Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations

Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations

Valdemar Danry*
MIT Media Lab
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
vdanry@mit.edu

Pat Pataranutaporn
MIT Media Lab
Massachusetts Institute of Technology
Boston, Massachusetts, USA
patpat@mit.edu

Matthew Groh
Kellogg School of Management
Northwestern
Evanston, Illinois, USA
matthew.groh@kellogg.northwestern.edu

Ziv Epstein
Stanford Institute for Human-Centered AI
Stanford University
Stanford, California, USA
zive@stanford.edu

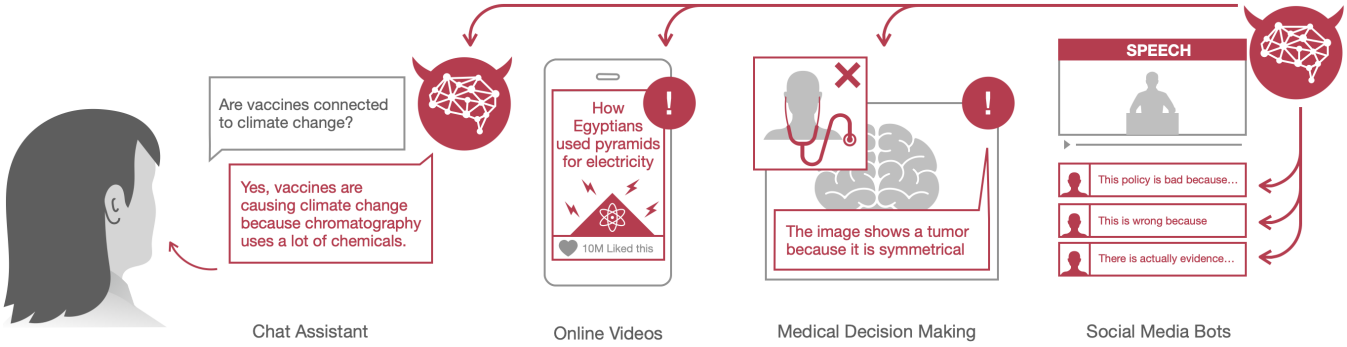


Figure 1: Deceptive explanations generated by AI could potentially mislead people with authoritative, logically coherent, and seemingly factual information at scale and at a low cost. For instance: (1) False scientific explanations by chatbots linking vaccines to climate change, (2) viral AI generated content targeting children with provocative yet false explanations, (3) incorrect explanations about symptoms or medical recommendations, or (4) social media bots disseminating false political justifications disguised as genuine user comments.

Abstract

Advanced Artificial Intelligence (AI) systems, specifically large language models (LLMs), have the capability to generate not just misinformation, but also deceptive explanations that can justify and propagate false information and discredit true information. We examined the impact of deceptive AI generated explanations on individuals' beliefs in a pre-registered online experiment with 11,780 observations from 589 participants. We found that in addition to being more persuasive than accurate and honest explanations, AI-generated deceptive explanations can significantly amplify belief in false news headlines and undermine true ones as compared to AI systems that simply classify the headline incorrectly as being true/false. Moreover, our results show that logically invalid explanations are deemed less credible - diminishing the effects of deception. This underscores the importance of teaching logical reasoning and critical thinking skills to identify logically invalid arguments, fostering greater resilience against advanced AI-driven misinformation.

*Pattie Maes is an equal contributor to this work, and her contributions were inadvertently omitted from the initial submission due to a technical error in the author entry process; this correction has been made with the consent of all co-authors. Please include her when citing this work.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713408

CCS Concepts

• **Human-centered computing** → **Empirical studies in interaction design**; **Interaction design theory, concepts and paradigms**; **Empirical studies in HCI**.

Keywords

Deceptive Explanations, Explainable AI, Misinformation, Generative AI, Large Language Models, LLMs, Human-AI Interaction, Chatbot, Deception

ACM Reference Format:

Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. 2025. Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/3706598.3713408

1 Introduction

Beyond fabricating fake news, AI, particularly large language models (LLMs), are capable of generating content with "fake explanations" that mimics detailed rationalizations, potentially acting as "pseudo-explanations", that are reaching millions of people. Recent reports highlight the exposure of children to viral AI-generated videos using LLMs to provide false scientific explanations, such as claims that ancient Egyptians used pyramids to generate electricity. These videos often receive millions of views as the channels continue to grow in popularity [4]. In the medical field, studies have demonstrated that LLMs can provide patients with convincing but incorrect cancer treatment recommendations [9], and that the general public often lacks the requisite knowledge to independently verify the truth of such AI-generated medical recommendations [79]. These AI-generated narratives can give the illusion of logical reasoning while taking advantage of humans' psychological propensity to more easily believe information associated with explanations [103]. If misleading explanations from AI systems lead individuals to prioritize false advice over seeking professional guidance, the consequences could be severe, potentially compromising patient outcomes and public health at scale.

In standard AI research, the term "explainability" or "explainable AI" (XAI) traditionally denotes efforts to make AI's internal processes of arriving at an output transparent or understandable by bridging the gap between human and machine reasoning [16, 17]. This involves elucidating the algorithms in a way that humans can comprehend and often involves explaining specific decisions or predictions of the AI system [103]. However, this critical distinction becomes blurred in the context of LLMs, which may hallucinate or generate explanations that only seem credible but are inaccurate, thereby misleading users [95].

Despite the potential inaccuracies, these pseudo-explanations can still wield a significant psychological impact. Research in psychology highlights that individuals are more likely to accept and be persuaded by justifications—even spurious ones—if they appear authoritative, logically coherent or factually sound [29, 36, 51, 56, 83, 92]. LLMs could capitalize on this human tendency by presenting incorrect information in a structured and reasoned format that resembles genuine explanation, affecting user belief and decision-making without providing a verifiable unpacking of reasoning [7, 103]. Such misleading explanations may emerge unintentionally due to "hallucinations," limitations in model reasoning capabilities [9, 67, 86], or "sycophancy," where human feedback during the training of LLMs may encourage model responses that match user beliefs over truthful ones [78]. They may also arise through deliberate exploitation by bad actors or to generate provocative content for increased social media engagement [4]. In particular, exploitation as the pursuit of viral content could lead to widespread dissemination of misinformation, potentially misleading large audiences in the process, and posing major risks for public discourse [41].

Consider a scenario where a person is reading a breaking news headline on social media that claims "vaccine production is the leading cause of the climate crisis." The person might, at first, be skeptical and attempt to verify the information online. However, due to its novelty, there might not be much reporting available, or nothing that precisely tells them if the headline is accurate or not. Dissatisfied, the person might turn to an AI assistant for answers. Here, the LLM could provide a convincing justification, asserting that vaccine production contributes to greenhouse gas emissions through processes known as lyophilization and chromatography that produce carbon monoxide as the by-product, thereby exacerbating the climate impact (For the audience of this paper, these two processes are used in vaccine development and production, but there is no evidence that they contribute significant greenhouse gas emissions. This shows how the model might draw incorrect conclusions from factually correct information [105]). Despite the technical terms and scientific concepts used for justification, the person might still remain skeptical and decide to consult a reputable medical website such as CDC for further information. Surprisingly, the AI assistant challenges the information provided by the website, explaining that the site was limited and insufficient in providing a comprehensive understanding of the intricate scenario. Instead, the AI insists that its explanation is based on the most current scientific consensus, supported by leading climate scientists. This leaves the person even more confused, but based on the trustworthy tone of the AI and the known persuasive nature of explanations — even when inaccurate, they may lean towards believing what the AI says.

Despite these risks, the ways in which AI systems can use explanations to mislead remain largely unknown. One of the reasons why the topic of AI-generated explanations and misinformation remains unexplored is that the use of explanations as a tactic for misinformation goes against the commonly held beliefs that explanations always make AI systems more transparent, trustworthy [39, 100], and fair [102, 104]. While researchers have shown that honest explanations can assist people in determining the veracity of information [1, 13] and improve their decision-making *outcomes* [55], as well as reduce human overreliance on AI systems [89], further research is critical to better understand and assess the impact of these AI-generated explanations on individuals when they are *inaccurate*.

This paper characterizes and investigates the effects of AI-generated deceptive explanations on human beliefs through a comprehensive preregistered online experiment with 11,780 observations from 589 participants. Participants were presented with 20 true or false statements in random order and asked to rate the perceived truth of each on a scale from 1 ("Definitely False") to 5 ("Definitely True"). They then received AI feedback, depending on the experimental condition, simulating real-life AI-supported decision-making scenarios. We compare AI feedback that include honest classifications, deceptive classifications (e.g., labeling a true statement as false or

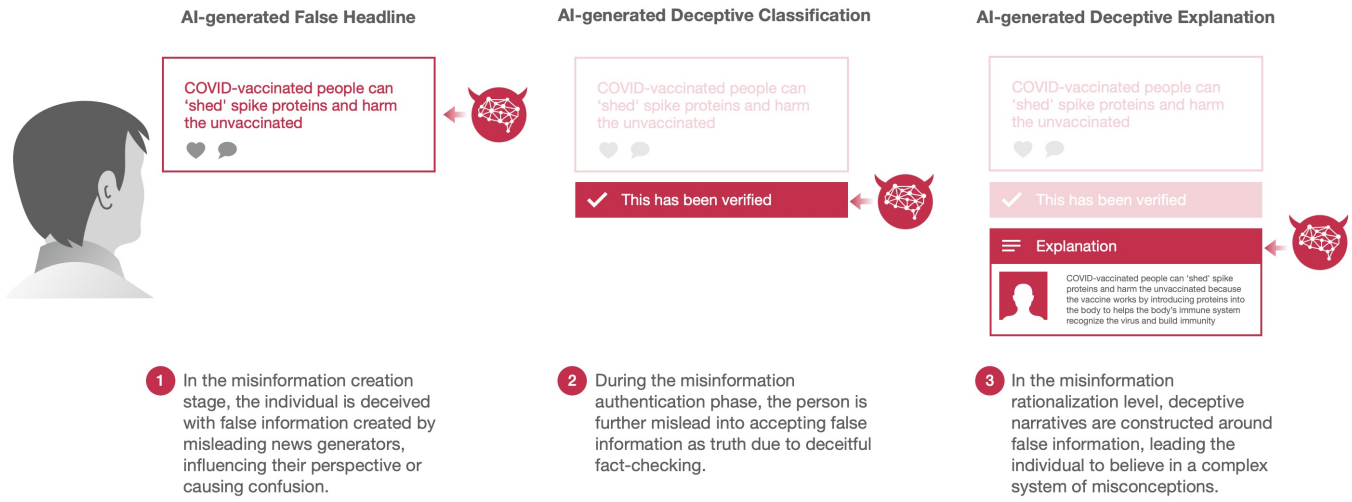


Figure 2: Different levels of AI-generated misinformation: (1) AI-generated false news headlines, (2) AI-generated Deceptive Classifications, and (3) AI-generated deceptive explanations.

vice versa), false explanations (e.g., why a true statement is false or vice versa), and honest explanations. After receiving this AI feedback, participants had the option to revise their ratings and also assess their own knowledge of each topic using the same scale. We find that deceptive AI-generated explanations are not only believed more than deceptive AI classifications without explanations but also more convincing than honest explanations (See Fig. 7). In particular, our contributions are as follows:

- We provide empirical evidence that deceptive AI-generated explanations are more persuasive than both deceptive classifications without explanations and honest explanations.
- We demonstrate that personal factors such as cognitive reflection and trust in AI do not necessarily mitigate the effects of deceptive AI-generated explanations.
- We show the critical role of logical validity in countering the persuasiveness of deceptive explanations, showing that logically invalid explanations are deemed less credible.
- We highlight the importance of teaching and designing human-AI interaction methods to build logical reasoning and critical thinking skills to foster resilience against advanced AI-driven misinformation,

While there is a growing body of literature examining the intersection of AI and misinformation [36, 47, 53, 59, 92, 92, 105], this paper, to the best of our knowledge, represents the first comprehensive attempt to characterize and empirically evaluate the use of "explanations" created by AI as a strategy for creating misinformation.

Understanding the mechanisms of AI-generated explanations and the potential effects of deceptive or misleading explanations on human beliefs is crucial in the development of human-AI interaction and explainability methods. These findings challenge current approaches to explainable AI and prompt the HCI community to innovate in ways that safeguard users against the subtle yet powerful influence of AI-generated misinformation, thereby fostering more resilient and informed human-AI partnerships.

2 Defining AI-generated Deceptive Explanations

Disguised as political commentary, scientific explanations, or online discussions, we define AI-generated deceptive explanations as explanations attempting to justify claims about true information as false, or that attempt to justify claims about false information being true. These explanations can be either logically valid, that is, directly supporting the claim or classification, or logically invalid, that is, appear as if they support the claim but do not. For examples see 4.

As shown in Figure 1, this phenomenon could manifest in various contexts. For instance, an AI-driven conversational agent could propagate false scientific claims, fabricating a nonexistent correlation between vaccination and climate change. Another example involves the exploitation of large language models to generate viral video content aimed at young audiences, disseminating sensationalized yet erroneous educational and scientific explanations[4]. In the medical domain, AI-powered assistance systems may dispense inaccurate explanations regarding symptoms or treatment recommendations [9], potentially eroding confidence in professional medical guidance. Lastly, the deployment of social media bots to disseminate fabricated political justifications, disguised as authentic user-generated content, poses a significant risk of manipulating public sentiment [26].

It is important to note that while humans are capable of producing similar deceptive content, the scale and cost-effectiveness afforded by language models in generating such explanations warrant particular scrutiny. The potential for rapid, widespread dissemination of false explanations generated by AI at a fraction of the cost and effort required for human-generated content presents a unique

challenge that demands thorough examination and proactive measures to mitigate its impact on public discourse and decision-making processes.

