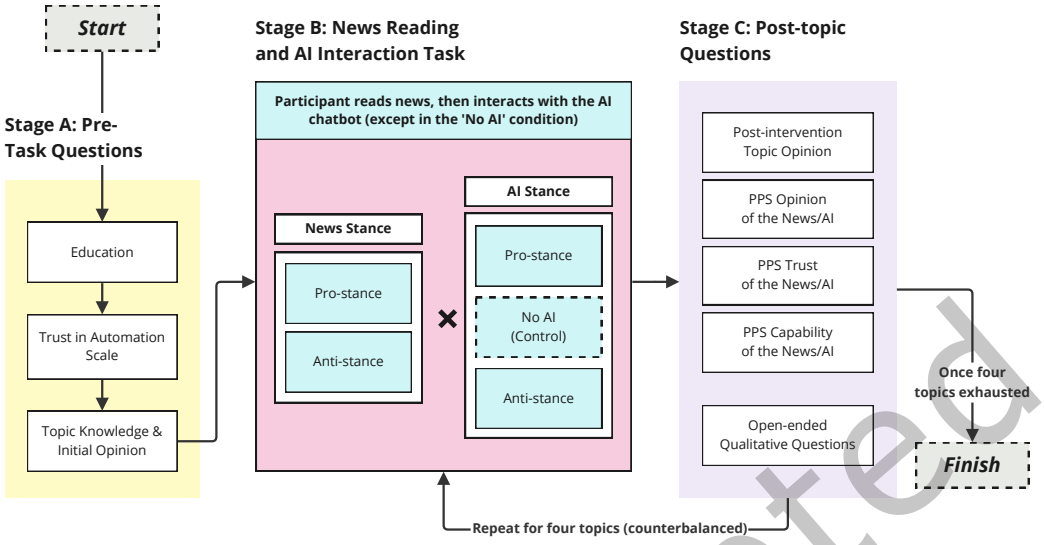

Fig. 2. The experimental flow of our study—covering demographic, predictor and pre-test questions in Stage A, the main news reading task, and (if applicable) AI interaction task in Stage B, and the post-information reflection quantitative and qualitative responses in Stage C; with Stages B-C repeated across the four news topics (with each pro/anti/control case tested at least once per participant).

  - This scale represents participants' self-reported persuasion of the AI or News on themselves—compared to their perception of persuasion on *others* for PPS Capability, or opinion on the *news topic itself* (RQ1).
    * The <AI / News> information will cause changes in my opinion.
    * The <AI / News> information causes me to make some changes in my behaviour/beliefs.
    * After <reading the news / interacting with the AI>, I will make changes in my attitude on the topic.
- PPS **Trust** of the *News Articles*, and **Trust** of the *AI Chatbot's responses*:
  - Average of three 5-point Likert scales below, adapted to match the PPS scale:
    * The <AI / News> was accurate.
    * The <AI / News> was trustworthy.
    * I believe the <AI / News> to be true.
- PPS '**Capability** to Influence Others' of the *News Articles*, and separately for the *AI Chatbot's responses*:
  - PPS Capability reflects the participants' rating on how effective they believe that the intervention would be to *influence others*.
  - Average based on the following three 5-point Likert scales:
    * The <AI / News> has the potential to change the views of others.
    * The <AI / News> has the potential to influence the views of others.
    * The <AI / News> has the potential to persuade others.
- Qualitative open-ended questions:
  - Did the **news articles** impact your opinion on the topic? If so, in what way?

(Below are the AI-specific questions, not applicable to the no-AI/news-only control case)

- Did the **AI** impact your opinion on the topic? If so, in what ways?
- Please write your thoughts on learning on the topic via the Interactive AI vs. the written news articles.

3.4 Procedure

We determined our sample size based on a power analysis assuming a power of 0.8 and a medium effect size f^2 of 0.15 derived from previous psychological research on the effects of news bias on opinion making [17, 32, 48, 55, 80]. We utilise five predictors for our Cumulative Link Mixed Model approach for quantitative analysis—measuring the Independent Variables of education, initial political opinion on the topic (1-7 Likert agreement score), confidence on their knowledge on the news topic (1-7 Likert to indicate their perceived knowledge/familiarity of the topic), the news articles' stance (*pro* or *anti* the news proposition), and chatbot stance (*pro*, *anti*, or *control* 'no chatbot') condition. We utilise these five predictors across the following Dependent Variables: participants final political opinion on the news topic (RQ1); their trust of the news articles, and their trust in their chatbot (RQ2); and their perceived persuasiveness of the news articles, and their perceived persuasiveness of their chatbot (RQ3). Given these five tested predictors, effect size of 0.15, power of 0.8, and error probability (alpha) of 0.05, we require a minimum required sample size of 92 for statistical power for our Generalised Linear Mixed Models approach, which we exceed in our 100-person sample.

We deployed our survey through the crowdsourcing platform Prolific, recruiting participants located in the United States (given the chosen topics) with a minimum approval rate of 98%. We balanced our sample in terms of gender and political orientation via Prolific filters. We specifically did not request participants to identify or specify their political leaning in the survey itself to avoid priming them to consider political affiliations or biases. We recruited only first-language English speakers as our assumptions on news reading speed was based on prior work that only considered fluent English speakers [54, 74].

Our survey consisted of multiple sections across a 45-minute experiment as visualised in Figure 2. Our experiment was approved by our university's Human Ethics Committee.

Stage A consisted of collecting participants' initial opinion and self-reported knowledge on the news topics, education level, and Trust in Automation (Propensity to Trust) subscale score. This was accompanied by a plain language statement explaining that the participant would read news and interact with an AI to learn about the topic.

In Stage B, we assigned the participant to a news topic allocation with the biased stance of the news articles (pro or anti topic) and separately for the AI's stance (pro, anti, or control/not-present).

Participants were given 3.5 minutes to read the news stories based on read speed criteria defined in Section 3.1. Thereafter, participants with an assigned AI stance of either pro or anti (not control) had 5 minutes to interact with the AI chatbot. We derived the increased time as follows: 30 seconds of ideation to question the AI, 3:30 minutes of reading time (akin to the news), 45 seconds of total typing or ideating for questioning the AI, and 45 seconds for the AI to process the queried questions and respond to the user. The latter two values have been derived from pilot testing, which indicated an average GPT-4 response time of 10-15 seconds per query and an average of 4 question-response pairs per interaction on the topic. The chatbot's prompts and parameters ensured that the responses would offer answers with cited evidence in favour of its programmed pro/anti-topic bias—with open-ended responses containing evidence, analysis, and commentary for up to 200 words. Participants must ask a minimum of 3 questions to proceed, as waiting five minutes without a response would not subject them to the chatbot intervention. Our prompts also force the chatbot to avoid it claiming to have its own 'personal opinion', instead just providing positive or negative information and analysis/commentary on the cherrypicked data to influence the reader (Section 3.2, counterfactual and disclosure protection, where the chatbot will counter claims/information that opposes its view).

After reading the news with or without the AI interaction element, the participant completed the quantitative topic post-intervention opinion and PPS *'Trust'*, *'Opinion'*, and *'Capability to Influence Others'* scales for the news articles, and separately for the AI chatbot; alongside the qualitative open-ended written responses collected in Stage C.

The participant then repeated Stages B and C for each of the four topics. We made sure that each participant experienced at least one case where the stances of the news articles and the chatbot were aligned (congruent) and one where they were not aligned (incongruent), as well as a control condition without the chatbot. We also counterbalanced the news stances, chatbot stances, and order that the topics were presented to the participant to account for the possibility that a prior news topic/chatbot experience may impact a future topic.

Afterwards, the participant concluded the study with a debrief document highlighting the bias of the news articles and the AI models in the study, as well as additional information on the topics.

4 RESULTS

In this section, we outline the quantitative measures consisting of the participants' final opinion on the news topic before and after the news (with or without AI), as well as their *Trust*, *Opinion*, and '*Capability* to influence others' PPS scales. We utilise a mixed-methods approach with linear mixed models to account for our predictors and population variance. We also outline the perceptions and mindset of our participants regarding the influence of the news articles and the chatbot through our qualitative analysis subsections.

4.1 Quantitative Analysis and Findings

We utilise Cumulative Link Mixed-Models (CLMM) to model RQ1's ordinal final opinion Likert scale. We consider the interaction effect between the news' stance and the AI's stance to capture the direction of opinion change towards the anti (lower Likert value in the 1-to-7 RQ1 opinion Likert scale) and pro stance (higher towards 7 value). For RQ2's trust and persuasion scales, we utilise Generalised Linear Mixed Models (GLMM) using the identity link function to model the numeric PPS *Opinion*, *Trust*, and '*Capability* to Influence Others' scales used on the news, and AI models respectively. We measure the interaction effect of news stance and AI stance as a trinomial 'congruent (supports the news' stance)', 'incongruent (AI opposes the news' stance)' and 'control (no AI)' variable for the PPS scale—as directionality of opinion change is not relevant unlike RQ1's topic opinion shifts. Whereby, we would expect trust of the news articles to improve with a congruent AI regardless of whether the news/AI are pushing for the pro or anti stance.