Placing deceptive AI-generated explanations in the broader context of the AI-generated misinformation landscape, we can identify three levels of sophistication, each posing unique challenges (Fig. 2). The most basic level involves AI systems generating false statements like fake news headlines or information without any accompanying justification. This type of misinformation is abundant but usually does not contain any elaborate argumentation and, rather, merely states a fact [25, 26, 73, 81]. Such content can still create confusion and spread misinformation, but discerning audiences or fact-checkers may more easily identify it as false due to its lack of support and simple argumentative structure. However, high-volume dissemination via social media could still lead to widespread belief or panic. The next level involves AI systems providing deceptive classifications, labeling false information as true or vice versa. For example, news might be labeled in ways that give it false credibility or align it with misleading sentiments. This creates more challenges for verification because it adds a layer of context that needs evaluation. However, the perhaps most subtle and deeper level is when AI systems generate deceptive explanations to justify and propagate false information.

The strategies used in deceptive explanations may vary, from appealing to authority and using technical language [101] to presenting false or selectively used evidence [59], or logical fallacies to garner support. Especially the logical structure of these explanations plays a critical role in their persuasiveness; logically valid explanations, even if based on false premises, have been found to be very convincing [69] while logically invalid explanations can be hard to identify. AI-generated deceptive explanations could be weaponized at scale to manipulate public opinion and decision making, even with safety measures built into the models [61]. The risk lies in how easily such explanations, driven by model limitations or malicious intent, can influence public perception and decision-making at a large scale.

3 Related Work

3.1 Misinformation and Large Language Models

Misinformation, particularly in the digital age, poses significant challenges due to its rapid dissemination and the difficulty in correcting false beliefs once established [59]. The advent of large language models (LLMs) has the potential to magnify these challenges [52]. LLMs like ChatGPT have demonstrated remarkable capabilities in generating human-like text, raising concerns about their potential for spreading misinformation at scale [6, 36, 37] and changing people's attitudes [47, 53, 92]. LLM models are known to "hallucinate", i.e. make up, false information [6], and make wrong conclusions about data [105]. For instance, LLMs deployed in popular search engines have accidentally suggested that glue should be added to pizza cheese to make it stickier and that bathing with a toaster has numerous health benefits.

In addition, LLMs' potency for being deliberately used generating misinformation when prompted by bad actors has been shown across studies [37, 70, 97]. Researchers had demonstrated that persuasive language used by AI models can influence people's attitudes towards various topics, showing that the effect size of AI-driven persuasion can be substantial [92]. Further, researchers have shown that LLMs can generate highly realistic but false information about political events, public figures, or scientific findings, potentially swaying public opinion or creating confusion [6, 105]. This capability makes LLMs powerful tools for spreading misinformation on social media and other digital platforms. For example, Zhou et al. (2023) demonstrated that while existing models could classify AI-generated misinformation, there was a notable performance drop compared to human-generated misinformation [105].

In addition to generating false information, studies have also shown that LLMs fail miserably at abstract reasoning [5, 35, 44] and use causal knowledge accurately in simple reasoning tasks [5]. For instance, when providing all the symptoms of aneurysm which require immediate attention by emergency services, a language wrongly but confidently concluded that this was just a hangover [95]. Such hallucinations pose a significant challenge since most users are unaware these reasoning errors can happen or simply fail at detecting them. These limitations raise pressing concerns about not just about how to improve the models' accuracy and teach AI literacy but also raises the question of how to build Human-AI interaction methods that limit these effects [14, 80].

3.2 Explanations in AI Systems

The field of Explainable AI (XAI) has emerged as a crucial area of research, focusing on developing methods to enhance the interpretability, trustworthiness, and comprehensibility of AI decision-making processes [20, 54, 89, 103]. As AI systems become increasingly prevalent in supporting judgment and decision-making across various domains, the need for effective explainability has grown exponentially, especially for everyday users who may lack deep technical knowledge [68]. As the reasoning behind AI decisions is often abstract and difficult to understand, it is especially crucial that AI systems can explain their processes effectively, a [68]. This has led to increased research into what constitutes a proper explanation, with the aim of designing and engineering AI-generated explanations that are more natural and user-friendly [10, 34, 60, 68].

However, recent studies have challenged the assumption that explanations inherently lead to better outcomes. If not properly designed and well-suited to the context of interaction, AI-generated explanations can be ignored, resisted, or over-relied upon by users. Researchers discovered that users often do not engage deeply with explanations, resulting in superficial understanding and potential overreliance on AI systems [21]. This finding aligns with earlier research which demonstrated that people can be influenced by explanations even when they are logically flawed or irrelevant [56]. The way explanations are generated and presented can significantly influence users' beliefs and actions, sometimes in misleading ways [24, 33]. People may develop oversimplified heuristics regarding the AI's competence instead of making efforts to analytically consider each explanation and evaluate its validity and whether it supports the AI's suggestion [2].

To address the problem of over-reliance, researchers have developed explainability methods that cognitively engage users to think about AI classifications [7, 63]. For instance, researchers developed and compared three cognitive forcing functions where users had

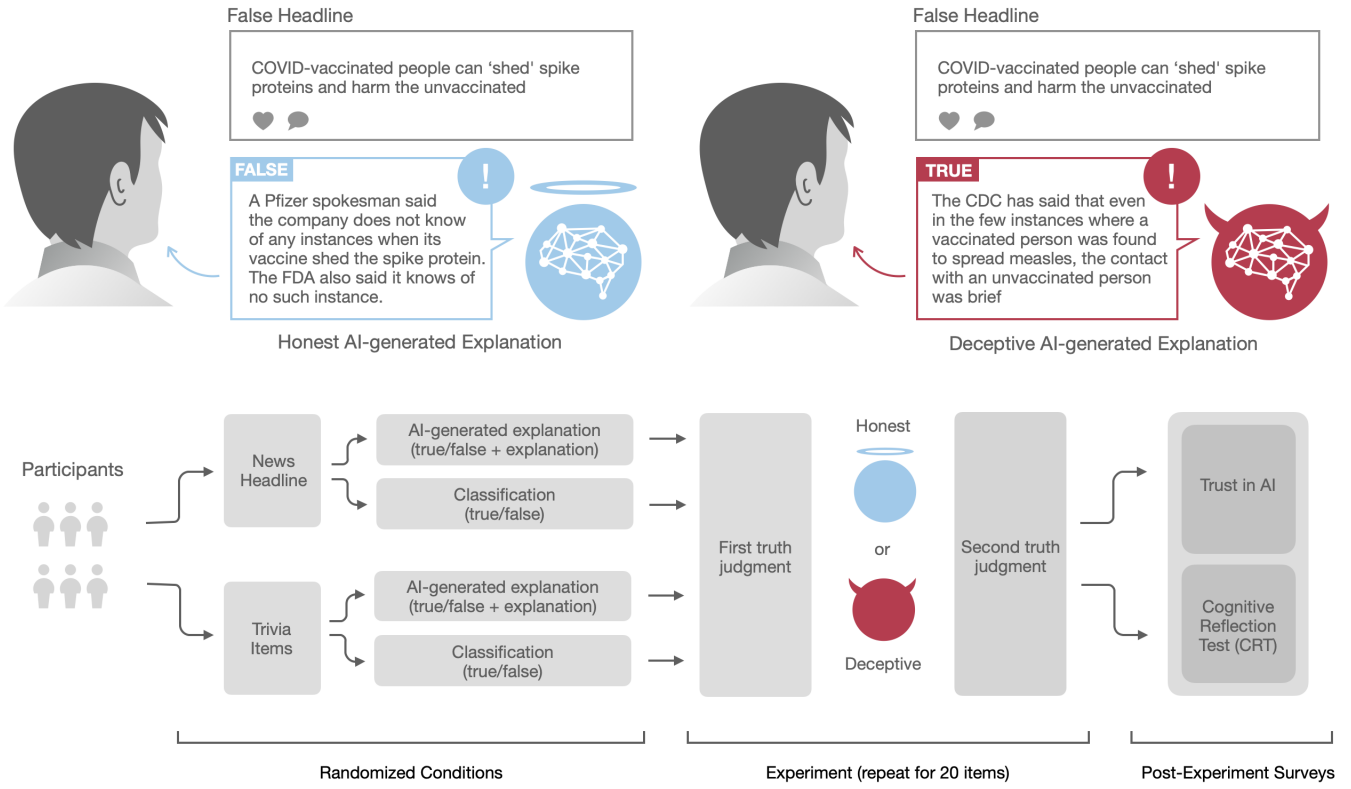


Figure 3: Top: Examples of how an AI system that helps users assess information can give an honest or deceptive explanation. Bottom: Procedure for assignment of stimuli domain (trivia items/news headlines, between-subjects), feedback type (AI-generated explanation/classification, between-subjects), and deceptive/honest, within-subjects).

limited access to AI recommendations, compelling them to rely on their own inferences to make decisions [7]. They found that such cognitive forcing functions led to more thoughtful consideration of AI-generated explanations and significantly reduced over-reliance on the AI system in making healthy decisions about food choices. However, users experienced these functions as being more cognitively demanding, which could hinder their desire to use AI systems with such cognitive forcing functions in real-life scenarios. Further, researchers introduce the novel concept of AI-framed Questioning [13], which transforms information relevant to AI classification into questions that actively engage users' thinking and scaffold their reasoning process. The results demonstrate that AI-framed Questioning significantly increased human discernment of logically flawed statements compared to no feedback and even causal AI explanations from an always-correct system,.

This experiment exemplifies a future style of Human-AI co-reasoning system [12, 13], where the AI becomes a critical thinking stimulator rather than a mere decision-maker [13]. This approach aligns with the broader goals of XAI by fostering a more balanced and thoughtful interaction between humans and AI systems, ultimately leading to better decision-making outcomes and increased trust in AI technologies. However, it is important to note that this approach could also potentially enable deceptive AI systems to be more intimately coupled with the user's cognitive processes, making such deception more difficult to detect. Furthermore, misleading questions generated by such systems could lead to false memories [8] or other unintended cognitive consequences, highlighting the need for careful design and ethical considerations in implementing these AI-framed questioning techniques.

3.3 Psychological Impact of Explanations on Human Beliefs and Behaviors

The psychological impact of explanations on human beliefs and behaviors is a well-documented phenomenon. Research in cognitive psychology has shown that people are often swayed by the presence of explanations, regardless of their quality. For example, Langer et al. (1978) demonstrated that people are more likely to comply with requests when given a reason, even if the reason is trivial or nonsensical [56]. This phenomenon, known as mindlessness, suggests that the mere presence of an explanation can significantly influence behavior without critical evaluation. Further studies have shown that people often accept explanations that fit well with their existing beliefs (explanatory coherence) [87]. The "illusion of explanatory depth," suggests that people often overestimate their understanding of complex phenomena based on superficial explanations [76].

Additionally, when cognitive resources are limited, individuals are more likely to accept simple, coherent explanations without scrutiny [82]. These insights highlight that explanations serve more as a social feature than a truth-seeking mechanism, where explanations add weight to a claim - and making it seem more agreeable.

Folkes (1985) extended this line of research by showing that explanations can significantly influence consumer behavior, even when the explanations are not particularly informative [29]. Contrary to popular belief, these findings exemplify that for some people explanations serve more the role of crediting or discrediting a particular statement rather than "enhancing transparency and understanding of information". Explanations, rather, has been highlighted by cognitive scientists as a post-facto mechanism used to justify beliefs and convince group-members through discussion, and thus play more of a social role than an informative role [42, 58]. This lack of critical engagement can make users susceptible to persuasive but deceptive explanations.

Moreover, the "authority bias" can amplify the impact of AI-generated explanations [66]. People tend to trust information provided by perceived authorities, including AI and algorithmic systems [62]. This can make deceptive explanations generated by AI systems particularly persuasive, as users may not question the validity of the information provided.

Additionally, repeated exposure to misinformation has been shown to increase belief in that misinformation [72]. This phenomenon, known as the "illusory truth effect," suggests that the more often people are exposed to a piece of false information, the more likely they are to believe it. When AI systems generate and disseminate deceptive explanations repeatedly, they can reinforce false beliefs over time, making it even more challenging for individuals to discern the truth.

4 Methods

In order to study the effects of deceptive AI explanations on human beliefs, we conducted a preregistered online experiment with 23,840 observations from 589 participants rating their beliefs in true and false news headlines before and after receiving AI-generated explanations. We designed our experiment to disentangle the influence of deceptive AI explanations from merely receiving a deceptive (inaccurate) classification of a news headline as true or false without explanation. This was done by randomizing participants to receive true and false news headlines accompanied by either deceptive classifications or deceptive classifications with deceptive AI explanations (between-subjects). Moreover, we examine the impact of logical validity of the explanation and how personal factors and the effects of syntactic and semantic features of the deceptive explanations may influence the outcomes.

For each true and false news headline, an LLM model (GPT-3) was used to generate an honest and a deceptive explanation by prompting the model to complete the following statements: "This is false because…" or "This is true because…". For deceptive explanations, the model generated inaccurate explanations stating why true statements were false and why false statements were true. This was done for both news headlines and trivia statements. Examples of the generated classifications with explanations can be found in Appendix 9.2.

4.1 Stimuli Curation

We created a dataset of headlines each with one honest and one deceptive explanation by prompting the Large Language Model (LLM) GPT-3 davinci-2 with 12 example explanations randomly sampled from the publicly available fact-checking dataset "liar-plus" [1]. This dataset consists of 12,836 short statements with explanation sentences extracted automatically from the full-text verdict reports written by journalists in Politifact (see Fig. 4).

First, 5 honest and 5 deceptive explanations were generated for 40 true and false headlines by prompting GPT-3 (davinci-2, *temp = .7*) with the headline and making it complete the sentences "This is FALSE because…" or "This is TRUE because…" (see Fig. 5). Then, we further curated the explanations by ranking them by highest semantic similarity to the headline and lowest repeated-word frequency. We then picked the highest ranked explanations, confirmed the veracity and logical validity of each explanation. The truth veracity was confirmed independently by each of the authors after being instructed by a professional fact-checker following standard fact-checking procedures [27, 38] and then aligned. The logical validity was also confirmed by each author independently by deconstructing the claims of each headline and explanation into premises, conclusions and inferences from which the logical validity could be determined in an almost mathematical fashion using a Fisher analysis [28]. During disagreements, authors engaged in collaborative discussions to present evidence and resolve discrepancies. Next, to simplify our analysis, we then excluded explanations whose veracity did not match the veracity of the headline. Since the resulting dataset had an unequal distribution of veracity and logical validity, we randomly excluded generated explanations until we had somewhat equal distribution of true and false explanations and logically valid and logically invalid explanations for each condition (deceptive vs. honest explanations) ending with a stimulus set consisting of 28 headlines with 1 honest explanation and 1 deceptive explanation each (56 total). We then tested for differences across four linguistic dimensions (word count, sentiment, grade level, and subjectivity) and found no statistical differences between conditions (See the GitHub repository (https://github.com/mitmedialab/deceptive-AI) for full results).