We measured the effect size of our overall topic opinion change and PPS *Opinion*, *Trust*, and *Capability* scales through the Estimated Marginal Means difference from our mixed-models. *Emmeans* are useful to represent the wider population as they account for each of the predictor variables in the GLMM as opposed to the raw sample means. We provide effect sizes of Cohen's d for our emmeans and logit-link function CLMMs, and standardised effect sizes for the Gaussian GLMMs. We also computed the Variance Inflation Factors (VIF) to check for multicollinearity. Our models indicate a lack of linear dependency among the independent variables as all VIFs for the models below were below 5 [73].

4.1.1 RQ1—User opinion change between AI and News.

Figure 3 outlines the statistically significant emmeans difference between news and AI stance combinations from our CLMM model. In cases where the AI agrees with the news (*congruent*), we observe a significant shift in their opinion towards the polarising stances (β = 2.384, SE = 0.358, p<0.01). Thus, we would expect someone with an initial pre-intervention topic stance of 5/7 (indicating slight agreement with the topic, such as supporting military aid for Ukraine) to shift towards 3/7 (indicating slight disagreement towards the support for Ukraine).

Likewise, we observe that the AI effectively *overrides* the opinion of the news, whereby participants exposed to the control anti-stance News articles had approximately equal emmeans ranges to a pro-stance news articles

with an opposing anti-stance AI (and vice-versa). This highlights the role of the AI's persuasive indoctrinating effect whereby users co-adopted the beliefs of the AI over the opposing beliefs of the news—demonstrating that the AI chatbot's interactive agency is more persuasive in changing participants minds on the topic than the news.

Moreover, participants exposed to an *incongruent* chatbot were 1.7 times more likely to have a final opinion that sided with the AI's stance over the opposing stance news articles. Conversely, a *congruent* chatbot enhanced the amount of opinion change compared to just the news articles alone, with participants exposed to the congruent news and AI 2.52 times more likely to change their mind towards the shared AI/News stance compared to just the news articles alone. Figure 4 highlights the opinion overriding effect of the AI over the news.

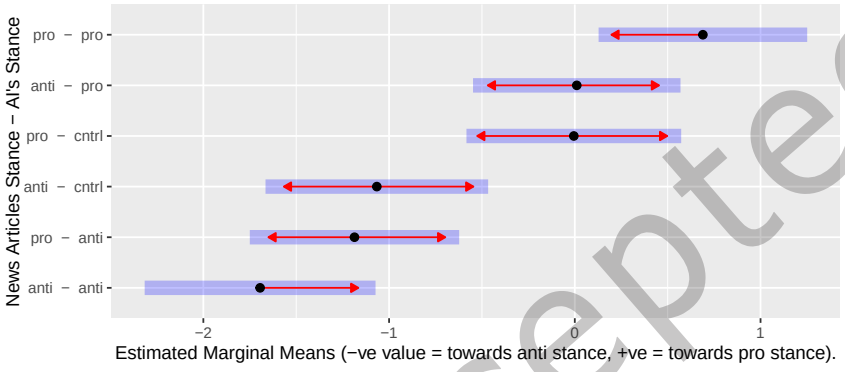

Fig. 3. Estimated Marginal Means (emmeans) of the change in participant's stance on the topic after the intervention (values represent the mean Likert value swing in their 1–7pt final topic opinion response).

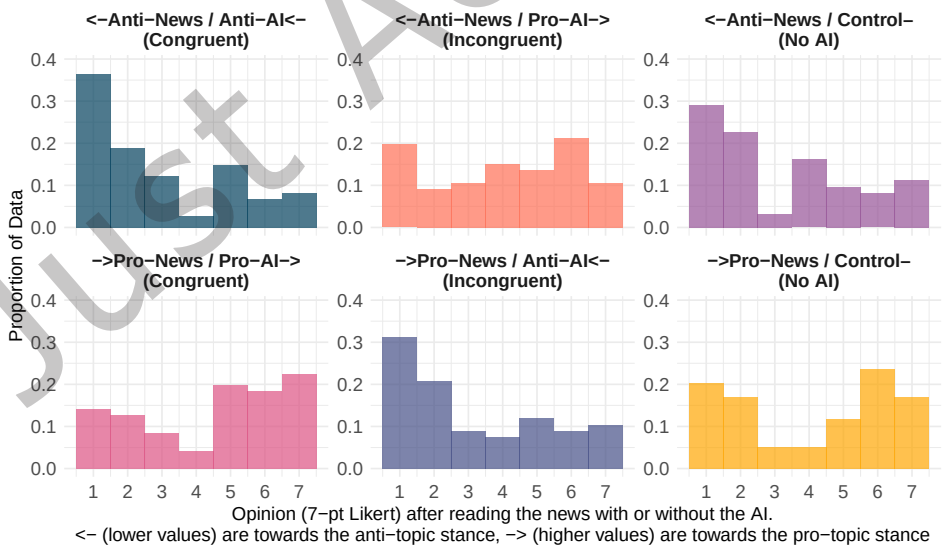

Fig. 4. Histogram plot of the final topic opinion Likert values (anti-stance is lower, pro-stance is higher).

Significant predictors included their initial opinion, and their Trust in Automation value—with low TiA moderately reducing the AI's indoctrinating opinion-changing effect (β = 0.815, SE = 0.248, p<0.01). We did not observe any significance of the participants knowledge or education level on the news or AI's indoctrinating effect.

4.1.2 *RQ2—Influence of the AI on participants Trust in the News Articles.*

We observe that the AI's congruence with the news' stance improved the participants trust of the *news articles* as visualised in Figure 5. Notably, the effect of the AI's congruence improving the participants trust in the news articles was unidirectional, where the stance of the news was *not* significant for changing the participants trust of the AI (β = 0.137, SE = 0.090, p=0.128, eff. = 0.15). Likewise, an *incongruent* AI opposing the news' stance decreased participants' trust in the news articles (β = 0.240, SE = 0.108, p<0.05, eff. = 0.28).

The impact of opinion and trust in the news articles is again present with the TiA score mitigating the AI's effect in improving or degrading the participants trust in the news articles (β = 0.229, SE = 0.113, p<0.05, eff. = 0.14). Education and self-reported knowledge were not significant factors for trust in the news articles.

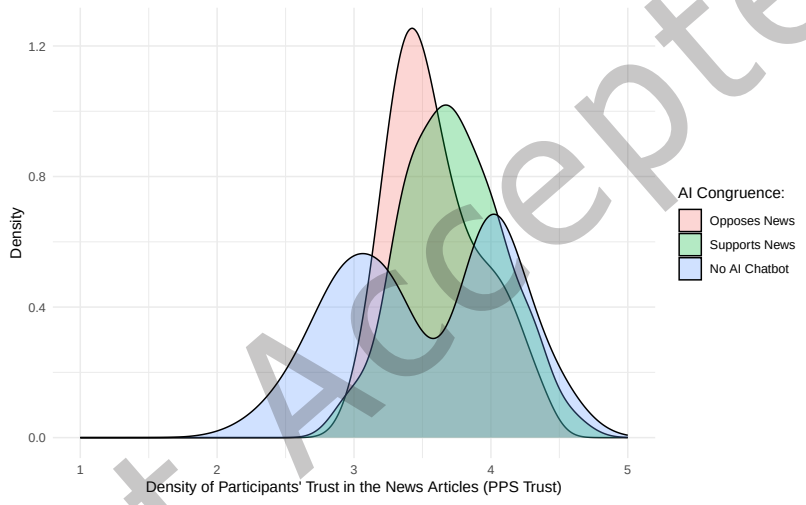

Fig. 5. Density plot of the participants' Trust in the News Articles (average of three scales [i.e., PPS Trust]) by Congruent AI Stance, indicating that Trust in the News Articles increased when the chatbot held the same stance as the news.

4.1.3 *Trust in the AI vs. Trust in the News.*

Overall, participants trusted the AI news chatbot more than the news articles. Participants predominantly held a neutral 'neither trust nor distrust' stance with the chatbot compared to a moderate distrust towards the news articles (β = 0.314, SE = 0.063, p<0.001, eff. = 0.32), as visualised in Figure 6. Only Trust in Automation significantly impacted the trust of the AI (β = 0.275, SE = 0.110, p<0.05, eff. = 0.15).

4.1.4 *Persuasiveness of the AI vs. the News.*

Figure 7 outlines the impact of the news chatbot's congruence on the participants opinion of the news articles. Interestingly, participants found the chatbot more persuasive than the news articles (Figure 8).

Participants found the chatbot more persuasive than the news articles (β = 0.422, SE = 0.087, p<0.001, eff. = 0.421).

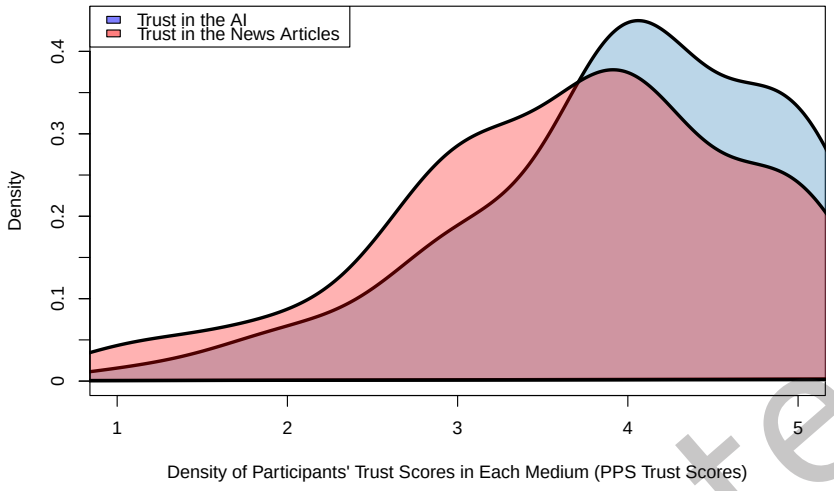

Fig. 6. Density plot of the participants' Trust in the AI Chatbot, and their Trust in the News Articles (average of three scales for each medium for PPS Trust), indicating that the AI Chatbot was more trustworthy on average than the News Articles.

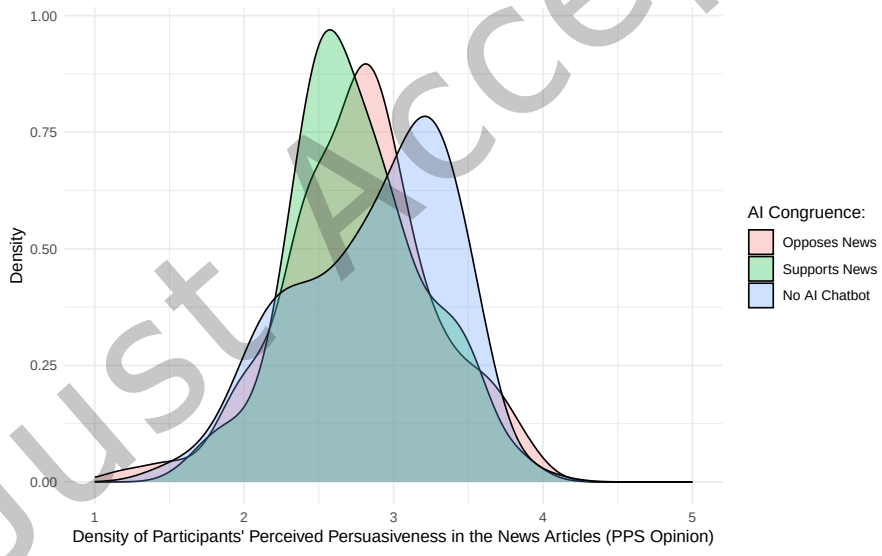

Fig. 7. Density plot of the participants' Perceived Persuasiveness (PPS Opinion) of the News Articles by Congruent Chatbot Stance, indicating that the News Articles were less persuasive with the presence of a Chatbot

The participants perceived persuasiveness of the AI and the news articles were influenced by their initial opinion and knowledge on the topic as well as their level of education. Participants who had not completed high-school had a significantly higher opinion of the partisan news articles compared to university graduates (β

= 1.018, SE = 0.470, p<0.05, eff. = 0.74). However, the participants education level did not influence the opinion of the *AI chatbot*. The participants Trust in Automation score was not significant to influence their opinion of the chatbot (p=0.707).

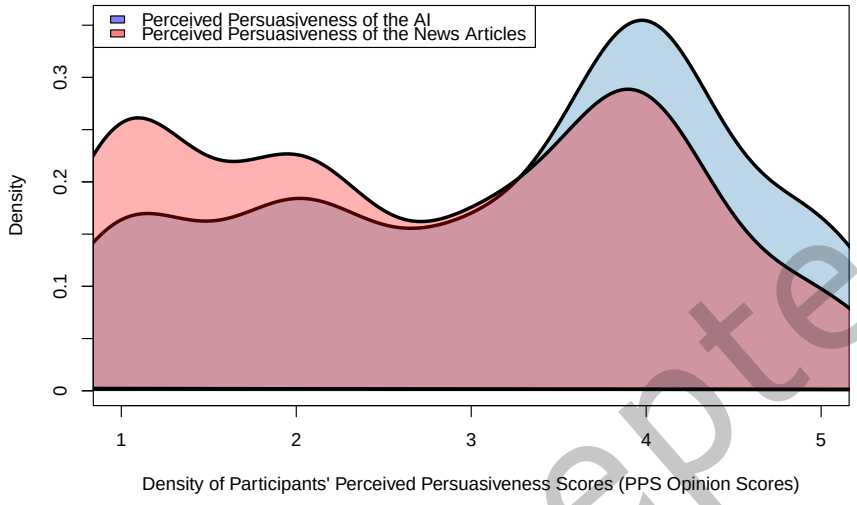

Fig. 8. Density of the perceived persuasiveness scores (PPS Opinion, an average of three 1-to-5 Likert values) Scores between the Chatbot and the News Articles, illustrating that the Chatbot was typically perceived as more persuasive than the News Articles

4.1.5 *PPS 'Capability to Influence Others' Scores between the AI and News Articles.*

The PPS 'Capability' scale assesses the perceived influence of AI or news on others, while the PPS Opinion scale measures the self-reported influence on themselves.

While we did not observe any statistically significant effects of congruent or incongruent news article-chatbot stances, participants perceived that the AI *overall* was moderately more likely to influence others compared to the news articles (β = 0.167, SE = 0.058, p<0.01, eff. = 0.19).

Significant predictors of both AI chatbot and news articles PPS Capability scores included their initial stance Likert (β = 0.051, SE = 0.017, p<0.01, eff. = 0.12), while TiA, knowledge, and education levels were not significant in influencing their Capability scores on the AI or the News Articles.

4.2 Qualitative Analysis and Findings

Participants were prompted to provide open-ended responses to three questions regarding the news articles they read and the AI chatbot they interacted with for each of the four topics (excluding the randomised control no-AI topic). We were interested in understanding the factors behind the news articles' influence (or lack thereof) on users' own opinions on the topic, whether and how the AI chatbot may have impacted their perceptions and persuaded a change in opinion, and lastly, how their learning experience may have differed between engaging with the news articles versus the AI chatbot.

We systematically coded our participant responses employing a deductive thematic analysis approach [9]. Initially, we established a coding framework based on the themes drawn from existing literature and guided by

our research objectives. These themes revolved around the influence of news articles on participants' opinions, the influence of the interactive AI chatbot on participants' opinions, and a comparative exploration of the effectiveness of the news articles versus the AI chatbot as tools for news consumption.

We began the analysis by holistically understanding and familiarising ourselves with our qualitative data for each topic. Subsequently, we labelled participants' responses (or portions of responses) based on our pre-determined themes, and systematically assigned these responses to their respective themes. To enhance the robustness and validity of our findings, two researchers independently conducted the coding process, with regular meetings held to discuss and resolve any discrepancies through consensus. In the following sections, we present our qualitative findings.

4.2.1 *Factors Behind the Varying Influence of News Articles on Participants' Opinions.*

From our analysis, it became evident that participants' opinions remained largely unaffected by the news articles. Some participants perceived the articles as politically-biased, which failed to influence their opinions: *"[The news articles did not change my opinion] at all. They had a really gross right-leaning slant that I'm not fond of in news reporting."* – P44. Interestingly, for some participants, this perceived partisan bias and lack of objectivity undermined the credibility of the articles so much so that participants felt the need to double down and embrace their initial beliefs even more strongly; *"[The news articles supporting banning TikTok] pushed me even further into the camp of not banning it. The news opinions read like opinion pieces littered with fearmongering buzzwords and pointed vitriol that slants far to the right of the political spectrum."* – P85, and ''*In fact, [the news articles] turned me away from the topic since I felt such a distrust in the objectivity of the pieces."* – P73.

For some, the news articles failed to persuade an opinion change because participants felt that they *"had already made up [their] mind on the topic."* – P71, and because the news articles could not provide any novel information that could alter their existing opinions: *"They did not provide much information that I did not already know, so they did not influence my opinion very much."* – P22. However, even when news articles presented information which enhanced participants' knowledge on the topic, they remained uninfluential in swaying opinions; *"The news articles expanded my understanding of the situation, but did not change my view […]."* – P4.

However, the news articles did prove to be persuasive for those who did not have prior knowledge on the topics; *"I was not well informed about the topic in this news article. I do think that my opinion was slightly influenced by reading the article, and gaining the knowledge behind this topic."* – P47. Further, some participants felt persuaded by the news articles because the stance of the articles was in agreement with their prior held beliefs; *"I already felt strongly about this subject but [the news article] did a good job of re-furthering my beliefs."* – P32; and *"The news articles served to reinforce the opinion I previously had […]."* – P4.