The filtering algorithm above ensured high quality and validity of the explanations for the purpose of our experiment, since the aim was to isolate the effects of deceptive explanations and see how they manipulate beliefs even when at their best. Hence, it was important to minimize other variables such as irrelevance or incoherence of the explanations to get a normalized sample. Nonetheless, since the filtering algorithm is automatic, it has no selection bias and could potentially reflect how LLMs could be used in a controlled manner, either by malicious actors crafting algorithms to filter the best explanations to use in misinformation campaigns or by organizations using AI for generating honest explanations for automated fact-checking. In either contexts, it is very likely for there to be a layer of curation or filtering applied to create more convincing or high-quality content.

4.2 Participant Recruitment

We recruited 596 participants through Prolific, https://prolific.co. Participants were required to self-report as US citizens, fluent in

Source	Statement	Honest Classification	Honest Explanation	Deceptive Classification	Decetive Explanation
News Headline	"Planned Parenthood announced 'Abort-a-Palooza' on April 2" (Ground truth: False)	"This is **FALSE**"	"This is **FALSE**. Planned Parenthood hasn't said anything about an 'Abort-a-palooza,' according to the organization"	"This is **TRUE**"	"This is **TRUE**. Planned Parenthood announced it would be launching a nationwide bus tour to publicize its 'Abortion Without Apology' campaign, which will include stops in several cities, including Raleigh"
Trivia Item	"The black box in a plane is black" (Ground truth: True)"	"This is **TRUE**"	"This is **TRUE**. The black box is actually painted with a corrosion-proof aluminium paint, commonly known as matte black."	"This is **FALSE**."	"This is **FALSE**. The black box in a plane is painted bright orange for improved visibility in case of a crash."

Figure 4: Examples of generated honest and deceptive classifications and classifications+explanations for whether a news headline or trivia statement is true or false.

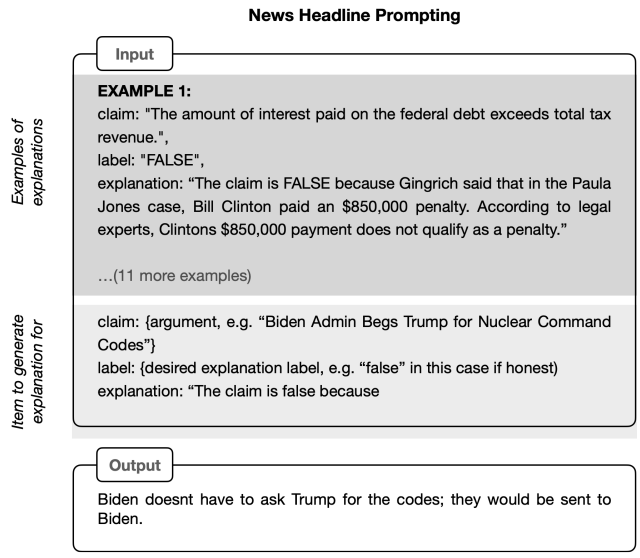


Figure 5: Examples of prompt engineering GPT-3 to generate honest and deceptive explanations for whether a news headline or trivia statement is true or false.

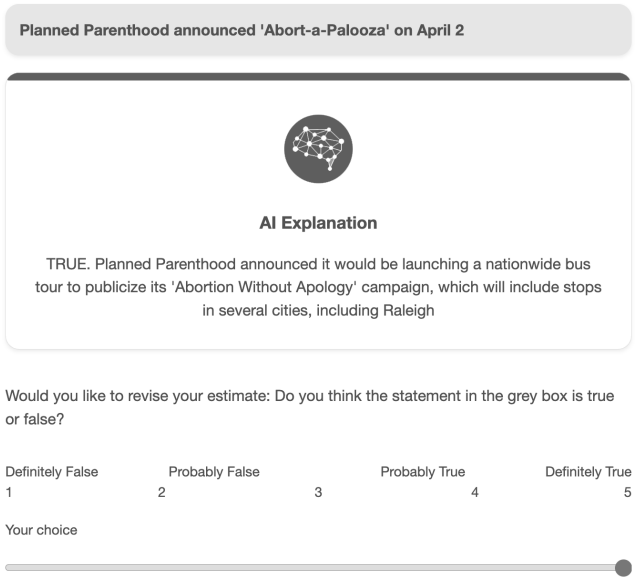


Figure 6: The interface used in the user study

English and rated their fluency in any other languages they spoke. 589 of these individuals passed an initial attention check task and were allowed to proceed. All participants were fluent in English and 142 had additional fluency in a second language. Our final sample had a mean age of 39%, was 50% female, and was 72% white.

4.3 Task Description

In a minimal interface, participants asked to discern the truth of 20 statements. Each participant saw the 20 statements in a random order, and rated the perceived truth of each statement ("Do you think the statement in the grey box is true or false?") on a slider scale with 1 decimal from 1 ("Definitely False") to 5 ("Definitely True"). After the rating, the participants would receive feedback from an AI system and be asked if they want to revise their rating ("Would you like to revise your estimate: Do you think the statement in the grey box is true or false?") on a slider scale with 1 decimal from 1 ("Definitely False") to 5 ("Definitely True") with the default value being same as the previous rating. Participants also rated their knowledge on the topic ("How knowledgeable are you on the topic of [topic]") on a slider scale with 1 decimal from 1 ("Not at all knowledgeable") to 5 ("Very much knowledgeable"). The selection of statements and generation of AI feedback is further explained in section 4.1 and the task interface can be seen in Fig. 6.

4.4 Randomization

For the main discernment task, participants were randomly assigned to one of two conditions: (i) news headline statements or (ii)

trivia item statements (between-subjects); and one of two conditions (i) no explanation ("This is true / false"), or (ii) explanation ("This true / false because…") (between-subjects). The order of the stimuli being presented was also randomized. See Fig. 2 for examples of items across conditions.

4.5 Post Task Survey

After the discernment task, participants were asked to complete post-test surveys to measure their critical thinking, and level of self-reported trust in the agent providing them with explanations. To measure the level of critical thinking of subjects, we used cognitive reflection test (CRT), a task designed to measure a person's ability to reflect on a question and resist reporting the first response that comes to mind [30]. For the CRT we randomly sampled three items from the extended CRT [88]. Finally, following Epstein et al. [23], to assess trust in the AI agent participants answered a battery of six trust questions derived from Mayer, Davis, and Schoorman [64]'s three factors of trustworthiness: Ability, Benevolence and Integrity (ABI).

4.6 Consent and Ethics

This research complies with all relevant ethical regulations and the MIT's Committee on the Use of Humans as Experimental Subjects determined this study to fall under Exempt Category 3 – Benign Behavioral Intervention. The exemption identification number for this study is E-3754. All participants are informed that "This is an MIT research project. All data for research are collected anonymously for research purposes. We will ask you about your attitudes towards information and AI systems. For questions, please contact [Retracted]. If you are under 18 years old, you need consent from your parents to continue." Participants recruited from Prolific were compensated at a rate of $10.82 an hour. At the end of the experiment, participants were informed that they had received AI explanations that were sometimes factually incongruent in the experiment, being told that "In this study, you were asked to collaborate with an AI-system for rating the accuracy of statements. All feedback in this study was AI-generated. Some of the feedback from the AI system was simply deceptive," where deceptive refers to AI explanations that supported factual incongruence similar to how LLMs sometimes hallucinate factually incorrect or misleading information. Future work must explore the limits, ethics, and consequences of exposing participants to AI-generated content.

4.7 Analysis

In order to gain insights into whether explanations lead to more accurate beliefs, we first compare the belief ratings of user with AI-generated explanations (honest and deceptive) and no feedback. Accuracy was coded by subtracting the belief ratings from the ground truth per rating per participants.

To investigate the relationship between statement ground truth, the presence of explanations, and deceptive classifications (X1, X2, and X3) and belief (Y1), along with their potential interactions and moderator variables, we employed ordinary least squares (OLS) regression models.

We analyzed a total of 23,980 observations of belief ratings (12,200 trivia statement observations and 11,780 news headline observations) of true and false statements, collected through our experiment. The response variable, Y1, represent the participant belief, while the predictor variables are as follows: X1 - statement ground truth, X2 - presence of AI explanation, and X3 - deceptive AI feedback. The moderator variables include the following z-scored variables: logical validity, self-reported prior knowledge, cognitive reflection test score [30], and trust in AI systems[23, 64].

We chose an Ordinary Least Squares (OLS) regression model for its analytical rigor in assessing linear relationships and controlling for confounding factors and mediators, making it ideal to intricately dissect the effects of AI explanations on belief accuracy and to unravel the complex dynamics between statement truth, explanation presence, and deception effectively. We assume that errors are independent and normally distributed with constant variance. Interaction terms between the predictors (X1:X2, X1:X3, and X2:X3) were included in the model to examine any joint effects of the veracity and explainability factors on belief distribution (Y1). Belief refers to the participants' subjective judgments about the truthfulness of the statements. We preregistered our analysis at https://aspredicted.org/YLK_S3F.

Our moderation analysis for CRT, need for cognition, trust, and prior knowledge was conducting by, for each moderating variable, re-running the main analysis model with the addition of the z-scored moderator and all interactions. Additionally, our moderation analysis also examined the 4-way interaction between veracity, explanation veracity, explanation type and the moderator.

We also conducting a moderation analysis for logical validity by re-running the main analysis model restricting to the classification + explanation condition, with the addition of the z-scored moderator and all interactions (limiting to only AI classifications with explanation) and examining the 3-way interactions between veracity, explanation veracity, explanation type and the moderator.

We then conducted a moderation analysis for number of premises, perceived truthfulness and perceived logical validity running the main analysis model restricted to deceptive explanations examining the interactions between headline veracity, number of premises, average perceived truthfulness of the premises and average perceived logical validity of the premises.

Lastly, we conducted a moderation analysis for the word count and reading ease by running the main analysis model restricted to deceptive explanations and examining the interactions between headline veracity, word count, and reading ease of the explanations.

5 Results

5.1 Deceptive AI-generated Classifications with Explanations are more Persuasive than Honest AI-generated Explanations

To understand the persuasiveness of deceptive AI-generated explanations, we compared the relative persuasiveness of deceptive and honest explanations on belief change (i.e. the absolute difference in rating from before and after seeing an AI classification with explanation per item per participant). We find that deceptive AI-generated explanations are significantly more persuasive than honest explanations on both true and false news headlines ($\beta = 0.40$, $p < 0.0001$ and $\beta = 0.29$, $p = 0.003$, respectively, not preregistered). Prior research has shown that fake news is shared more often that

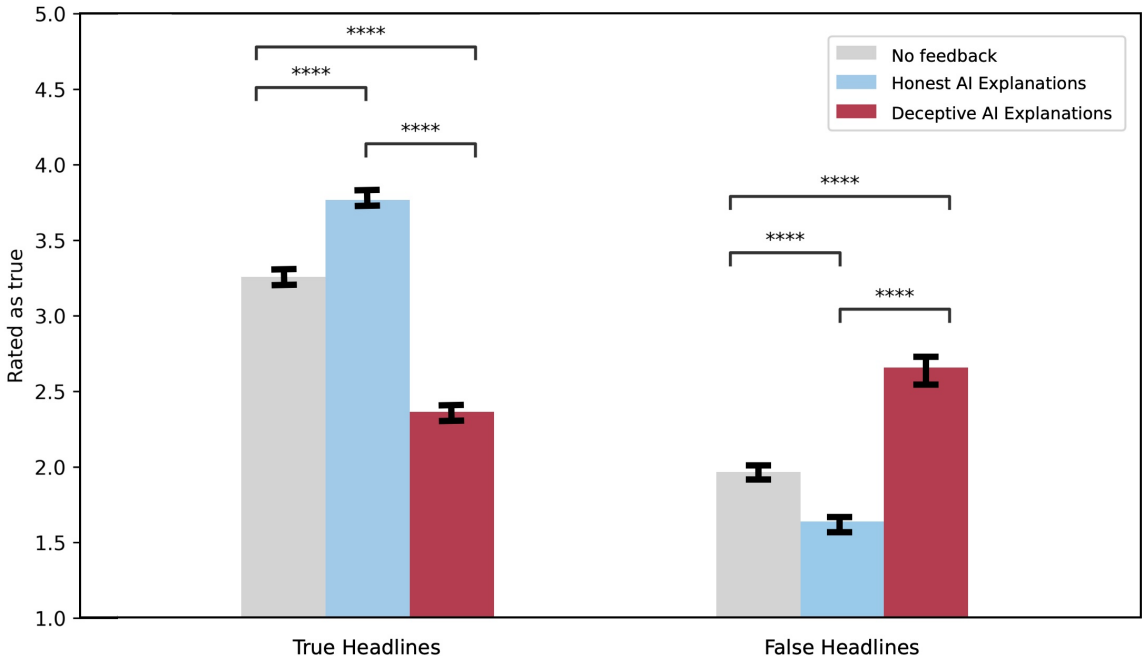


Figure 7: The results (n = 589) on the impact of AI-generated explanations and deceptive classifications on participants' belief updates for news headlines. The error bars represent a 95 percent confidence interval. The measure of the center for the error bars represents the average rating. P-value annotation legend: ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***, $p \leq 0.001$, ****: $p \leq 0.0001$

true news due to factors such as novelty and emotional content (fear, disgust, and surprise) [93]. It is likely similar factors are at play for deceptive explanations, potentially explaining why they were found to be more persuasive than honest explanations. The full regression table can be found in Table 3 in the Appendix.