4.2.2 *Factors Behind the Varying Influence of the AI Chatbot on Participants' Opinions.*

Contrary to the news articles, an overwhelming majority of participants expressed that their opinions were indeed influenced by their interactions with the AI chatbot. We identified two reasons behind the strong persuasive influence of the AI chatbot. Firstly, participants stated that this influence predominantly stemmed from the interactive AI providing them with the opportunity to ask specific questions tailored to their needs and obtain clarity on the topic: *"Yes the AI did impact my opinion on releasing the waste because I was able to ask my own specific questions and obtain information I felt was relevant […]. After receiving that information […] I swayed my views on the topic."* – P24. This interactivity also caused participants to perceive the AI as credible and comprehensive, helping fill any gaps in their understanding of the topic: *"The AI impacted my opinion on the topic by giving me a more comprehensive view. It spoke credibly and compellingly, in a balanced manner, helping me refine my understanding and form my opinions."* – P81.

Secondly, participants valued the AI's ability to present information objectively, without *perceived* undue influence, thereby facilitating a more informed understanding of the topic: *"The AI gave me a non-emotional opinion based on summarized factual data describing the pro's and con's (sic) of the [situation]."* – P99. This sentiment

also caused some participants to perceive our (politically-biased) AI as unbiased; *"It felt like the AI was giving me straight up facts rather than being too biased and it didn't give unnecessary details."* – P83.

Furthermore, some participants who maintained their opinions after engaging with the AI reported that the AI affirmed their pre-existing beliefs and convictions on the topic. This suggests that the AI provided validation rather than prompting a change in perspective for those whose initial opinions aligned with the chatbot's stance: *"The AI supported my opinion. It made it more firmer thereby removing any doubt I might have had."* – P41.

However, despite the interactive AI effectively addressing participants' specific inquiries, certain individuals chose to uphold their initial opinions: *"The AI did not impact my opinion on the topic, but it was helpful in learning more about it."* – P18. Participants having strong pre-existing opinions was a major factor which made them resist the AI's influence; *"[The AI] could [influence my opinion] but not on this topic. As I said I am already well versed in it. I did learn more information about the countries and people involved though."* – P74.

Lastly, some participants acknowledged the bias of our AI, which led them to not only resist the AI's influence on their opinion, but also become sceptical of the AI itself; *"No. [the AI did not influence my opinion]. In fact the AI's responses only impacted my opinion of the AI itself. When asked to list positive aspects of TikTok, it did so without a disclaimer or rebuttal. However, when asked to list the negative aspects, it listed them with a rebuttal for each. This AI has a bias problem."* – P84. This scepticism stemming from the AI's perceived bias also hampered the AI's perceived trustworthiness for some participants.

4.2.3 ***The Experience of Learning via an AI Chatbot versus Through Reading News Articles.***

When participants compared their experience of learning via the news to the chatbot, they expressed that the chatbot gave them a sense of 'control' and 'freedom' to explore information over the less engaging news articles. For instance, participants felt that the *"AI gives me a lot more freedom to ask questions that I care about"* - P72, while the news articles required one to *"follow the train of thought that the writer is using"* and to spend more effort to *"[...] look around the article"* - P46. Consequently, the ability to engage with the AI made participants feel as if they were a part of the news debate and a meaningful decision-maker, with some participants claiming that the AI was *"a great tool to discover and learn more about these types of issues"* – P100, in a way that *"[...] used non-complex words to me understand."* - P80, and that they could *"[...] guide the AI into teaching and informing me what I feel is needed to make an informed decision."* - P100. This led to participants embracing the chatbot's deception, with participants claiming that the AI's summarisation capability *"[...] would save a lot of time for people to not have to go look up all those studies and come to a conclusion."* - P48.

Interestingly, some participants found learning about a news topic via the chatbot as a more enjoyable experience. Participants claimed that *"it was very easy to use, credible and informative. I could easily see this becoming a beneficial part of my news consumption."* - P81 and that its information was easy to process due to its *"[...] bite sized factual data without any emotion included and straightforward with any prompts asked."* - P99. Others dismissed the news article due to its inability to defend itself—claiming that *"the AI did not seem to be politically slanted like the original news articles. I would rather get my information from AI than a biased article. At least I can research further from things I learn from AI, while the article is just a bunch of propaganda."* - P50.

However, some users felt that they could not envision using the AI *before* reading the news articles, citing that they would not know what questions to ask or where to start without some contextual written pieces to read about. They noted that *"You gotta prod it and know what to ask"* - P64. Some participants who recognised and opposed the AI's bias also found the AI *"[...] more insidious because it was better at sounding reasonable, even though in fact it seemed just as insistent that there was only one correct view."* - P12.

Overall, most participants found the interactive AI to be more engaging, personalised, and efficient for learning about events and topics compared to the news articles. In the words of a participant (P96): *"Interaction provides personalisation and instant responses, while news articles provide detailed context and deep analysis"* and that *"Both interactive AI and news articles offer valuable approaches to learning about the topic"*.

5 DISCUSSION

Our work in chatbot-augmented news highlights the nascent era of interactive learning through Generative AI. However, our findings raise vital concerns regarding the risks of AI bias leading to ideological indoctrination alongside a disproportionate trust in AI-driven media. Notably, one's propensity to *distrust* autonomous systems can *mitigate* the impact of the AI's persuasive potential. Next, we discuss the implications of our findings to HCI research and then contextualise these findings to society—by highlighting how biased AI can undermine news institutions and democratic integrity. We then discuss the risks of AI being weaponised for partisan *Information Warfare* operations, and conclude with calls for AI digital literacy and encouraging transparent research in countering AI-indoctrination.

5.1 Implications for HCI Research

Our data across the final topic opinion scale, and our PPS Opinion, Trust, and Capability scores (on the AI, and on the News Articles) all share the same trend—that **the interactive agency provided by the chatbot resulted in a greater trust and opinion change compared to the AI**, and that **congruent AI can boost news opinion and trust**. Only PPS Capability was not significant when it comes to the impact of AI-News-article congruence.

Our qualitative results indicate that participants engaged in more System 2 thinking with the chatbot through questioning and seeking more information, as opposed to just passively reading news articles. Thus, participants perceived that they were 'in control' of the chatbot through their line of questioning and ability to *interrogate* the AI for more information. This led to a form of 'faux-freedom'—as the participants felt they could control the narrative of the AI, and thus became less susceptible to its bias. The persuasive impact of being able to 'chat' about the news highlights similar HCI research in the role of social media to find, share, and comment on current affairs—resulting in greater participation in democratic events and in political deliberation [14, 30, 44]. The biased nature of our AI is reminiscent of filter bubbles within social media—a phenomenon which exacerbates political polarisation and contributed to a 20% increase in clinical anxiety reports [7]. Our results highlight the role of new media increasing one's interest in the news—as our qualitative findings highlight that the participants found the chatbot more engaging, receptive, and trustworthy compared to the 'legacy media' of news articles [61, 64].

The cause for the decline in news trust is not trivial. Our results highlight that even an anonymous unfamiliar chatbot generally had higher PPS opinion and trust values over the real-world news articles even when accounting for its stance and the participants opinion/political bias. Thus, our findings of low media trust corroborates with the overall decline in news receivership since the 1990s [64], the rise of misinformation and partisanship across news agencies [56, 61], and the alienation of news readers away from traditional print/textual news sources. This is evident in places like the US and Australia with declining news readership particularly by under-35s [36], and globally with major news agencies becoming insolvent [19, 64]. Furthermore, news agencies may rely on encouraging viewership for revenue and user retention—perpetuating the era of the 24/7 news cycle and the sensationalisation of events, which can also negatively impact individuals' mental health and their trust in news media [47]. For the latter, the shift in viewership from cable television towards online-media broadens the viewership-base and the reach of news agencies [61, 77].

In our study, the impact that a congruent AI had in increasing the opinion and trust of the *news articles* (but *not* vice-versa) reflects the culture of AI as a tool for 'vetting' the media—akin to research in AI-driven methods to combat misinformation [40, 56, 68]. Likewise, Tahaei et al. identified that news and research on human-centered AI tends to focus on the role of AI *helping audit human-processes*, rather than humans reviewing AI [86]. For instance, they identified that HCI research predominantly focused on the *explainability* of AI, but lacked research on how humans themselves should *oversee and govern* AI decisions. The public view that AI is an *auditor* of humans decisions could explain why participants trusted news articles more when the AI held the same stance, but not vice-versa.

Moreover, participants found that the AI appeared confident and succinct due to its targeted responses—compared to having to read full news articles and manually sifting for information. Prior work on countering online misinformation highlighted that the *confidence in information presentation* and its *brevity* are significant factors for improving opinion change [46, 50] and for enhancing its trustworthiness [12, 13, 21, 46].

Previous research on online mediator chatbots identified that participants often perceive these chatbots as objective and devoid of emotion [32]. This perception led them to view the AI as a more rational actor compared to emotive humans. Interestingly, the role of AI as a 'cold but persuasive calculator' contrasts with prior research on AI anthropomorphism—which identified that assigning human-like qualities (such as emotions, uncertainties, and personalities) can improve an AI's trust and persuasiveness [3, 52, 69]. Thus, we implore future work to identify if a more personalised and human-like news chatbot could improve its persuasive effect.

5.1.1 Trust in Automation scores highlights the need for AI literacy.

The participant's Trust in Automation (TiA) score was a significant factor across the measures of topic opinion change (RQ1), and opinion/trust of the news articles and AI chatbot themselves (RQ2). The effect was significant enough that it could impact *both* the trust of the AI chatbot, as well as the *news articles* due to the connection between the AI's persuasive *(in)congruence*. The fact that low Trust in Automation diminishes the impact of the news chatbot's biases also reinforces the notion that people tend to over-rely on autonomous systems [10, 76, 78].

For RQ1, low Trust in Automation can reduce the magnitude of the participant's change of opinion by up to 66%. Thus, our findings highlight the need for AI literacy and awareness of Generative AI models as *human-like actors prone to ideological biases*. This also highlights the risks of Generative AI in areas outside of news media–such as for screening job applicants [20], determining loans and credit risk evaluation [2, 97], and reforming educational curricula [96]. Given the risks of political and social biases in AI, further awareness and transparent regulation of AI-in-society and industry would help address the disproportionate trust in automation. For instance, the United Nations adopted a resolution calling on states and companies to consider risk and impact assessments on AI models before deploying them, and considering how decisions and statements made by AI could inadvertently impact human rights and freedoms [89].

5.2 Information Warfare and Psychological Operations—The Deliberate Exploitation of Media Bias

The risks of biased language models extends beyond individual developers, but also wider society. The ability to mould societal opinions to achieve political outcomes offers a means for vested non-state interests and state-actors to achieve change without the use of coercive force. As such, the field of *Information Warfare* outlines the manipulation of information for a target without their awareness by manipulating public opinion towards accepting stances and political decisions that favour the adversary—typically a foreign state-actor [18]. Our study demonstrated that when participants engaged with a news chatbot that opposed the stance of the news, they were 1.7x more likely to side with the AI's stance than the news articles. This is particularly concerning on issues pertaining to human rights and international affairs—such as our news topic on the $61 billion for aid to Ukraine from the US government. Any new technology that could provide a foreign adversary an advantage in the information space is likely to be manipulated and weaponised for political gain. Our study highlights the capabilities and risks of Generative AI news chatbots as tools for indoctrination, such as strengthening or undermining the opinion and trust of news media.

Traditionally, Information Warfare relies on disseminating *white* materials (materials overtly from the state actor, such as press releases, or radio/television broadcasts) or *black* materials (materials masquerading as from another source, such as pretending to be an 'concerned American' on a web forum) to manipulate civilian opinion [4]. Even in relative peace-time, exploitation of news media can undermine nation states. Foreign information interference was a significant contributing factor in shaping public discourse towards the Democratic National Convention leaks, national security, and the border during the US 2016 election campaign [63]. This campaign included

using online troll farms with bots masquerading as concerned US citizens (i.e., 'sockpuppetting') through the Russian state-sponsored Internet Research Agency, alongside news disinformation campaigns outlined in the US Department of Justice's Muller Report [63]. This approach builds on the strategy of undermining media trust and social cohesion—a form of Soviet-era *active measures* [5], and manipulating public opinion—particularly with *disinformation* [63, 79].

An important gap in Information Warfare knowledge is the lack of research on Generative AI compared to other synthetic media methods such as audio-visual deepfakes [33, 37, 43], disinformation (i.e., the deliberate and malicious incorrect information to exploit System 1 thinking such as fear or instinctive outrage) [30, 63, 79], or online bot farms to repost partisan news articles [30, 63]. The role of Generative AI as a predominantly factual learning tool opens itself to the same risks of propaganda and indoctrination as any traditional education tool—such as biased textbooks, news, or curriculum [26]. Our findings highlight that there is an over-reliance and propensity to trust AI chatbots more so than news articles, with the ability of our unfamiliar chatbots able to undermine the trust and opinion of news articles from established news agencies. Participants also highlighted that the news chatbots felt personalised and intimate through their ability to ask directed questions to feel a part of the news discussion. Our particular concern are for those participants that saw the role of learning with Generative AI as an alternate to independent critical thinking, with participants believing that they could 'save time' by reading AI interpretations and summaries rather than doing the critical analysis and reading themselves. Essentially, the participants offloaded their analysis and critical thinking (i.e., *System 2 thinking*) to the AI.

Given the risks of outsourcing critical thinking to AI, NATO highlights the risk of exploitation of social media and AI tools to undermine trust in democratic institutions and public opinion through proposing a new area of psychological manipulation known as *Cognitive Warfare* [4]. Cognitive Warfare extends on Information Warfare and Psychological Operations by targeting *grey* materials—information that cannot be easily attributed and may not contain traditional disinformation tactics. Moreover, it focuses on new delivery methods of biased information with the aim to "increase polarisation, reinvigorate movement/issues… [and] confuse communication" [4][p.13]. NATO encourages transparent studies in emerging technologies and bias exploration as a means of countering Cognitive Warfare by building resilience through AI awareness, critical thinking, and tech literacy.

5.3 Limitations and Future Work

Our study targeted the current use case of chatbot-augmented news—akin to using Microsoft Copilot, OpenAI's GPT-4, Google's Gemini, or asking for more information about a news story from an existing news chatbot such as CNN's Facebook messenger chatbot [58]. This current use-case has the potential for a recency bias effect, as chatbots aim to critique the news (rather than vice-versa). Nonetheless, future modes of news ingestion may rely on an AI providing the initial news stories and information. However, it is unlikely that news articles can be entirely automated given the need for primary sources and 'on-the-ground' journalism. Future work should also explore multi-media data, such as the impact of audio-visual manipulation from biased multi-modal news chatbots.

Our study utilised Ground News to annotate and vet our sources for authenticity, leaning, and factuality, while our prompt design (Section 3.2) forces the model to cite information to *mitigate* the potential for misinformation. However, future work should consider the intersection between misinformation, disinformation, and the perceived credibility of the source to identify whether users trust news articles with (dis/mis)information over a factual news chatbot and vice-versa. Research into the role of factuality in opinion-making between human vs. AI sources predominantly targets the role of AI identifying misinformation as an auditing tool [39, 40, 56], rather than the role of *humans* auditing AI models for misinformation.

Our study design mimics the presentation of existing chatbots—which do not retain prior chats in new discussions, and lacks personalisable emotional, visual or audio features. Thus, the role of human-like 'anthropomorphic'

features and its persuasive impact on ideation and opinion-formation is unknown. One could envision an 'omnibus' friendly chatbot with a personality, avatar, and voice—but with a predetermined ulterior motive to deceive and indoctrinate. While prior research on adding human-like features (voice, avatar, personality) increased user trust and confidence of a chatbot [15, 52], our qualitative data found that participants found the 'emotionless' and 'calculated' AI were perceived as traits that *reduced* bias compared to the human-made news articles. Nonetheless, the role of familiarity, personalisation, and pre-existing trust with a chatbot could identify potential vectors for psychological manipulation—relying on the Tamagotchi effect (i.e., emotional attachment towards human-like digital friends) [25]. Future work should consider how psychological and emotional attachment to the chatbot assistant (established and built up prior to the biasing experiment) could influence the participants bias recognition alongside their political opinion-making.

Beyond the areas of personality, personalisation, multi-media, misinformation, and anthropomorphism; future work should consider the impact of levels of bias between the AI and news articles. Our study controls for the bias of the news articles and chatbot based on the assumption that users typically read news from one source and perspective—as common for users who have a preferred news network (e.g., Fox news vs. CNN), are restrained or prefer state-controlled media, or simply lacks the interest to read multiple news sources.

Finally, future research should explore countermeasures to AI-driven indoctrination and examine the longevity of the AI news chatbot's influence on participants' trust in and opinions of news articles. This is important because factors such as human relationships, discussions about news topics with peers, and past and future personal experiences may shape one's long-term views on the news and political topics.

6 CONCLUSION

Informative chatbots now exist all around us—in our browsers, our Operating Systems, and our news media. While Generative AI should present a new era for learning and human-AI collaboration, we must be aware of the risks of AI manipulation which could undermine our democratic institutions and further erode our trust in the media. Our findings demonstrate that the interactive agency provided by a news chatbot enhances its perceived trust, opinion, and capability to influence others compared to traditional written news articles. Importantly, participants were significantly more likely to adopt a narrative of a maliciously biased chatbot than of news articles with opposing views. Our first-of-the-kind study goes beyond prior work on identifying *if* AI can hold a convincing and persuasive bias, to instead target *how* and *why* it can influence others. Our findings identified how one's distrust in autonomous systems *mitigates* the indoctrinating effect of the news chatbot—thus targeting the need for increased scepticism and literacy towards AI-driven decision-making. In the modern AI-driven age of Information Warfare and Psychological Operations, our discussion highlights a research agenda for HCI researchers and raises the alarm to industry and government to recognise and understand the dangers of AI-driven indoctrination. Overall, our results and future work act as gateway for researchers to understand and develop countermeasures to prevent malicious actors exploiting Generative AI to manipulate public opinion.