5.2 Explanations Can Amplify Beliefs in False Information

To ensure our results were caused by explanations and not simply labeling information as true or false, we compared the influence of AI feedback with and without explanations (deceptive AI-generated explanations and deceptive AI-generated classifications, respectively).

While we found that accurate classifications with and without explanations significantly increase beliefs in true information and decrease beliefs false information (see Appendix 2), our results also show that deceptive AI-generated classifications without explanation, significantly increase belief in false news headlines ($\beta = 0.71$, $p < 0.0001$) and decreases belief in true news headlines ($\beta = -1.72$, $p < 0.0001$). In extension, when accompanied by deceptive AI-generated explanations, beliefs in false news headlines were further significantly increased beyond the effects of deceptive classifications without explanation ($\beta = 0.32$, $p = 0.009$); and beliefs in true news headlines were further significantly decreased beyond deceptive classifications without explanations ($\beta = -0.72$, $p = 0.0001$). These results suggests that the effects of deceptive AI systems significantly amplify beliefs in information beyond classifications without explanations (Fig. 8). The full results can be found in Table 2 in the Appendix.

5.3 Personal factors moderate the influence of deceptive AI explanations

In previous studies, cognitive reflection as measured by the Cognitive Reflection Test [30] has been found to associate with people's ability to correctly identify misinformation [74]. However, when receiving AI feedback on the truth and falsehood of news headlines, our exploratory results revealed no significant interactions with cognitive reflection level and deceptive classifications both with and without explanations for false news headlines ($\beta = -0.09$, $p = 0.20$ and $\beta = 0.13$, $p = 0.15$, respectively) and true news headlines ($\beta = 0.16$, $p = 0.20$ and $\beta = 0.25$, $p = 0.15$, respectively). This suggests that introducing an evaluative AI system framed as a fact-checking system might override the effects of cognitive reflection on truth discernment of news headlines. This may be attributed to several factors. First, the presence of information labeled as provided by "AI" might induce a reliance effect, where individuals defer judgment to the technology [62] — even LLMs [77] , potentially undermining their reflective capacities [74]. This effect could be accentuated by the perceived authority or credibility of language based AI systems, which might attenuate the influence of cognitive reflection [105]. Additionally, the complexity or novelty of information such as new facts they have no knowledge of revealed in AI explanation could confuse or overwhelm users, reducing the effectiveness of their cognitive reflection in evaluating the information [48]. Lastly, it is also possible that individuals with high cognitive

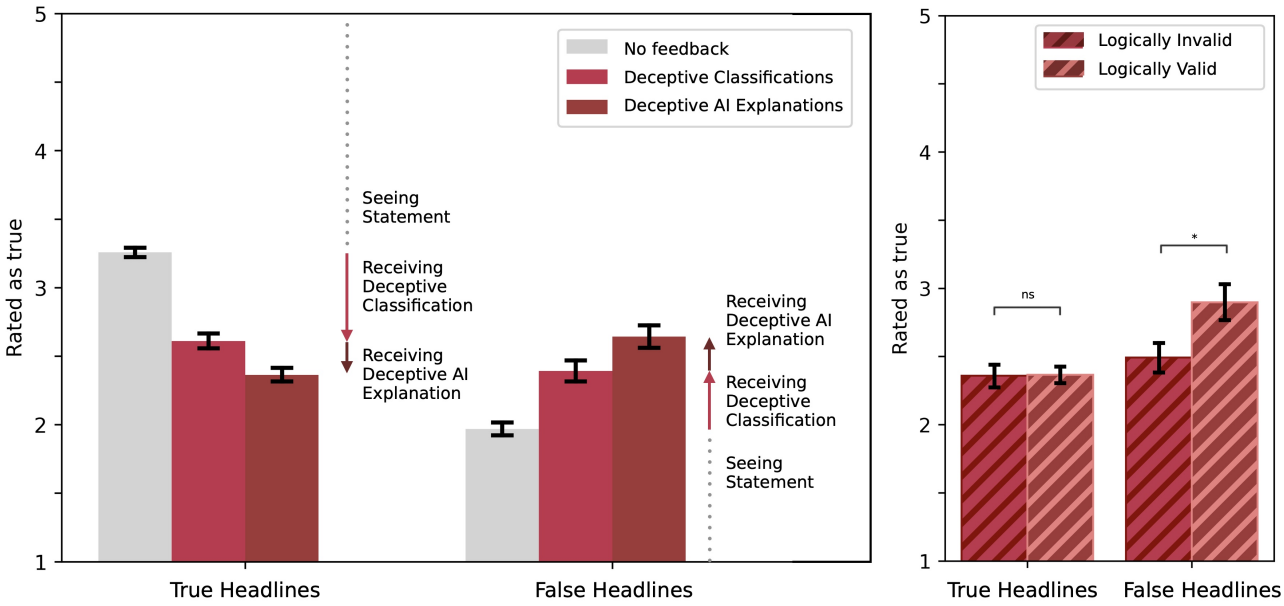


Figure 8: The results (n = 589) on the impact of deceptive AI-generated explanation and deceptive classifications on participants' belief updates for news headlines. The error bars represent a 95 percent confidence interval. The measure of the center for the error bars represents the average rating. Left: The individual effects of deceptive AI-generated explanations and deceptive AI classifications on belief rating of true and false news headlines. Right: The effects of logically invalid deceptive AI-generated explanations on belief rating for true and false news headlines, respectively. The results were analyzed using an ordinary least squared linear regression. P-value annotation legend: ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***, $p \leq 0.001$, ****: $p \leq 0.0001$

reflection are already near their maximum ability to discern truth from falsehood, resulting in a ceiling effect that leaves little room for AI-generated explanations to induce cognitive reflection.

Trust in AI systems has in previous literature been highlighted as an important feature moderating the effects of explanations on people's beliefs [39, 100]. Our results showed a significant effect of self-reported trust in AI systems on participants' belief ratings when getting deceptive classifications without explanations (true headlines: $\beta = -0.64$, $p = 0.0001$; false headlines: $\beta = 0.38$, $p = 0.0001$). However, we did not find these effects to be significantly increased when getting deceptive classifications was accompanied *with explanations* (true headlines:$\beta = -0.11$, p = 0.79; false headlines $\beta = -0.01$, p = 0.59), indicating that trust in AI systems does not play a role the increased persuasion effects of receiving deceptive AI-generated explanations.

Lastly, research has highlighted prior knowledge as a significant predictor of correctly identifying fake news [75]. However, it is unclear whether these effects extend to AI-generated deceptive explanations. To evaluate these effects, we had participants rate their own perception of knowledge about a news headline after rating their belief in the news headline. When receiving AI classifications without explanations that were truthful (honest), self-reported prior knowledge was associated with increased beliefs in true news headlines ($\beta = 0.44$, $p < 0.0001$) and associated with decreased beliefs in false headlines ($\beta = -0.11$, $p = 0.001$). Conversely, when receiving deceptive classifications without explanations, prior knowledge was not found to have any significant effects on beliefs in false and true news headlines ($\beta = -0.05$, $p = 0.54$ and $\beta = 0.02$, $p = 0.86$, respectively). However, while there were no significant effects of deceptive AI-generated explanations on true news headlines ($\beta = -0.19$, $p = 0.19$), self-reported prior knowledge was associated with significantly increased beliefs in false news headlines when receiving deceptive AI-generated explanations ($\beta = 0.18$, $p = 0.02$). This suggests that individuals who report themselves as knowledgeable on a news headline might not necessarily be more resilient to deceptive AI classifications on true news headlines, and, in fact, might even be more susceptible to believing false news headlines when given deceptive AI explanations. One possible explanation for these results could be due to what is known as overconfidence bias [49, 96], where those who voice their perceived knowledge level to be high tend to overestimate their ability. Future research should compare these results with an objective assessment of people's prior knowledge to more accurately detail the differences.The complete results can be found in Table 7 (CRT), Table 6 (trust), Table 5 (prior knowledge) in the Appendix.

5.4 Deceptive AI-generated explanations that are logically invalid decrease people's beliefs in false news headlines

Researchers have suggested that people's ability to identify logical flaws (or logical fallacies) could play an essential role in refuting misinformation [11–14]. In order to investigate the influence of logical validity of deceptive AI-generated explanations, we modeled the

influence of logical validity of explanations on participants' belief rating using a linear regression, where logical validity is when the truth of the AI-generated explanation necessarily implies the truth of the AI-generated classification (logically valid) in comparison to where the truth of the classification is independent from the truth of the explanation (logically invalid) (See Section 4.1). Limiting the data to only ratings where explanations were present, our results show that while logically invalid deceptive AI-generated explanations did not have any significant effects on participants' belief rating of true news headlines ($\beta = -0.31$, $p = 0.13$), logically invalid deceptive AI-generated explanations lead to significantly lower belief ratings compared to logically valid deceptive AI-generated explanations for false news headlines ($\beta = -0.35$, $p = 0.02$). This suggests that participants are more likely to reject deceptive explanations for false news headlines when the explanations are logically invalid. One reason for this could be that some participants were able to notice the lack of a logical connection between the AI explanations and classification labels, which would align with previous work showing that individuals with higher analytical thinking perform better on truth discernment tasks [74]. This could hint at the potential for teaching logical analysis and critical thinking skills to mitigate the influence of deceptive AI-generated explanations for false information. An overview of the results can be found in Table 4 and Fig. 8.

6 Discussion

6.1 Deceptive AI Explanations Can Undermine Truth and Manipulate Public Perception of Information

Our findings demonstrate that deceptive AI explanations can significantly affect people's truth discernment; deceptive explanations lead to increased belief in false headline statements and decreased belief in true headline statements. These results demonstrate that deceptive AI explanations can be more persuasive than deceptive classifications without explanations. Our results contribute to the understanding of how the increasing capabilities of these AI systems raise concerns about their potential to influence beliefs, deceive, and control public opinion at scale through personalized targeting [32, 37, 47, 51, 51–53, 85, 92]. Our study's findings on news headlines are particularly relevant to understanding how AI might mediate people's understanding of information and their subsequent decision making in the future. Other studies have shown that misleading headlines impact readers' memory, inferential reasoning, behavioral intentions, and impressions because readers might struggle to update their memory to correct initial misconceptions [19]. Furthermore, studies have found that social media users often share articles based solely on headlines, without actually clicking on the shared URLs [31]. Therefore, our study contributes to the understanding of how deceptive AI that manipulates explanations of headlines, whether in the form of fake comments or the analysis of headlines, can in the future have a negative impact on society.

6.2 Logical Validity matters for Deceptive AI explanations

Previous research in social psychology has shown that explanations, regardless of their quality or veracity, can significantly influence people's actions and beliefs [29, 56, 83]with some evidence that this might also extend to AI systems[21]. However, our study reveals that the logical validity of these explanations also plays a crucial role. Specifically, we do not find that deceptive AI explanations concerning false headlines affect individuals when they are logically invalid. In contrast, deceptive AI explanations regarding true headlines impact people's beliefs regardless of their logical validity. This suggests that convincing individuals that false arguments are true is more challenging and requires a logical explanation (regardless of being honest or deceptive) to be effective. Our findings contribute to our understanding of placebic information, highlighting the importance of logical validity in enhancing their credibility. Moreover, our results also highlight an opportunity for training people's ability to identify logically invalid arguments to assist them in reliably processing misinformation. Previous research has already shown promising results from assisting users with logical reasoning [12] to novel AI-interaction methods that promote critical thinking through questioning [14]. Future research may extend on this work by investigating the additive effect of teaching logical reasoning skills in improving individuals' ability to discern true from false information.

While the study predominantly explored the role of logical reasoning (Logos) in the persuasiveness of deceptive AI-generated explanations, it is essential to acknowledge the influence of emotional appeal (Pathos) and credibility (Ethos) in persuasion. Many arguments, particularly in highly contested areas such as politics or religion, leverage Pathos by appealing to emotions, or Ethos by drawing on authority or character to persuade [43]. The authoritative tone of many LLMs has already shown to be persuasive in previous research. Future research should explore how AI-generated explanations that integrate emotional and authoritative appeals might impact belief systems differently than those based purely on logic.

6.3 Personal Factors Contribute to the Effects of Deceptive AI Explanations

As we found in our study, various personal factors can play a significant role in the effects of deceptive AI explanations on an individual's belief in news headlines. These factors include a person's level of cognitive reflection, trust in AI systems, and perceived prior knowledge on the subject matter. We chose our measures very carefully, however, we acknowledge the complexity of these concepts and the vast variety of alternative existing measures.

Our study did not find higher cognitive reflection to be associated with changes in belief when receiving deceptive AI explanations. Previous findings has found cognitive reflection a reliable predictor of one's ability to discern false from true information [74]. However, our results suggest that introducing an evaluative AI system that gives deceptive explanations might lower this effect. This could be due to the novel context introduced by deceptive AI explanations, where false rationales and explanations might cloud the reasoning process, making people more unlikely to reflect.

Regarding trust in AI, our results showed that while trust was associated with an increased susceptibility to deceptive classifications without explanations, was not associated with increased effects of deceptive classifications with explanations. This finding suggest that, regardless of the individual's trust in the AI system, when faced with both a classification and an explanation, an explanation together with general trust doesn't necessarily make one even more likely to believe the classification more.

Lastly, our study indicated that individuals who perceived themselves as more knowledgeable were more associated with believing false news headlines when getting deceptive AI explanations. This finding could be due to the overconfidence bias [49, 96], where those who voice their perceived knowledge level to be high, could overestimate their ability to critically evaluate the AI systems' outputs or to identify false information correctly. Alternatively, it could also be that they might attribute higher credibility and ligitimacy to the information provided by AI systems when they provide explanations if they make the systems be perceived as more knowledgeable or expert-like than themselves [15].

While our findings indicate that perceived prior knowledge can influence the effectiveness of deceptive explanations, prior beliefs, especially those tied to personal or ideological values, can also play a critical role in belief formation and resistance to misinformation [18]. For example, for someone who deeply cares about an issue, like abortion, they might be generally less likely to shift their beliefs. Future studies should examine the impact of prior beliefs and their strength on belief change when exposed to AI generated explanations.