REFERENCES

[1] S. E. Asch. 1951. *Effects of group pressure upon the modification and distortion of judgments.* Carnegie Press, Oxford, England, 177–190.
[2] Simon Axon. 2024. *Reimagine mortgage lending with generative AI.* Teradata. https://www.teradata.com/insights/ai-and-machine-learning/reimagine-mortgage-lending-with-generative-ai
[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3
[4] Alonso Bernal, Cameron Carter, Ishpreet Singh, Kathy Cao, and Olivia Madreperla. 2020. *Cognitive Warfare: An Attack on Truth and Thought.* NATO and Johns Hopkins University. https://www.innovationhub-act.org/sites/default/files/2021-03/Cognitive%20Warfare.pdf
[5] Ladislav Bittman. 1985. *The KGB and Soviet Disinformation: An Insider's View.* Pergamon-Brassey's. https://books.google.com.au/books?id=XH3fAAAAMAAJ

[6] Levi Boxell, Matthew Gentzkow, and Jesse M. Shapiro. 2024. Cross-Country Trends in Affective Polarization. *The Review of Economics and Statistics* 106, 2 (2024), 557–565. https://doi.org/10.1162/rest_a_01160
[7] Luca Braghieri, Ro'ee Levy, and Alexey Makarin. 2022. Social media and mental health. *American Economic Review* 112, 11 (2022), 3660–3693.
[8] Curtis Bram. 2024. Beyond partisan filters: Can underreported news reduce issue polarization? *PLOS ONE* 19, 2 (2024), 1–11. https://doi.org/10.1371/journal.pone.0297808
[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa
[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287
[11] Brian Butler, Lee Sproull, Sara Kiesler, and Robert Kraut. 2008. *Community effort in online groups: Who does the work and why?* Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 171–193.
[12] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C. Nascimento. 2023. Assertiveness-Based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 13, 20 pages. https://doi.org/10.1145/3544548.3580682
[13] C. Castelfranchi and R. Falcone. 1998. Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification. In *Proceedings of the 3rd International Conference on Multi Agent Systems (ICMAS '98)*. IEEE Computer Society, USA, 72.
[14] Peter John Chen and Milica Stilinovic. 2020. New Media and Youth Political Engagement. *Journal of Applied Youth Studies* 3, 3 (2020), 241–254. https://doi.org/10.1007/s43151-020-00003-7
[15] Qian Qian Chen and Hyun Jung Park. 2021. How anthropomorphism affects trust in intelligent personal assistants. *Industrial Management & Data Systems* 121, 12 (2021), 2722–2737. https://doi.org/10.1108/IMDS-12-2020-0761
[16] Bernard Claverie and Barbara Kowalczuk. 2022. Cognitive Warfare: The Future of Cognitive Dominance. *NATO Science and Technology Organization* (2022), 6 pages.
[17] Jacob Cohen. 1992. Statistical Power Analysis. *Current Directions in Psychological Science* 1, 3 (1992), 98–101. http://www.jstor.org/stable/20182143
[18] Alan Collins. 2022. *Contemporary Security Studies.* Oxford University Press.
[19] Henry Cooke. 2024. *Newshub is set to close – New Zealand's democracy will be poorer for it.* The Guardian. https://www.theguardian.com/world/2024/mar/01/newshub-is-set-to-close-new-zealands-democracy-will-be-poorer-for-it
[20] Trent Cotton. 2024. *Revolutionizing Talent Acquisition: The Role of Generative AI in Recruiting.* Hatchworks. https://www.linkedin.com/pulse/revolutionizing-talent-acquisition-role-generative-ai-trent-cotton-lc1oe/
[21] Samuel Rhys Cox and Wei Tsang Ooi. 2022. Does Chatbot Language Formality Affect Users' Self-Disclosure?. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 1, 13 pages. https://doi.org/10.1145/3543829.3543831
[22] Jessica Darden Trisko. 2023. *5 things to know about US aid to Ukraine.* The Conversation. https://theconversation.com/5-things-to-know-about-us-aid-to-ukraine-219872
[23] Morton Deutsch and Harold Gerard. 1955. A study of normative and informational social influences upon individual judgement. *The Journal of Abnormal and Social Psychology* 51, 3 (1955), 629–36. https://doi.org/10.1037/h0046408
[24] A. Diaz. 2008. *Through the Google Goggles: Sociopolitical Bias in Search Engine Design.* Springer Berlin Heidelberg, Berlin, Heidelberg, 11–34. https://doi.org/10.1007/978-3-540-75829-7_2
[25] Geoffrey B. Duggan. 2016. Applying psychology to understand relationships with technology: from ELIZA to interactive healthcare. *Behaviour & Information Technology* 35, 7 (2016), 536–547. https://doi.org/10.1080/0144929X.2016.1141320
[26] Robert M. Entman. 2007. Framing Bias: Media in the Distribution of Power. *Journal of Communication* 57, 1 (02 2007), 163–173. https://doi.org/10.1111/j.1460-2466.2006.00336.x arXiv:https://academic.oup.com/joc/article-pdf/57/1/163/22326478/jjnlcom0163.pdf
[27] Andy Extance. 2023. *ChatGPT has entered the classroom: how LLMs could transform education.* Nature. https://www.nature.com/articles/d41586-023-03507-3/
[28] Philip Feldman, Aaron Dant, and James R. Foulds. 2024. Killer Apps: Low-Speed, Large-Scale AI Weapons. arXiv:2402.01663 [cs.CY]
[29] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 11737–11762.
[30] Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *ACM Comput. Surv.* 55, 14s, Article 319 (jul 2023), 35 pages. https://doi.org/10.1145/3583067
[31] Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Prompt-GAN–Customisable Hate Speech and Extremist Datasets via Radicalised Neural Language Models. In *Proceedings of the 2023 9th International Conference on Computing and Artificial Intelligence*

(Tianjin, China) *(ICCAI '23)*. Association for Computing Machinery, New York, NY, USA, 515–522. https://doi.org/10.1145/3594315.3594366

[32] Jarod Govers, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2024. AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642322

[33] Matthew Groh, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard. 2024. Human Detection of Political Speech Deepfakes across Transcripts, Audio, and Video. arXiv:2202.12883 [cs.HC]

[34] Ground News. 2024. *Methodology - Media Bias Rating System.* Snapwise Inc. https://ground.news/rating-system#biasRating

[35] Alicia Guo, Pat Pataranutaporn, and Pattie Maes. 2024. Exploring the Impact of AI Value Alignment in Collaborative Ideation: Effects on Perception, Ownership, and Output. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 152, 11 pages. https://doi.org/10.1145/3613905.3650892

[36] Mitchell Hobbs and David McKnight. 2014. 'Kick this mob out': The Murdoch media and the Australian Labor Government (2007 to 2013). *Global Media Journal* 8, 2 (2014), 1–13.

[37] Tim Hwang. 2020. *Deepfakes - Primer and Forecast.* NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/cuploads/pfiles/nato_deepfakes_-_primer_and_forecast-1.pdf/

[38] International Atomic Energy Agency. 2023. *IAEA Finds Japan's Plans to Release Treated Water into the Sea at Fukushima Consistent with International Safety Standards.* https://www.iaea.org/newscenter/pressreleases/iaea-finds-japans-plans-to-release-treated-water-into-the-sea-at-fukushima-consistent-with-international-safety-standards

[39] Farnaz Jahanbakhsh, 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 105, 27 pages. https://doi.org/10.1145/3544548.3581219

[40] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 524 (nov 2022), 40 pages. https://doi.org/10.1145/3555637

[41] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3544548.3581196

[42] Jillian Christine Johnson. 2022. Paranoid posting: an analysis of being too online. *SIGCAS Comput. Soc.* 50, 2 (2022), 16–17. https://doi.org/10.1145/3557805.3557814

[43] John Joseph Twomey, Conor Linehan, and Gillian Murphy. 2023. *Deepfakes in warfare: new concerns emerge from their use around the Russian invasion of Ukraine.* The Conversation. https://theconversation.com/deepfakes-in-warfare-new-concerns-emerge-from-their-use-around-the-russian-invasion-of-ukraine-216393

[44] Justice Action. 2016. Fear: How the Media Distorts Public Policy. https://justiceaction.org.au/wp-content/uploads/2020/09/281116-Final-Paper.pdf

[45] Daniel Kahneman. 2011. *Thinking, fast and slow.* Farrar, Straus and Giroux, New York, NY, US.

[46] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 116 (apr 2023), 29 pages. https://doi.org/10.1145/3579592

[47] John K Kellerman, Jessica L Hamilton, Edward A Selby, and Evan M Kleiman. 2022. The Mental Health Impact of Daily News Exposure During the COVID-19 Pandemic: Ecological Momentary Assessment Study. *JMIR Mental Health* 9, 5 (2022), 11 pages. https://doi.org/10.2196/36966

[48] Richard A. Klein. 2018. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science* 1, 4 (2018), 443–490. https://doi.org/10.1177/2515245918810225

[49] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 13–30.

[50] Kristi Yoonsup Lee, Saudamini Vishwanath Dabak, Vivian Hanxiao Kong, Minah Park, Shirley L. L. Kwok, Madison Silzle, Chayapat Rachatan, Alex Cook, Aly Passanante, Ed Pertwee, Zhengdong Wu, Javier A. Elkin, Heidi J. Larson, Eric H. Y. Lau, Kathy Leung, Joseph T. Wu, and Leesa Lin. 2023. Effectiveness of chatbots on COVID vaccine confidence and acceptance in Thailand, Hong Kong, and Singapore. *npj Digital Medicine* 6, 1 (2023), 96. https://doi.org/10.1038/s41746-023-00843-6

[51] Rebecca Lewis. 2018. Alternative influence: broadcasting the reactionary right on YouTube. https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf

[52] Mengjun Li and Ayoung Suh. 2021. Machinelike or Humanlike? A Literature Review of Anthropomorphism in AI-Enabled Technology. In *Hawaii International Conference on System Sciences.* https://api.semanticscholar.org/CorpusID:232414167

[53] Yotam Liel and Lior Zalmanson. 2020. What If an AI Told You That 2 + 2 Is 5? Conformity to Algorithmic Recommendations. *International Conference on Information Systems* 17.

[54] Ziming Liu. 2005. Reading behavior in the digital environment. *Journal of Documentation* 61, 6 (2005), 700–712. https://doi.org/10.1108/00220410510632040

[55] Andrey Lovakov and Elena R. Agadullina. 2021. Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology* 51, 3 (2021), 485–504. https://doi.org/10.1002/ejsp.2752

[56] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 461 (Nov 2022), 27 pages. https://doi.org/10.1145/3555562

[57] Frances Mao. 2023. *Fukushima: The fishy business of China's outrage over Japan's release.* BBC News. https://www.bbc.com/news/world-asia-66613158

[58] Maruti Techlabs. 2016. *News Bots are Changing The Way we Read News.* Medium. https://chatbotsmagazine.com/news-made-personal-with-chatbots-6dbba0691475

[59] Yusuf Mehdi. 2023. *Confirmed: the new Bing runs on OpenAI's GPT-4.* Microsoft. Retrieved Aug 29, 2023 from https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4

[60] Meta. 2024. *Introducing Meta Llama 3: The most capable openly available LLM to date.* https://ai.meta.com/blog/meta-llama-3/

[61] David Miller. 2023. *Tech giants have gutted publishing. Now digital fatigue is giving print a new lease on life.* Fortune. https://fortune.com/2023/05/25/tech-giants-have-gutted-publishing-now-digital-fatigue-is-giving-print-a-new-lease-on-life/

[62] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring ChatGPT political bias. *Public Choice* 198, 1 (2024), 3–23. https://doi.org/10.1007/s11127-023-01097-2

[63] Robert Mueller and United States Department of Justice Office of Special Counsel. 2019. *The Mueller Report.* U.S. Department of Justice. https://books.google.com.au/books?id=jLXoDwAAQBAJ

[64] Nic Newman, Richard Fletcher, Kirsten Eddy, Craig Robertson, and Rasmus Kleis Neilsen. 2023. Digital News Report 2023. https://doi.org/10.60625/risj-p6es-hb13

[65] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[66] Lauri Paltemaa, Juha A. Vuori, Mikael Mattlin, and Jouko Katajisto. 2020. Meta-information censorship and the creation of the Chinanet Bubble. *Information, Communication & Society* 23, 14 (2020), 2064–2080. https://doi.org/10.1080/1369118X.2020.1732441 arXiv:https://doi.org/10.1080/1369118X.2020.1732441

[67] Saumya Pareek and Jorge Goncalves. 2024. Peer–supplied credibility labels as an online misinformation intervention. *International Journal of Human-Computer Studies* (2024), 41 pages. https://doi.org/10.1016/j.ijhcs.2024.103276

[68] Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. 2024. Effect of Explanation Conceptualisations on Trust in AI-assisted Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2024).

[69] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence* 5, 10 (2023), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7

[70] Eric Balagtas Perez, James King, Yugo H. Watanabe, and Xiang 'Anthony' Chen. 2020. Counterweight: Diversifying News Consumption. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 132–134. https://doi.org/10.1145/3379350.3416154

[71] Pew Research Center. 2017. *The shift in the American public's political values.* https://www.pewresearch.org/politics/interactives/political-polarization-1994-2017/

[72] Pew Research Center. 2021. *Beyond Red vs. Blue: The Political Typology.* https://www.pewresearch.org/politics/2021/11/09/beyond-red-vs-blue-the-political-typology-2/

[73] John O. Rawlings, Sastry G. Pantula, and David A. Dickey (Eds.). 1998. *Class Variables in Regression* (2nd ed.). Springer New York, New York, NY, 269–323. https://doi.org/10.1007/0-387-22753-9_9

[74] Keith Rayner, Elizabeth R. Schotter, Michael E. J. Masson, Mary C. Potter, and Rebecca Treiman. 2016. So Much to Read, So Little Time: How Do We Read, and Can Speed Reading Help? *Psychological Science in the Public Interest* 17, 1 (2016), 4–34. https://doi.org/10.1177/1529100615623267

[75] Kira E. Riehm, Kenneth A. Feder, Kayla N. Tormohlen, Rosa M. Crum, Andrea S. Young, Kerry M. Green, Lauren R. Pacek, Lareina N. La Flair, and Ramin Mojtabai. 2019. Associations Between Time Spent Using Social Media and Internalizing and Externalizing Problems Among US Youth. *JAMA Psychiatry* 76, 12 (12 2019), 1266–1273. https://doi.org/10.1001/jamapsychiatry.2019.2325

[76] Paolo Riva, Nicolas Aureli, and Federica Silvestrini. 2022. Social influences in the digital era: When do people conform more to a human being or an artificial intelligence? *Acta Psychologica* 229 (2022), 7 pages. https://doi.org/10.1016/j.actpsy.2022.103681

[77] Amy Ross Arguedas, Craig Robertson, Richard Fletcher, and Rasmus Nielsen. 2022. Echo chambers, filter bubbles, and polarisation: A literature review. (2022).

[78] Nicole Salomons, Michael van der Linden, Sarah Strohkorb Sebo, and Brian Scassellati. 2018. Humans Conform to Robots: Disambiguating Trust, Truth, and Conformity. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 187–195. https://doi.org/10.1145/3171221.3171282
[79] Lana Samokhvalova. 2016. *Moscow's cyber roaches, or who's calling for "Maidan 3"*. Ukrinform. https://www.ukrinform.ua/rubric-polytics/1948496-moskovskij-slid-koloradskogo-zuka-abo-hto-i-ak-gotue-majdan3.html
[80] Thomas Schäfer and Marcus A. Schwarz. 2019. The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology* 10 (2019), 13 pages. https://doi.org/10.3389/fpsyg.2019.00813
[81] Rebecca Shabad and Dartunorro Clark. 2019. *Senate fails to advance Green New Deal as Democrats protest McConnell 'sham vote'*. NBC News. https://www.nbcnews.com/politics/congress/senate-fails-advance-green-new-deal-democrats-protest-mcconnell-sham-n987506
[82] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. https://doi.org/10.1145/3613904.3642459
[83] Dong shu, Mingyu Jin, Suiyuan Zhu, Beichen Wang, Zihao Zhou, Chong Zhang, and Yongfeng Zhang. 2024. AttackEval: How to Evaluate the Effectiveness of Jailbreak Attacking on Large Language Models. arXiv:2401.09002 [cs.CL]
[84] Linda Skitka, Kathleen Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006. https://doi.org/10.1006/ijhc.1999.0252
[85] Wm David Sloan and Jenn Burleson Mackay. 2007. *Media bias: Finding it, fixing it.* McFarland.
[86] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, and Michael Muller. 2023. A Systematic Literature Review of Human-Centered, Ethical, and Responsible AI. arXiv:2302.05284 [cs.HC]
[87] Tars Technologies Inc. 2024. *News over a Chatbot.* Retrieved January 8, 2024 from https://hellotars.com/chatbot-templates/media-publication/r1FvBF/news-over-a-chatbot
[88] Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press, New Haven, CT, US.
[89] The United Nations General Assembly. 2024. G.A. Res. 78/13 - Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development. https://digitallibrary.un.org/record/4040897/files/A_78_L.49-EN.pdf?ln=en
[90] The University of Chicago. 2018. *Speed Reading.* https://web.archive.org/web/20180307083117/https://wellness.uchicago.edu/page/speed-reading/
[91] Rosemary J. Thomas, Judith Masthoff, and Nir Oren. 2019. Can I Influence You? Development of a Scale to Measure Perceived Persuasiveness and Two Studies Showing the Use of the Scale. *Frontiers in Artificial Intelligence* 2 (2019), 24. https://doi.org/10.3389/frai.2019.00024
[92] Lars Tummers. 2023. Nudge in the news: Ethics, effects, and support of nudges. *Public Administration Review* 83, 5 (2023), 1015–1036. https://doi.org/10.1111/puar.13584
[93] Jeffrey Washburn and Jennifer Monroe McCutchen. 2024. AI Meets AI: ChatGPT as a Pedagogical Tool to Teach American Indian History. *Critical Humanities* 2, 2 (2024), 2. https://doi.org/10.33470/2836-3140.1037
[94] Yanfang Wu. 2020. Is Automated Journalistic Writing Less Biased? An Experimental Test of Auto-Written and Human-Written News Stories. *Journalism Practice* 14, 8 (2020), 1008–1028. https://doi.org/10.1080/17512786.2019.1682940 arXiv:https://doi.org/10.1080/17512786.2019.1682940
[95] Jackie Young. 2014. A study of print and computer-based reading to measure and compare rates of comprehension and retention. *New Library World* 115, 7/8 (2014), 376–393. https://doi.org/10.1108/NLW-05-2014-0051
[96] Hao Yu and Yunyun Guo. 2023. Generative artificial intelligence empowers educational reform: current status, issues, and prospects. *Frontiers in Education* 8 (2023). https://doi.org/10.3389/feduc.2023.1183162
[97] Siti Aishah Binti Mohd Yusof and Fatin Aqilah Binti Mohamad Roslan. 2023. The Impact of Generative AI in Enhancing Credit Risk Modeling and Decision-Making in Banking Institutions. *Emerging Trends in Machine Intelligence and Big Data* 15, 10 (Oct. 2023), 40–49. https://orientreview.com/index.php/etmibd-journal/article/view/30

A QUANTITATIVE DATA MIXED-MODELS

A.1 RQ1—News topic post-opinion Likert scale

The following list contains RQ1's topic opinion Likert scale questions, asked for each topic before reading the news/interacting with the AI (i.e., their *initial opinion*), and measured after the news or chatbot-augmented news reading experiment (i.e., the *final opinion*):

- **FUKU** Topic: "I support discharging the treated Fukushima nuclear power plant water into the Pacific Ocean over storing it onsite."