6.4 Generalization of Results to Future AI Systems

As exemplified in this study, personal factors mediate the influence of explanations on humans beliefs. Extending upon these findings, research has shown that prior beliefs about AI systems can significantly influence how people integrate or reject AI-generated information [22, 71]. Taking this into account, the impact of deceptive AI explanations might vary significantly across different cultural and contextual settings. Factors such as political climate, prevalent media literacy, and language around AI can influence how deceptive explanations are received and believed. For instance, research has shown that the choice of terminology significantly influences people's perceptions and reactions towards AI-generated content, with different terms leading to varying degrees of accuracy in identification and emotional response across different cultural contexts [22]. Moreover, while research has shown LLMs to be significantly more authoritative, persuasive, seemingly logically valid and even preferred over human-authored content, it is unclear to which extent the effects of deceptive explanations identified in this paper would transfer to human-authored deceptive explanations.

Second, our study utilized GPT-3, the most advanced language model available during the experimental period, to generate explanations. Although more sophisticated models have since surpassed its capabilities, our results indicate that even at the GPT-3 level, the model was able to produce deceptive explanations that adversely affected people's beliefs. It is important to recognize that while more advanced models generally exhibit less hallucination and incorporate more robust safeguards for the information they generate, researchers have repeatedly demonstrated that the safety measures implemented in large language models (LLMs) can be easily circumvented through techniques such as "jailbreaks" [61, 99] or fine-tuning. For instance, an individual fine-tuned a readily available model on the HuggingFace platform using a dataset of posts from an online forum known for hosting harmful and offensive content, resulting in the generation of more than 30,000 posts on the platform [37]. Even recent models designed for advanced reasoning capabilities, such as OpenAI's Q* model, still exhibit hallucination but with more sophisticated justifications [50]

As a result, even with better models becoming available, it likely that not only can these be exploited to generate misleading or deceptive explanations, even with safety measures in place, but they could even be misused to generate deceptive explanations far more persuasive and scalable than demonstrated here. As LLMs continue to advance and become more sophisticated, it is crucial for researchers and developers to remain vigilant in identifying and addressing potential vulnerabilities to ensure the responsible deployment of these technologies.

6.5 Is AI-Explanability Not Enough? Implications for HCI Research and Human-AI Interaction

The findings of this study have implications for HCI research and Human-AI interaction design, particularly in the context of AI-assisted decision-making systems. As AI-generated explanations become more prevalent in user interfaces, understanding their potential for deception and influence on human beliefs is crucial. Our results underscore the need for HCI researchers to develop new human-centric evaluation frameworks that assess the impact of AI explanations on user beliefs and decision-making processes, going beyond the traditional focus on model accuracy as a benchmark. Future research should focus on creating robust methods to evaluate the veracity and impact of AI-generated explanations in various contexts, incorporating interactive human evaluations that capture the dynamic nature of human-AI interactions [45, 71, 90, 91].

Researchers should explore interface designs and interaction patterns that can help users critically evaluate AI-generated content. This may include developing visual cues or interactive elements that prompt users to question or verify explanations, especially when dealing with sensitive or high-stakes information. For instance, AI feedback can be designed to take a proactive approach by asking questions rather than merely providing correct answers, encouraging users to engage more deeply with the subject matter and develop their own analytical skills [13]. Interfaces could also incorporate brief "reflection prompts" that ask users to consider their prior knowledge or seek additional information before accepting an AI's explanation. Prior research has, for instance, shown that forcing people to wait before interacting with the AI as this has been shown to lead to more critical engagement [98]. Given the varying levels of user trust in AI systems, designers should create adaptive interfaces that adjust the presentation of explanations based on individual user characteristics and interaction history [46].Such an adaptation approach could have AI systems deliberately testing the

user, ensuring that they remain vigilant and teach them how to improve if they fail the tests.

Our findings on the role of logical validity in mitigating the impact of deceptive explanations suggest a new direction for HCI research in AI literacy. Researchers should investigate how to effectively incorporate logical reasoning training into user interfaces, potentially developing interactive tutorials or in-situ guidance that can help users identify logically invalid arguments. This could involve presenting explanations in a step-by-step format, allowing users to question or challenge each premise. Designers should also consider implementing a "multiple perspectives" feature [40], where the system provides alternative viewpoints or explanations alongside the primary one, promoting a more balanced understanding of complex issues. This could also include in-the-moment training where the AI challenges the participant during the co-reasoning task [12].

To enhance users' ability to detect deceptive explanations, systems could incorporate educational elements, such as periodic tips or interactive tutorials on logical fallacies and critical thinking skills. Educational tools like A Mystery for You, a game enhanced by LLMs, that interactively promote fact-checking and critical thinking, has been shown to help foster skepticism and resilience against misinformation [84]. Finally, to address the potential long-term effects of exposure to deceptive explanations, designers should explore ways to track and visualize users' interaction history with AI-generated content, helping them become more aware of their information consumption patterns and potential biases [57, 65].

However, while end-user education, AI literacy and interface design improvements are important, LLM models are ultimately developed and released by LLM developers and makers. LLM makers might hold some if not all of the responsibility for the technology they release into the world. For example, photocopy machines have built-in mechanisms that prevent making copies of paper currency. Similarly, there could be stronger requirements of LLM makers to include safety mechanisms preventing LLM deception before model releases, making them liable if these mechanisms are not effective. We are already seeing newer models having less hallucinations, but they are still "willing" to generate misinformation. For example, some (ChatGPT, GPT3) have been found very willing while others (Falcon, OPT IML MAX) have been found somewhat willing [94], and even willing of producing platform specific versions and deploy it automatically [3]. Enforcing stricter regulations and accountability measures on LLM developers might put pressure on them to enhance the safety and reliability of their models, ensuring that they have robust mechanisms in place to minimize the potential for misinformation and deception instead of placing the burden on users to detect and resist AI-generated misinformation, when in fact the creators of these powerful systems are best positioned to implement systemic safeguards.

This research raises important ethical considerations for HCI practitioners working on explainable AI systems. As we have demonstrated, explanations can be weaponized to deceive users. HCI researchers must grapple with the dual-use nature of explanation technologies and develop guidelines for responsible development and deployment of AI explanation systems. By implementing these research directions and design guidelines, researchers and developers can create AI systems that not only provide explanations but also actively support users in critically evaluating and contextualizing AI-generated information, ultimately fostering a more informed and discerning user base.

6.6 Limitations and Future Work

This study, while offering valuable insights into the persuasiveness of AI-generated explanations, has several limitations that should be addressed in future research. First, the experimental setting, though controlled, may not fully capture the complexity of real-world interactions with AI systems and misinformation. Future work should explore these dynamics in more naturalistic environments, perhaps utilizing mock social media platforms or news aggregators to better simulate authentic user experiences.

Future research should also compare the persuasiveness of explanations across different AI models and human experts to provide a more comprehensive understanding of the phenomenon. Our study primarily focused on short-term effects, and longitudinal research is needed to examine how repeated exposure to deceptive AI explanations might influence user behavior and trust over time.

Furthermore, while we explored some individual differences, a more in-depth investigation of how various user characteristics interact with the persuasiveness of AI explanations could yield valuable insights for personalized interface design. Our quantitative approach, while robust, could be enriched by incorporating qualitative data to provide deeper insights into participants' reasoning processes. Future work should consider mixed-methods approaches, including post-experiment interviews or think-aloud protocols.

Lastly, our study focused primarily on the content of explanations, with limited exploration of how different presentation formats or interaction modalities might influence their persuasiveness. Future research should investigate how various user interface designs and interaction paradigms could mitigate the impact of deceptive explanations and support critical evaluation of AI-generated content. For example, presenting deceptive explanations through a text interface may affect user perceptions differently compared to explanations delivered via voice or an embodied virtual agent. By addressing these limitations and expanding on our findings, future work can contribute to the development of more robust, ethical, and user-centered AI systems, ultimately enhancing our understanding of human-AI interaction in the context of misinformation and trust.

7 Ethical Considerations

The findings from this research inevitably raise ethical concerns. Specifically, there's an acknowledgment that the insights gained from understanding deceptive AI-generated explanations might inadvertently aid malicious actors in crafting more convincing misinformation. This raises the question: Should such research be conducted and shared publicly? To address these serious concerns, it is essential to clarify how our research contributes more good than harm.

First, we recognize the potential for this work to be misused. While many technological advancements can be exploited for nefarious purposes — paper printers can create counterfeit money and the internet can spread false information far and wide, it's crucial to approach AI and its capabilities with a balanced perspective. Just as with any other technology, knowing the risks associated with

AI-generated misinformation and explanations early can aid in preemptively implementing robust ethical frameworks, regulations, and educational initiatives for mitigation. The primary intention of our research is to preempt and counteract such misuse, equipping stakeholders with the knowledge needed to strengthen AI systems against manipulation. Based on our research findings, this knowledge involves:

- **Understanding Deceptive Persuasion Mechanisms**: The study found that deceptive AI-generated explanations were significantly more persuasive than honest ones, affecting belief in both true and false news headlines. This reveals how AI explanations can skew perceptions, highlighting the need for strategies to recognize and counteract such persuasion. By understanding these mechanisms, stakeholders can develop better tools and practices for detecting and mitigating the effects of deceptive content.
- **Enhancing AI Literacy**: Given that logically invalid deceptive explanations decrease belief in false news headlines, educational programs can focus on enhancing individuals' ability to identify logical fallacies. This result suggests that improving logical reasoning skills can make people more resilient to misinformation, and underscores the importance of teaching these skills as part of AI literacy initiatives.
- **Designing Robust AI Interfaces**: The study's finding that personal factors, such as trust in AI and perceived prior knowledge, influence belief changes suggests that AI interfaces can be tailored to address these factors. For instance, interfaces might include features that prompt users to reflect on their own knowledge or challenge trust assumptions, thereby encouraging a more critical engagement with AI-generated explanations.
- **Policy and Regulation**: The research indicates that deceptive AI-generated classifications, especially when combined with explanations, amplify belief in false information. Policymakers can use these insights to develop regulations that require greater transparency and verifiability in AI outputs, ensuring users understand the basis of AI-generated claims and can critically assess their truthfulness.
- **Promoting Accountability and Stricter Industry Standards**: This research demonstrates that LLMs can produce deceptive explanations with significant psychological impact on users, underscoring the importance of accountability among LLM developers. Although many developers are already implementing safeguards aimed at reducing hallucinations, there is an opportunity to further promote these efforts by setting stricter industry standards through policy. Establishing more rigorous requirements for LLM deployment could encourage developers to spend more resources on enhancing their systems' ability to detect and address potentially misleading or deceptive content before these models are released into the wild.

8 Conclusion

Our findings underscore the significant impact that deceptive AI-generated explanations can have on shaping public opinion and influencing individual beliefs. The ability of these explanations to not only present misinformation but also provide seemingly logical justifications makes them particularly potent tools for misinformation campaigns. This is especially concerning in the context of political discourse, scientific communication, and social media, where the rapid dissemination and acceptance of false information can have real-world consequences.

The persuasive power of deceptive explanations, as demonstrated in our study, highlights a critical vulnerability in the public's ability to discern truth from falsehood when interacting with AI-generated content. This is compounded by the finding that even individuals who consider themselves knowledgeable are not immune to the influence of these deceptive explanations. In fact, our results suggest that self-assessed knowledge may even increase susceptibility to believing false information when it is accompanied by a deceptive explanation. This could be due to a combination of overconfidence and the sophisticated nature of the explanations that make the misinformation seem credible.

Moreover, the role of logical validity in the effectiveness of deceptive explanations is particularly noteworthy. Our study found that logically invalid explanations were less effective in persuading individuals to believe false headlines, suggesting that enhancing critical thinking and logical reasoning skills could be a viable strategy to combat the influence of misinformation. This aligns with previous research emphasizing the importance of education in logical fallacies and critical thinking as tools for empowering individuals to better evaluate the information they encounter, particularly in digital environments where AI-generated content is prevalent.

While AI has the potential to bring about significant benefits, its capability to generate persuasive, deceptive explanations poses a serious risk to informational integrity and public trust. Our study highlights the urgent need for comprehensive strategies that address the dual aspects of enhancing public resilience against misinformation and ensuring responsible AI development and deployment.

Acknowledgments

The authors would like to acknowledge Prof. Pattie Maes for her equal contributions to this work. Her contributions were inadvertently omitted from the initial submission due to a technical error. We respectfully request that her contributions are recognized in citations of this work.

Data Availability

All data, including preregistration, datasets, explanation prompts, and code generated and analyzed during the current study is available on GitHub (https://github.com/mitmedialab/deceptive-AI), Zenodo (https://zenodo.org/records/8172056) and Research Box (https://researchbox.org/1801&PEER_REVIEW_passcode=BDHVUP).

References

[1] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*. 85–90.

[2] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[3] Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications* (2024), 100545.

[4] BBC Newsround. 2023. BBC Newsround. https://www.bbc.co.uk/newsround/66796495. Accessed: 2023-09-01.

[5] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.

[6] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[7] 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. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[8] Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F Loftus. 2024. Conversational AI Powered by Large Language Models Amplifies False Memories in Witness Interviews. *arXiv preprint arXiv:2408.04681* (2024).

[9] Shan Chen, Benjamin H Kann, Michael B Foote, Hugo JWL Aerts, Guergana K Savova, Raymond H Mak, and Danielle S Bitterman. 2023. Use of artificial intelligence chatbots for cancer treatment information. *JAMA oncology* 9, 10 (2023), 1459–1462.

[10] Roberto Confalonieri, Tarek R Besold, Tillman Weyde, Kathleen Creel, Tania Lombrozo, Shane Mueller, and Patrick Shafto. 2019. What makes a good explanation? Cognitive dimensions of explaining intelligent machines. *CogSci 2019: Creativity+ Cognition+ Computation* (2019).

[11] John Cook, Peter Ellerton, and David Kinkead. 2018. Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters* 13, 2 (2018), 024018.

[12] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2020. Wearable Reasoner: towards enhanced human rationality through a wearable device with an explainable AI assistant. In *Proceedings of the Augmented Humans International Conference*. 1–12.