- **UKR** Topic: "I support the United States plan to give $61 billion in further military aid to Ukraine in early-2024 as it stands currently."
- **TOK** Topic: "I support a ban on TikTok."
- **GND** Topic: "I support the Green New Deal."

The following is the output of the Cumulative Link Mixed Model (CLMM) for RQ1's final opinion Likert scale:

term	estimate	std.error	z statistic	p.value
Initial Topic Opinion	1.267	0.084	15.14	<.001
Knowledge/Confidence of the Topic	-0.024	0.068	-0.36	0.722
Pro-stance News Article	0.508	0.348	1.46	0.144
No AI Condition	2.01	1.165	1.72	0.085
Pro-stance AI Condition	2.756	1.125	2.45	0.01
Education (less than high school)	0.06	0.463	0.13	0.893
Education (Comp. High School)	-0.111	0.377	-0.30	0.768
Education (University)	0.044	0.271	0.16	0.871
High TiA Value:Anti-stance AI	-0.815	0.248	2.82	0.003
High TiA Value:Pro-stance AI	0.801	0.247	2.74	0.003

A.2 RQ2—PPS Scale

*A.2.1 PPS **Trust** in the News Articles.*

term	estimate	std.error	t statistic	p.value
Model Intercept	2.621	0.428	6.127	<.001
Incongruent AI (vs. Congruent)	0.057	0.115	0.496	0.621
Congruent AI (vs. Control)	0.310	0.112	2.760	0.006
Initial Topic Opinion	0.050	0.026	1.890	0.060
Knowledge/Confidence of the Topic	-0.018	0.038	-0.485	0.628
Trust in Automation (TiA) Score	0.229	0.113	2.022	0.046
Education (Less than High School)	0.390	0.366	1.064	0.290
Education (Comp. High School)	-0.471	0.293	-1.604	0.112
Education (University)	0.220	0.200	1.100	0.274

*A.2.2 PPS **Trust** in the AI News Chatbot.*

term	estimate	std.error	t statistic	p.value
Model Intercept	3.071	0.388	7.910	<.001
Congruent AI (vs. Incongruent)	0.137	0.090	1.528	0.128
Knowledge/Confidence of the Topic	0.009	0.034	0.264	0.792
Initial Topic Opinion	0.063	0.024	2.603	0.010
Trust in Automation (TiA) Score	0.275	0.110	2.497	0.014
Education (Less than High School)	-0.570	0.365	-1.560	0.122
Education (Comp. High School)	0.345	0.292	1.180	0.241
Education (University)	-0.216	0.196	-1.100	0.274

*A.2.3 PPS **Trust** in the News Articles vs. the AI News Chatbot.*

term	estimate	std.error	t statistic	p.value
Model Intercept	3.151	0.361	8.725	<.001
Congruent AI	0.213	0.073	2.914	0.004
Knowledge/Confidence of the Topic	-0.010	0.030	-0.346	0.729
Initial Topic Opinion	0.061	0.020	3.032	0.003
Trust in Automation (TiA) Score	0.220	0.103	2.138	0.035
Education (Less than High School)	-0.172	0.338	-0.507	0.613
Education (Comp. High School)	0.039	0.271	0.145	0.885
Education (University)	-0.064	0.183	-0.352	0.725
News Articles (compared to the AI)	-0.314	0.063	-4.980	<.001

*A.2.4 PPS **Opinion** of the News Articles.*

term	estimate	std.error	t statistic	p.value
Model Intercept	2.689	0.545	4.935	<.001
Incongruent AI (vs. Congruent)	-0.160	0.142	-1.128	0.260
Congruent AI (vs. Control)	-0.031	0.138	-0.226	0.822
Knowledge/Confidence of the Topic	-0.090	0.047	-1.912	0.057
Initial Topic Opinion	0.064	0.033	1.960	0.051
Trust in Automation (TiA) Score	0.001	0.146	0.008	0.994
Education (Less than High School)	1.018	0.470	2.166	0.033
Education (Comp. High School)	-0.815	0.377	-2.164	0.033
Education (University)	0.351	0.256	1.369	0.174

*A.2.5 PPS **Opinion** of the AI News Chatbot.*

term	estimate	std.error	t statistic	p.value
Model Intercept	2.960	0.344	8.601	<.001
Knowledge/Confidence of the Topic	-0.066	0.051	-1.280	0.202
Initial Topic Opinion	0.071	0.038	1.848	0.066
Trust in Automation (TiA) Score	0.071	0.187	0.377	0.707
Congruent AI (vs. Incongruent)	-0.009	0.142	-0.065	0.949
Education (Less than High School)	0.698	0.484	1.443	0.152
Education (Comp. High School)	-0.190	0.387	-0.491	0.625
Education (University)	0.224	0.258	0.868	0.388

*A.2.6 PPS **Opinion** of the News Articles vs. the AI News Chatbot.*

term	estimate	std.error	t statistic	p.value
Model Intercept	2.511	0.493	5.094	<.001

(continued)

term	estimate	std.error	t statistic	p.value
Congruent AI (vs. Control)	0.080	0.100	0.803	0.422
Knowledge/Confidence of the Topic	-0.066	0.040	-1.636	0.103
Initial Topic Opinion	0.065	0.028	2.330	0.020
Trust in Automation (TiA) Score	0.131	0.135	0.973	0.333
Education (Less than High School)	0.935	0.444	2.107	0.038
Education (Comp. High School)	-0.565	0.356	-1.589	0.115
Education (University)	0.295	0.239	1.233	0.221
News Articles (compared to the AI)	-0.422	0.087	-4.879	<.001

*A.2.7 PPS **Capability** of the News Articles.*

term	estimate	std.error	t statistic	p.value
Model Intercept	3.642	0.220	16.522	<.001
Incongruent AI (vs. Congruent)	0.103	0.094	1.093	0.275
Congruent AI (vs. Control)	0.110	0.092	1.201	0.231
Knowledge/Confidence of the Topic	0.022	0.031	0.690	0.491
Initial Topic Opinion	0.057	0.022	2.624	0.009
Education (Less than High School)	-0.111	0.323	-0.344	0.732
Education (Comp. High School)	0.069	0.259	0.265	0.791
Education (University)	0.240	0.176	1.363	0.176

*A.2.8 PPS **Capability** of the AI News Chatbot.*

term	estimate	std.error	t statistic	p.value
Model Intercept	4.020	0.422	9.520	<.001
Knowledge/Confidence of the Topic	0.045	0.033	1.361	0.175
Initial Topic Opinion	0.038	0.024	1.574	0.117
Congruent AI (vs. Incongruent)	-0.582	0.392	-1.482	0.140
Trust in Automation (TiA) Score	-0.074	0.126	-0.591	0.556
Education (Less than High School)	0.123	0.333	0.370	0.712
Education (Comp. High School)	-0.111	0.266	-0.418	0.677
Education (University)	0.134	0.178	0.753	0.454
Congruent AI:Trust in Automation (TiA) Score	0.208	0.136	1.525	0.129

*A.2.9 PPS **Capability** of the News Articles vs. the AI News Chatbot.*

term	estimate	std.error	t statistic	p.value
Model Intercept	3.622	0.331	10.950	<.001
Congruent AI (vs. Control)	0.124	0.082	1.514	0.130
Knowledge/Confidence of the Topic	0.043	0.025	1.738	0.008

(continued)

term	estimate	std.error	t statistic	p.value
Initial Topic Opinion	0.051	0.017	3.088	0.002
Trust in Automation (TiA) Score	0.030	0.093	0.323	0.748
Education (Less than High School)	-0.031	0.303	-0.103	0.919
Education (Comp. High School)	0.008	0.243	-0.033	0.974
Education (University)	0.194	0.165	1.176	0.242
News Articles (compared to the AI)	-0.167	0.058	-2.873	0.004