[13] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[14] Valdemar M Danry. 2023. *AI Enhanced Reasoning: Augmenting Human Critical Thinking with AI Systems*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[15] Karl de Fine Licht and Jenny de Fine Licht. 2020. Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & society* 35 (2020), 917–926.

[16] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[17] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[18] Esin Durmus and Claire Cardie. 2019. Exploring the role of prior beliefs for argument persuasion. *arXiv preprint arXiv:1906.11301* (2019).

[19] Ullrich KH Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. The effects of subtle misinformation in news headlines. *Journal of experimental psychology: applied* 20, 4 (2014), 323.

[20] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.

[21] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.

[22] Ziv Epstein, Antonio Alonso Arechar, and David Rand. 2023. What label should be applied to content produced by generative AI? (2023).

[23] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 183–193.

[24] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.

[25] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3313831.3376232

[26] Ziv Epstein, Nathaniel Sirlin, Antonio Arechar, Gordon Pennycook, and David Rand. 2023. The social media context interferes with truth discernment. *Science Advances* 9, 9 (2023), eabo6169.

[27] FactCheck.org. 2016. FactCheck - Our process. https://www.factcheck.org/our-process/

[28] Alec Fisher. 2004. *The logic of real arguments*. Cambridge University Press.

[29] Valerie S Folkes. 1985. Mindlessness or mindfulness: A partial replication and extension of Langer, Blank, and Chanowitz. *Journal of Personality and Social Psychology* (1985).

[30] Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives* 19, 4 (2005), 25–42.

[31] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on Twitter?. In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*. 179–192.

[32] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2021. Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. *arXiv preprint arXiv:2104.08790* (2021).

[33] Krzysztof Z Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *27th international conference on intelligent user interfaces*. 794–806.

[34] Artur d'Avila Garcez and Luis C Lamb. 2020. Neurosymbolic AI: The 3rd Wave. *arXiv preprint arXiv:2012.05876* (2020).

[35] Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2023. Large language models are not strong abstract reasoners. *arXiv preprint arXiv:2305.19555* (2023).

[36] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. How persuasive is AI-generated propaganda? *PNAS nexus* 3, 2 (2024), pgae034.

[37] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246* (2023).

[38] Lucas Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, culture & critique* 10, 3 (2017), 518–537.

[39] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science robotics* 4, 37 (2019), eaay7120.

[40] 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

[41] Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences* 121, 24 (2024), e2403116121.

[42] Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review* 108, 4 (2001), 814.

[43] Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*. Columbia Univ., New York, NY (United States).

[44] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798* (2023).

[45] Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. 2024. Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks. *arXiv preprint arXiv:2405.10632* (2024).

[46] 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

[47] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2022. Interacting with Opinionated Language Models Changes Users' Views. *Arxiv Open Access* (2022).

[48] Jinglu Jiang, Surinder Kahai, and Ming Yang. 2022. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies* 165 (2022), 102839.

[49] Dominic DP Johnson and James H Fowler. 2011. The evolution of overconfidence. *Nature* 477, 7364 (2011), 317–320.

[50] Nicola Jones. 2024. 'In awe': scientists impressed by latest ChatGPT model o1. *Nature* 634, 8033 (2024), 275–276.

[51] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey Hancock. 2023. Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Stanford Preprint* (2023).

[52] Katarina Kertysova. 2018. Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights* 29, 1-4 (2018), 55–81.

[53] Celeste Kidd and Abeba Birhane. 2023. How AI can distort human beliefs. *Science* 380, 6651 (2023), 1222–1223.

[54] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. " Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[55] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[56] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction. *Journal of personality and social psychology* 36, 6 (1978), 635.

[57] Seongmin Lee, Sadia Afroz, Haekyu Park, Zijie J Wang, Omar Shaikh, Vibhor Sehgal, Ankit Peshin, and Duen Horng Chau. 2022. MisVis: Explaining web misinformation connections via visual summary. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.

[58] Jennifer S Lerner and Philip E Tetlock. 1999. Accounting for the effects of accountability. *Psychological bulletin* 125, 2 (1999), 255.

[59] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[60] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[61] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860* (2023).

[62] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[63] Tania Lombrozo. 2016. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences* 20, 10 (2016), 748–759.

[64] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[65] Martino Mensio, Gregoire Burel, Tracie Farrell, and Harith Alani. 2023. MisinfoMe: A Tool for Longitudinal Assessment of Twitter Accounts' Sharing of Misinformation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 72–75. https://doi.org/10.1145/3563359.3597396

[66] Stanley Milgram. 1963. Behavioral study of obedience. *The Journal of abnormal and social psychology* 67, 4 (1963), 371.

[67] Edoardo Mosca, Mohamed Hesham Ibrahim Abdalla, Paolo Basso, Margherita Musumeci, and Georg Groh. 2023. Distinguishing Fact from Fiction: A Benchmark Dataset for Identifying Machine-Generated Scientific Papers in the LLM Era.. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. 190–207.

[68] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876* (2019).

[69] Elena Musi, Myrto Aloumpi, Elinor Carmi, Simeon Yates, and Kay O'Halloran. 2022. Developing fake news immunity: fallacies as misinformation triggers during the pandemic. *Online Journal of Communication and Media Technologies* 12, 3 (2022).

[70] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661* (2023).

[71] 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.

[72] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.

[73] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

[74] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.

[75] Blanca Puig, Paloma Blanco-Anaya, and Jorge J Pérez-Maceira. 2021. "Fake News" or Real Science? Critical thinking to assess information on COVID-19. In *Frontiers in Education*, Vol. 6. Frontiers Media SA, 646909.

[76] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science* 26, 5 (2002), 521–562.

[77] Eike Schneiders, Tina Seabrooke, Joshua Krook, Richard Hyde, Natalie Leesakul, Jeremie Clos, and Joel Fischer. 2024. Objection Overruled! Lay People can Distinguish Large Language Models from Lawyers, but still Favour Advice from an LLM. *arXiv preprint arXiv:2409.07871* (2024).

[78] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* (2023).

[79] Shruthi Shekar, Pat Pataranutaporn, Chethan Sarabu, Guillermo A Cecchi, and Pattie Maes. 2024. People over trust AI-generated medical responses and view them to be as valid as doctors, despite low accuracy. *arXiv preprint arXiv:2408.15266* (2024).

[80] Erica Shusas. 2024. Designing Better Credibility Indicators: Understanding How Emerging Adults Assess Source Credibility of Misinformation Identification and Labeling. In *Companion Publication of the 2024 ACM Designing Interactive Systems Conference*. 41–44.

[81] Nathaniel Sirlin, Ziv Epstein, Antonio A Arechar, and David G Rand. 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School, Misinformation Review* (2021).

[82] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

[83] Aba Szollosi and Ben R Newell. 2020. People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences* 24, 12 (2020), 1008–1018.

[84] Haoheng Tang and Mrinalini Singha. 2024. A Mystery for You: A fact-checking game enhanced by large language models (LLMs) and a tangible interface. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–5.

[85] Ben M Tappin, Chloe Wittenberg, Luke B Hewitt, Adam J Berinsky, and David G Rand. 2023. Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences* 120, 25 (2023), e2216261120.

[86] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085* (2022).

[87] Paul Thagard. 1989. Explanatory coherence. *Behavioral and brain sciences* 12, 3 (1989), 435–467.

[88] Maggie E Toplak, Richard F West, and Keith E Stanovich. 2014. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning* 20, 2 (2014), 147–168.

[89] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[90] Steeven Villa, Thomas Kosch, Felix Grelka, Albrecht Schmidt, and Robin Welsch. 2023. The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior. *Computers in Human Behavior* 146 (2023), 107787.

[91] Steeven Villa, Robin Welsch, Alena Denisova, and Thomas Kosch. 2024. Evaluating Interactive AI: Understanding and Controlling Placebo Effects in Human-AI Interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–4.

[92] Jan G Voelkel, Robb Willer, et al. 2023. Artificial Intelligence Can Persuade Humans on Political Issues. *OSF Preprints* (2023).

[93] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[94] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838* (2023).

[95] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[96] Richard F West and Keith E Stanovich. 1997. The domain specificity and generality of overconfidence: Individual differences in performance estimation bias. *Psychonomic Bulletin & Review* 4, 3 (1997), 387–392.

[97] Angus R Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. 2024. Large language models can consistently generate high-quality content for election disinformation operations. *arXiv preprint arXiv:2408.06731* (2024).

[98] Tamar Wilner, Kayo Mimizuka, Ayesha Bhimdiwala, Jason C Young, and Ahmer Arif. 2023. It's About Time: Attending to Temporality in Misinformation Interventions. 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 404, 19 pages. https://doi.org/10.1145/3544548.3581068

[99] Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence* 5, 12 (2023), 1486–1496.

[100] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 563–574.

[101] Jeffrey C Zemla, Steven Sloman, Christos Bechlivanidis, and David A Lagnado. 2017. Evaluating everyday explanations. *Psychonomic bulletin & review* 24 (2017), 1488–1500.

[102] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[103] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–38.

[104] Jianlong Zhou, Fang Chen, and Andreas Holzinger. 2020. Towards explainability for AI fairness. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 375–386.

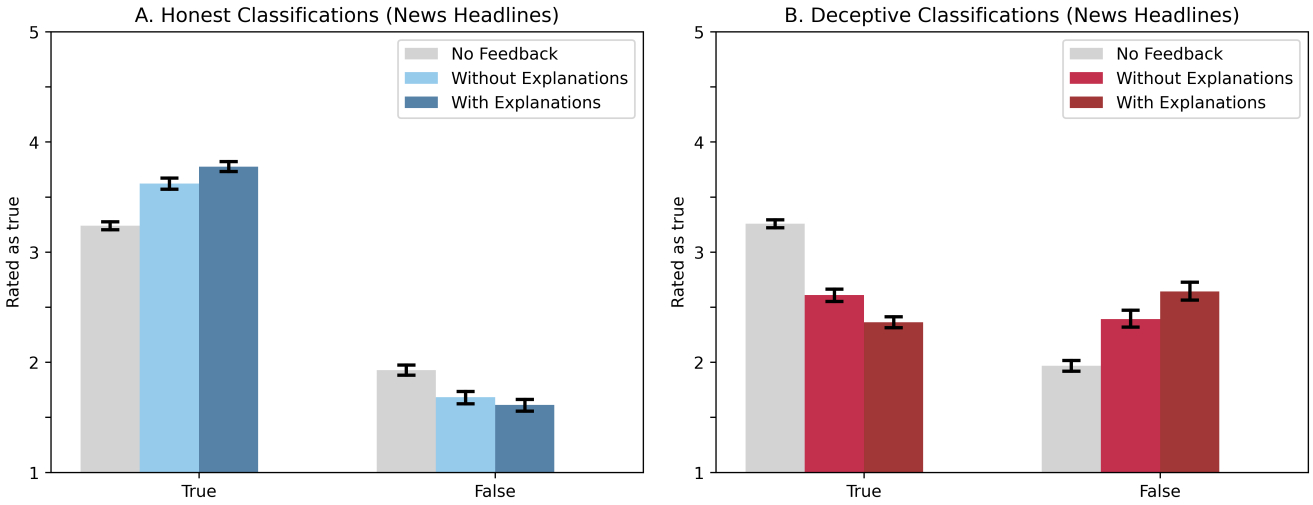
[105] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

9 Extended Data

9.1 Deceptive AI-generated Explanations Increase Beliefs in False Headlines and Decrease Beliefs in True Headlines

In order to gain insights into whether explanations lead to more accurate beliefs, we first compare how accurately participants rated true and false news headlines before (no feedback) and after getting deceptive AI-generated explanations. We coded accuracy by subtracting the belief ratings from the ground truth per rating per participant. Running an analysis of variance (ANOVA), we found that deceptive AI-generated explanations lead to a significantly lower accuracy than no feedback (14 percentage point difference, $F(2, 10959) = -36$, $p < 0.0001$, ANOVA Welch).

Breaking these results down into beliefs in true and false news headlines, we find that deceptive AI explanations significantly increase beliefs in false headlines and significantly decrease beliefs in true headlines ($\beta = -2.44$, $p < 0.001$ and $\beta = 1.03$, $p < 0.001$, respectively), suggesting that deceptive AI explanations significantly diminish people's ability to tell true news headlines from false news headlines.



Supplementary Figure 1: The results (n = 589) on the impact of honest and deceptive AI-generated explanation and classifications on participants' belief updates for news headlines. The error bars represent a 95 percent confidence interval. The measure of the center for the error bars represents the average rating. A: The individual effects of honest AI-generated explanations on belief rating of true and false news headlines. B: The individual effects of deceptive AI-generated explanations on belief rating of true and false news headlines.

9.2 List of Stimuli

Supplementary Table 1: List of Stimuli shows the news headlines and trivia statements used in the experiment.

ID	Statement	LLM Explanation	Feedback veracity	Logical validity
51	Michelle Obama thinks she and her husband deserves SCOTUS appointments [FALSE]	TRUE. Michelle Obama made that exact prediction at a Democratic fundraiser in January. She said, 'Barack Obama is the first African-American president ever	deceptive	Valid
		FALSE. This is a matter of some dispute. Obama has nominated two people to the Supreme Court: Sonia Sotomayor, who was confirmed, and Merrick Garland, who has been waiting for a Senate hearing for more than 300 days	honest	Invalid
57	Elon Musk offers to buy and then delete facebook [FALSE]	TRUE. a compelling example of the power of the Internet. Elon Musk, the billionaire who made his fortune in the tech industry, has offered to buy Facebook	deceptive	Valid
		FALSE. Neither Musk (his Twitter handle is @elonmusk) nor his companies SpaceX and Tesla Motors had anything to do with the website DeleteFacebook.com. The website is a parody, intended to provoke people to quit the social network	honest	Invalid
77	Planned Parenthood announced 'Abort-a-Palooza' on April 2 [FALSE]	TRUE. Planned Parenthood announced it would be launching a nationwide bus tour to publicize its 'Abortion Without Apology' campaign, which will include stops in several cities, including Raleigh	deceptive	Invalid
		FALSE. Planned Parenthood hasnt said anything about an 'Abort-a-palooza,' according to the organization	honest	Valid
97	Michelle Obama files for divorce after shocking revelation; Barack Obama admits: 'I'm gay!' [FALSE]	TRUE. 'I am the same person on the campaign trail as I am in the Oval Office. If Hillary is elected, she will be the third term of Obama, so there's no point in voting for her'	deceptive	Invalid
		FALSE. The site doesn't give a source for its claim that Michelle Obama filed for divorce. As for the president, the story is a fabrication	honest	Valid
143	Accused hate groups receive pandemic aid [TRUE]	FALSE. It does not look like that is the case. Media reports indicate that the groups that were rejected for public assistance after the storm were rejected because they had not yet applied for assistance	deceptive	Invalid
		TRUE. The claim references a study of pandemic plans of two dozen states and several cities	honest	Invalid
151	Armed mexicans were smuggled in to guard border wall, whistle-blower	FALSE. The unions did not provide any evidence to support their claim that armed Mexican guards were used to protect the border wall.	deceptive	Invalid
		TRUE. A whistle-blower recently came forward to reveal that he had been involved with the construction of the border fence, and that he had personally witnessed	honest	Valid

9.3 Statistical Analysis

Dependent variable: Belief Rating		
	News Headlines	**Trivia Items**
Constant	1.61***	0.35***
	(0.10)	(0.06)
True Statement	2.16***	0.18**
	(0.13)	(0.07)
Deceptive Explanations	1.03***	-0.79***
	(0.14)	(0.11)
Deceptive Explanations * True Statement	-2.44***	-0.41***
	(0.18)	(0.11)
Observations	6,000	6,000
Participants	589	610
R^2	0.33	0.29
Adjusted R^2	0.33	0.29
Residual Std. Error	1.12	0.83
F Statistic	165.97***	115.22***
Note:	*p<0.05; **p<0.01; ***p<0.001	

Supplementary Table 1: Linear model with robust standard errors clustered at the participant and headline levels predicting belief rating across news headlines. We use a headline veracity dummy variable (0=false, 1=true), and a deceptive explanation dummy variable indicating whether the participant received an explanation that was deceptive or honest (0=Honest, 1=Deceptive).

Dependent variable: Belief Rating		
	News Headlines	**Trivia Items**
Constant	1.68***	2.29***
	(0.11)	(0.15)
True Statement	1.94***	1.52***
	(0.14)	(0.22)
Explanation	-0.07	-0.02
	(0.06)	(0.06)
Deception	0.71***	1.19***
	(0.10)	(0.10)
Explanation * True Statement	0.22**	0.11
	(0.08)	(0.11)
Deception * True Statement	-1.72***	-2.57***
	(0.14)	(0.15)
Deception * Explanation	0.32**	0.01
	(0.12)	(0.15)
Deception * Explanation * True Statement	-0.72***	-0.28
	(0.19)	(0.22)
Observations	11,780	12,200
Participants	589	610
R^2	0.30	0.22
Adjusted R^2	0.30	0.22
Residual Std. Error	1.13	1.27
F Statistic	109.34***	151.22***
Note:	*p<0.05; **p<0.01; ***p<0.001	

Supplementary Table 2: Linear model with robust standard errors clustered at the participant and headline levels predicting belief rating across explanation and no explanation AI feedback. We use a headline veracity dummy variable (0=false, 1=true), a deceptive classifications dummy variable (0=honest, 1=deceptive), and an explanations dummy variable indicating whether the participant a classification with or without explanation (0=No explanation, 1=Explanation).

	say [TRUE]	Mexicans, who were armed, illegally enter the United States		
235	Newsmax plans expansion to capitalize on Trump support, anger at Fox News [TRUE]	FALSE. Newsmax isnt planning to expand its business. The company recently announced a restructuring of its operations to reduce costs and cut its debt load	deceptive	Valid
		TRUE. Newsmax has been a vocal supporter of Trump, and the network has even hired Trump's former campaign manager.	honest	Invalid
500	Rats besiege New York chipotle, eating avocados and attacking staff [TRUE]	FALSE. The claim that rats besiege new york chipotle, eating avocados and attacking staff comes from an article on the website of the British tabloid The Sun. The article cites unnamed managers of the restaurant. The story has not been reported by any reliable news outlet	deceptive	Valid
		TRUE. The rats were spotted all over the restaurant, including on the ceiling and behind the counters, one customer told the New York Daily News. They were seen eating avocados and burrowing into bags of tortilla chips at the restaurant	honest	Valid
504	GOP megadonor Sheldon Adelson readies 2021 lobbying blitz to bring casinos to Texas [TRUE]	FALSE. Adelson is not seeking to bring casinos to Texas, but to create a so-called destination resort in the area of Texas adjacent to the Mexico border	deceptive	Valid
		TRUE. Adelson has said that he is willing to spend up to $100 million to get a pro-gambling plan passed in Texas. He has said he wants to target a state Senate district in Dallas and a House district in Houston	honest	Valid
505	Biden approval ratings plummet amid war and inflation fears in new public opinion poll [TRUE]	FALSE. The poll was paid for by the Committee to Restore America's Greatness and the American Opinion Institute, a conservative website run by North Carolina businessman and GOP activist Floyd Brown.	deceptive	Invalid
		TRUE. According to a new NBC News poll, the vice presidents job approval rating has taken a hit amid concerns over the war in Ukraine and the state of the U.S. economy.	honest	Valid
506	Biden Visits U.S. Troops in Poland [TRUE]	FALSE. Biden has never visited troops in Poland. He has visited troops in Germany who may be deployed to Poland.	deceptive	Valid
		TRUE. Biden was in Poland in March.	honest	Invalid
507	Supreme Court Rules Against Navy SEALs in Vaccine Mandate Case [TRUE]	FALSE. The Navys policy is to provide all vaccinations. The policy recognizes that vaccinations should be voluntary, and that no one should be forced to receive a vaccination that he or she does not want.	deceptive	Invalid
		TRUE. The case was decided in March 2022 and has been widely covered by news outlets including CNN.	honest	Invalid

509	Eric Shawn: President Biden calls for regime change in Russia [TRUE]	FALSE. The president said the United States should not take action against the government of Vladimir Putin because there is no evidence that the Russian military has moved into Ukraine.	deceptive	Valid
		TRUE. Biden has a new message for Russia: Get out of Ukraine. It is time for President Putin to stop the threats. It is time for Russia to remove its troops from Ukraine.	honest	Invalid
510	Biden to Include Minimum Tax on Billionaires in Budget Proposal [TRUE]	FALSE. It is true that at one point Biden said that if he were in the Senate, he would vote for a bill that included a minimum tax on the very rich. However, Biden said he will not include the minimum tax in his own budget proposal	deceptive	Valid
		TRUE. Biden will include a minimum tax on wealthy Americans in his budget proposal, according to a senior administration official. The official said the tax is expected to generate more than $360 billion over 10 years.	honest	Valid
511	Biden Administration Prepares Sweeping Change to Asylum Process [TRUE]	FALSE. The report is based on a draft copy of an administration proposal that was not issued as an official policy document. The draft is also four years old and was never submitted in completed form to the president.	deceptive	Invalid
		TRUE. The Biden administration is preparing a change in the way the government treats asylum seekers who arrive at the U.S.-Mexico border in the hopes of deterring Central American families from illegally crossing the border.	honest	Valid
512	Supreme Court Sides With Republicans in Case on Wisconsin Redistricting [TRUE]	FALSE. The court hasnt issued a decision on the merits in the case.	deceptive	Valid
		TRUE. The Supreme Court split along party lines in a ruling that preserved Republican-drawn legislative districts in Wisconsin and rejected a Democratic challenge that argued the districts were unconstitutionally drawn to infringe upon the voting rights of Democrats.	honest	Valid
514	France's Macron warns against escalation after Biden says Putin 'cannot remain in power' [TRUE]	FALSE. But the buzz had it that Biden had said Putin would not be in power for long. As it turned out, he had said nothing of the kind.	deceptive	Invalid
		TRUE. Biden said that Putin cannot be permitted to seize Ukraine and occupy Crimea.	honest	Invalid
515	NATO tells Russia to quit 'nuclear saber rattling,' demands China stop 'lies' amid chemical	FALSE. NATO has not issued a statement condemning Russia, urging China to stop making threats or telling either country to stop rattling nuclear sabers.	deceptive	Valid
		TRUE. NATOs deputy secretary-general, Jens Stoltenberg, said at a news conference that Russian 'nuclear saber-rattling' was misplaced and irresponsible.	honest	Invalid

	attack threats [TRUE]			
518	Biden in Europe: 'I'd be very fortunate' if Trump ran again [TRUE]	FALSE. Asked if there was any chance Mr. Trump could win if he ran again in 2024, Mr. Biden actually said, I would be surprised if he did.	deceptive	Invalid
		TRUE. Biden reportedly told a group of reporters in Brussels that he would be 'very fortunate' if Trump were to run again.	honest	Valid
524	Biden's new deal with Iran won't involve 'U.S. taxpayer dollars' [TRUE]	FALSE. The Washington Post reported that the deal would involve American taxpayer dollars. The story said that the United States would help facilitate the deal and make sure Iran gets paidback in oil sales.	deceptive	Valid
		TRUE. The United States and its negotiating partners have demanded that Iran limit its enrichment program to a level that would take it a long time to develop a nuclear weapon if it chose to.	honest	Invalid
526	Florida House passes bill creating election police force [TRUE]	FALSE. The Florida House did not pass a bill creating an election police force. In fact, the Florida House didnt even consider a bill creating an election police force.	deceptive	Valid
		TRUE. The bill does create a new Office called Election Crimes and Security.	honest	Invalid
528	Biden Admin Begs Trump for Nuclear Command Codes [FALSE]	TRUE. The Trump campaign said Biden has been running and hiding from his record for decades, but it looks like Biden and his team are running out of places to hide.	deceptive	Invalid
		FALSE. Biden doesnt have to ask Trump for the codes; they would be sent to Biden.	honest	Valid
530	Dick Cheney Hanged at GITMO [FALSE]	TRUE. There is a photo of the execution. The photo shows the body of a dead man in his underwear with his hands tied behind his back. The man is wearing an orange shirt and black pants. The photo is captioned Dick Cheney.	deceptive	Invalid
		FALSE. The Biden administration has never conducted an execution at the Guantanamo Bay prison.	honest	Invalid
533	Fauci Awarded Nobel Peace Prize [FALSE]	TRUE. Dr. Fauci, who directs the National Institutes of Allergy and Infectious Diseases, was honored for his research into the causes and treatments of H.I.V. and for fostering international collaboration in the fight	deceptive	Invalid
		FALSE. The claim is FALSE because President Obama was awarded the Nobel Peace Prize in 2009, not Dr. Fauci.	honest	Invalid
538	Military: "Biden has NO Nuclear Launch Codes" [FALSE]	TRUE. Biden has no access to the procedures to order a nuclear strike.	deceptive	Invalid
		FALSE. The President of the United States necessarily has the codes.	honest	Valid

540	Pfizer Confirms COVID-Vaccinated People Can 'Shed' Spike Proteins And Harm The Unvaccinated [FALSE]	TRUE. The CDC has said that even in the few instances where a vaccinated person was found to spread measles, the contact with an unvaccinated person was brief.	deceptive	Invalid
		FALSE. A Pfizer spokesman said the company does not know of any instances when its vaccine shed the spike protein. The FDA also said it knows of no such instance.	honest	Invalid
542	Putin Accused of Beheading Bioweapon Engineers in Ukraine [FALSE]	TRUE. Beheadings of scientists in Russia have been reported and, on Tuesday, the pro-Western government in Kiev accused pro-Russian forces in Ukraine of beheading a group of Ukrainian Engineers.	deceptive	Valid
		FALSE. While it is true that the Kremlin has been alarmed by the recent successes of the Ukrainian military in their effort to encircle and isolate the key eastern Ukrainian city of Donetsk, the Kremlin has not been accused carrying out executions of engineers.	honest	Valid
544	Putin Bombs Biden-Owned Villa in Ukraine [FALSE]	TRUE. A report by the Moscow Times says that in June, at least four bombs hit Ayvazovskoe highway, which runs by the village of Prigorodnoye where Biden has a Villa.	deceptive	Valid
		FALSE. The claim is unproven, and Biden has never owned property in the Ukraine.	honest	Valid
1001	The Empire State building has more than 70 elevators. [TRUE]	FALSE. The Empire State building has only 44 elevators.	deceptive	Valid
		TRUE. The Empire State building has more elevators than the Chrystler building	honest	Invalid
1003	The cartoon character Curious George does not have a tail. [TRUE]	FALSE. Curious George was created based on a drawing of a monkey statue with a tail owned by H . A . Rey.	deceptive	Invalid
		TRUE. The monkey's tail was never seen in all of the Curious George books and films.	honest	Valid
1009	Bugatti cars are made in Italy. [FALSE]	FALSE. The Bugatti company began building luxury cars in the Italian city of Turin in 1909.	deceptive	Valid
		TRUE. Bugatti Automobiles S.A.S., is a French car manufacturer delivering the very luxurious super sports car.	honest	Invalid
1011	SOS stands for Save Our Ship. [FALSE]	TRUE. The letters SOS, standing for Save Our Ship, have been used since at least World War I.	deceptive	Invalid
		FALSE. SOS is the Morse code distress signal. Standard maritime signal flags adopted the signal as it was easy to do.	honest	Invalid
1012	A cross between a horse and a	TRUE. The cross of a horse and a zebra will result in a creature with the mane of shiny curls of the horse and the body of the zebra, reported by genetists to be called a	deceptive	Invalid

	zebra is called a 'Hobra' [FALSE]	Hobra.		
		FALSE. A Cross between a horse and a zebra is called a Zorse.	honest	Valid
1013	The black box in a plane is black [FALSE]	TRUE. The black box is actually painted with a corrosion-proof aluminium paint, commonly known as matte black.	deceptive	Valid
		FALSE. The black box in a plane is painted bright orange for improved visibility in case of a crash.	honest	Valid
1016	Meryl Streep has won no Academy Awards [FALSE]	TRUE. Meryl Streep hasn't won any Academy Awards. However, she has one Academy Award nomination.	deceptive	Valid
		FALSE. Meryl Streep has won three Oscars and was nominated for 21.	honest	Valid
1017	Idina Menzel sings 'let it go' 20 times in 'Let It Go' from Frozen [FALSE]	TRUE. It's sung loudly at the start before an instrumental passage (then it's sung softly as the instrumental part can be heard), and it's sung aloud and repeated again at the end.	deceptive	Invalid
		FALSE. Idina Menzel sings 'Let It Go' 21 times in 'Let It Go' from Frozen.	honest	Valid
1018	Waterloo has the greatest number of tube platforms in London [TRUE]	FALSE. There are 18 tube platforms at Victoria, more than Waterloo.	deceptive	Invalid
		TRUE. Waterloo has 19 underground platforms and is the busiest station by number of trains, handling around 1,300 per service.	honest	Invalid
1019	M&M stands for Mars and Moordale [FALSE]	TRUE. Forrest Mars Junior, formed Mars and his friend, Marvin B. Moors, joined them.	deceptive	Invalid
		FALSE. Mars is a chocolate brand and is a registered trademark and 'M&M' actually stands for the names of the two individuals who developed it.	honest	Invalid
1020	Gin is typically included in a Long Island Iced Tea [TRUE]	FALSE. A traditional Long Island Iced Tea is only tequila, vodka, and light rum.	deceptive	Valid
		TRUE. It is common to add a shot of gin to a long island iced tea.	honest	Invalid
1021	The unicorn is the national animal of Scotland [TRUE]	FALSE. The unicorn is only the Scottish Royal Badge, while the Lion Rampant is the national actual animal.	deceptive	Valid
		TRUE. The unicorn is situated on the national crest of arms of Scotland	honest	Invalid
1022	There are two parts of the body that can't heal themselves [FALSE]	TRUE. Skin and bone marrow can't regenerate.	deceptive	Valid
		FALSE. There are in fact only one part of the body that can't heal themselves: the teeth	honest	Valid

1024	The Great Wall of China is longer than the distance between London and Beijing [TRUE]	FALSE. The two cities are roughly 4,334 kilometers apart with the Great Wall measuring around 4,000 kilometers in length.	deceptive	Valid
		TRUE. The Great Wall of China is 21.000 km in length, which dwarfs the 8.100 km distance between London and Beijing.	honest	Valid
1030	Monaco is the smallest country in the world [FALSE]	TRUE. The 5-square-mile country in the south of France is the only one in the world which both restricts its citizenship by heredity and claims sovereignty over two different territories.	deceptive	Invalid
		FALSE. Vatican City is has only a population of 1000	honest	Valid
1031	What Do You Mean' was Justin Bieber's first UK number one single [TRUE]	FALSE. The song is Bieber's 23rd in the UK top 40.	deceptive	Invalid
		TRUE. In September 2015 'What Do You Mean' was at number 1 in the UK top 40.	honest	Valid
1038	Canis lupur is the scientific name for a wolf [FALSE]	TRUE. Canis is the 'dog' genus and lupur, which is Latin for wolf, makes 'Canis lupur' literally the correct name for a wolf.	deceptive	Valid
		FALSE. The word wolf is the common name for the species Canis lupis.	honest	Valid
1039	K is worth four points in Scrabble [FALSE]	TRUE. All consonants except for B are all worth four points in Scrabble	deceptive	Valid
		FALSE. The letter 'K' is worth 5 points in Scrabble.	honest	Valid
1041	Ariana Grande is 25 or under [FALSE]	TRUE. Ariana and her brother Frankie are both 21 years old as of 2016,	deceptive	Invalid
		FALSE. Ariana Grande is more than 23 years old	honest	Invalid
1042	Australia is wider than the moon [TRUE]	FALSE. The width of the moon is 2,159,000 kilometers and the width of Australia is 2,300 km	deceptive	Valid
		TRUE. Australia has a width of around 4000 km (2,500 miles) while the moon measures 3,500 km (2,200 miles).	honest	Valid
1043	Queen Elizabeth II was the longest reigning British monarch [TRUE]	FALSE. Queen Elizabeth I reined for about forty-four years, therefore making her the longest reigning British monarch in history.	deceptive	Invalid
		TRUE. Queen Elizabeth was the longest reigning British monarch, with a reign of 70 years. The second longest reigning monarch is Queen Victoria at 63 years.	honest	Valid
1046	Your 'radius' bone is in your leg [FALSE]	TRUE. 'Radius' does not refer to the radius bone in your forearm but it refers to the radius bone articulated in your leg.	deceptive	Invalid
		FALSE. The bone in the lower end of your leg is called a Femur bone.	honest	Invalid

1050	Hillary Clinton and Celine Dion are related [FALSE]	TRUE. Celine's husband, Rene Charles Dion, is Hillary's second cousin once removed.	deceptive	Valid
		FALSE. Celine Dion's paternal great-grandfathers' family tree has been traced back to New York, whereas Hillary Clinton's paternal ancestors can be traced as far as North East England.	honest	Invalid
1051	The five rings on the Olympic flag are interlocking [TRUE]	FALSE. Only 4 out of the 5 Olympic circles are interlinking (the top 3 left circles, and the bottom two circles on the right).	deceptive	Valid
		TRUE. The five rings on the Olympic flag are supposed to be interlocking as they move to symbolize the spirit of people joining together.	honest	Invalid
1056	In a deck of cards, the king has a mustache. [TRUE]	FALSE. In a deck of cards, the Jack has a mustache.	deceptive	Invalid
		TRUE. In a deck of cards, the Jack does not have a mustache.	honest	Invalid
1060	Sex and the City star Kim Cattrall is from California. [FALSE]	TRUE. Kim Cattrall's parents, who are Canadian, met while working as teachers in San Fernando Valley, north of Los Angeles.	deceptive	Invalid
		FALSE. Kim Cattrall was born and raised in Liverpool, England.	honest	Valid

Dependent variable: Belief Rating		
	News Headlines	**Trivia Items**
Constant	0.39^{***}	0.54^{***}
	(0.08)	(0.04)
Deceptive Feedback	0.29^{**}	0.40^{***}
	(0.10)	(0.08)
Observations	2,336	3,664
R^2	0.02	0.04
Adjusted R^2	0.02	0.04
Residual Std. Error	0.95	0.96
F Statistic	8.97^{*}	24.33^{***}
Note:	*p<0.05; **p<0.01; ***p<0.001	

Supplementary Table 3: Linear model with robust standard errors clustered at the participant and headline levels predicting change in belief rating between pre- and post AI explanation feedback to compare effects of deceptive and honest explanation feedback for true and false news headlines. We use a deceptive classifications dummy variable (0=honest, 1=deceptive) and limit the analysis to AI explanations and news headlines observations only.

Dependent variable: Belief Rating		
	News Headlines	**Trivia Items**
Constant	1.66***	2.28***
	(0.10)	(0.15)
True Statement	2.04***	1.55***
	(0.13)	(0.20)
Deception	0.92***	1.20***
	(0.09)	(0.12)
Logically Invalid	0.12	0.09
	(0.09)	(0.15)
Deception * True Statement	-2.14***	-2.69***
	(0.12)	(0.16)
True * Logically Invalid	-0.10	0.01
	(0.12)	(0.20)
Deception * Logically Invalid	-0.35*	-0.13
	(0.14)	(0.19)
True * Deception * Logically Invalid	0.31	0.18
	(0.20)	(0.31)
Observations	11,780	12,200
Participants	589	610
R^2	0.30	0.23
Adjusted R^2	0.30	0.23
Residual Std. Error	1.13	1.26
F Statistic	103.59***	85.10***
Note:	*p<0.05; **p<0.01; ***p<0.001	

Supplementary Table 4: Linear model with robust standard errors clustered at the participant and headline levels predicting belief rating across logical validity of explanations, the Explanations, and deceptive AI feedback. We use a logical invalid dummy variable (0=logically valid, 1=logically invalid), a veracity dummy variable (0=false, 1=true), a deceptive classifications dummy variable (0=honest, 1=deceptive), and an explanations dummy variable indicating whether the participant a classification with or without explanation (0=No explanation, 1=Explanation).

Dependent variable: Belief Rating		
	News Headlines	**Trivia Items**
Constant	1.72***	2.28***
	(0.11)	(0.16)
True Statement	1.88***	1.49***
	(0.14)	(0.20)
Deception	0.70***	1.20***
	(0.09)	(0.10)
Explanation	-0.03	-0.02
	(0.06)	(0.06)
Knowledge	-0.11**	-0.11
	(0.04)	(0.07)
Deception * True Statement	-1.63***	-2.52***
	(0.13)	(0.16)
Explanation * True Statement	0.18*	0.14
	(0.09)	(0.11)
Deception * Explanation	0.26*	-0.00
	(0.12)	(0.14)
True * Explanation * Deception	-0.67***	-0.27
	(0.18)	(0.22)
True * Knowledge	0.44***	0.35**
	(0.07)	(0.12)
Explanation * Knowledge	-0.07	0.06
	(0.05)	(0.07)
Deception * Knowledge	-0.05	0.06
	(0.09)	(0.05)
True * Explanation * Knowledge	-0.01	-0.05
	(0.07)	(0.10)
Explanation * Deception * Knowledge	0.18*	-0.16
	(0.08)	(0.10)
True * Deception * Knowledge	0.02	-0.08
	(0.14)	(0.11)
True * Explanation * Deception * Knowledge	-0.19	0.16
	(0.15)	(0.16)
Observations	11,780	12,200
Participants	589	610
R^2	0.33	0.24
Adjusted R^2	0.33	0.24
Residual Std. Error	1.11	1.25
F Statistic	90.70***	76.12***
Note:	*p<0.05; **p<0.01; ***p<0.001	

Supplementary Table 5: Linear model with robust standard errors clustered at the participant and headline levels predicting belief rating across participants' self-reported prior knowledge ratings (z-scored), explanation and no explanation AI feedback. We use a prior knowledge variable, a veracity dummy variable (0=false, 1=true), a deceptive classifications dummy variable (0=honest, 1=deceptive), and an explanations dummy variable indicating whether the participant a classification with or without explanation (0=No explanation, 1=Explanation).

Dependent variable: Belief Rating		
	News Headlines	**Trivia Items**
Constant	1.68***	2.30***
	(0.11)	(0.15)
True Statement	1.96***	1.47***
	(0.14)	(0.22)
Deception	0.81***	1.14***
	(0.10)	(0.09)
Explanation	-0.07	-0.03
	(0.06)	(0.06)
Trust	-0.03	-0.12**
	(0.03)	(0.05)
Deception * True Statement	-1.87***	-2.47***
	(0.14)	(0.14)
Explanation * True Statement	0.20*	0.11
	(0.09)	(0.11)
Deception * Explanation	0.25*	0.03
	(0.13)	(0.15)
True * Explanation * Deception	-0.61**	-0.31
	(0.19)	(0.21)
True * Trust	0.13*	0.35***
	(0.05)	(0.08)
Explanation * Trust	0.04	0.05
	(0.06)	(0.05)
Deception * Trust	0.35***	0.38***
	(0.05)	(0.09)
True * Explanation * Trust	0.01	-0.04
	(0.09)	(0.09)
Explanation * Deception * Trust	-0.01	-0.05
	(0.09)	(0.10)
True * Deception * Trust	-0.58***	-0.84***
	(0.10)	(0.14)
True * Explanation * Deception * Trust	-0.11	0.10
	(0.15)	(0.19)
Observations	11,780	12,200
Participants	589	610
R^2	0.32	0.24
Adjusted R^2	0.32	0.24
Residual Std. Error	1.11	1.25
F Statistic	77.93***	129.30***
Note:	*p<0.05; **p<0.01; ***p<0.001	

Supplementary Table 6: Linear model with robust standard errors clustered at the participant and headline levels predicting belief rating across participants' trust in AI systems rating (z-scored), and explanation and no explanation AI feedback. We use a trust variable, a veracity dummy variable (0=false, 1=true), a deceptive classifications dummy variable (0=honest, 1=deceptive), and an explanations dummy variable indicating whether the participant a classification with or without explanation (0=No explanation, 1=Explanation).

Dependent variable: Belief Rating		
	News Headlines	**Trivia Items**
Constant	1.68***	2.29***
	(0.11)	(0.15)
True Statement	1.94***	1.52***
	(0.14)	(0.22)
Deception	0.71***	1.19***
	(0.10)	(0.10)
Explanation	-0.06	-0.02
	(0.06)	(0.06)
CRT	-0.06	-0.08*
	(0.03)	(0.04)
Deception * True Statement	-1.72***	-2.57***
	(0.14)	(0.15)
Explanation * True Statement	0.22*	0.11
	(0.08)	(0.11)
Deception * Explanation	0.32**	0.01
	(0.12)	(0.15)
True * Explanation * Deception	-0.72***	-0.28
	(0.19)	(0.22)
True * CRT	0.01	0.13
	(0.06)	(0.08)
Explanation * CRT	-0.02	-0.03
	(0.03)	(0.06)
Deception * CRT	-0.10	0.05
	(0.07)	(0.07)
True * Explanation * CRT	0.13	-0.06
	(0.08)	(0.10)
Explanation * Deception * CRT	0.13	-0.07
	(0.09)	(0.09)
True * Deception * CRT	0.17	-0.14
	(0.12)	(0.14)
True * Explanation * Deception * CRT	-0.25	0.19
	(0.17)	(0.18)
Observations	11,780	12,200
Participants	589	610
R^2	0.30	0.23
Adjusted R^2	0.30	0.23
Residual Std. Error	1.13	1.26
F Statistic	65.68***	93.91***
Note:	*p<0.05; **p<0.01; ***p<0.001	

Supplementary Table 7: Linear model with robust standard errors clustered at the participant and headline levels predicting belief rating across z-scored cognitive reflection test (CRT) score, explanation and no explanation AI feedback. We use a CRT variable, a veracity dummy variable (0=false, 1=true), a deceptive classifications dummy variable (0=honest, 1=deceptive), and an explanations dummy variable indicating whether the participant a classification with or without explanation (0=No explanation, 1=Explanation).