# Things Are Clear to Me Now! – Exploring the Effectiveness of AI-Generated Contextual Explanations in Digitally Altered and Synthetic Images

**Saquib Ahmed, Tejo Gayathri Busireddy, Sanorita Dey**

University of Maryland, Baltimore County
saquiba1@umbc.edu, tejogab1@umbc.edu, sanorita@umbc.edu

## Abstract

In the digital era, the exponential growth of images and videos on social platforms has transformed how individuals perceive information and form opinions. However, the escalating prevalence of altered and synthetic visuals poses significant challenges to media trust. These altered visuals often mislead viewers, propagate confusion, and distort public perception. Social media algorithms, optimized for engagement, can inadvertently amplify the dissemination of such content, making simple tagging insufficient to distinguish authentic from altered visuals. Contextual explanations present a promising approach by offering audiences deeper insights and encouraging more informed interpretations. In this study, we developed contextual explanations for 15 altered and synthetic images and conducted a user study to evaluate their effectiveness. Our findings show that contextual explanations consistently outperformed non-contextual ones across all evaluated metrics. We also assessed the capability of large language models (LLMs) to generate these explanations for diverse audiences. While LLM-generated explanations were generally comparable to those created by human experts, the models exhibited limitations in conveying intrinsic motivations in complex scenarios. We conclude with a discussion of the design implications and ethical considerations of this work.

## Introduction

In the digital era, images and videos have grown exponentially on social platforms. Posts containing visuals attract significantly more engagement than text-only posts (Li and Xie 2020), reshaping the way people perceive information and form opinions. Younger audiences, in particular, view social platforms as equivalent to traditional news sources, trusting its content as much as national news agencies (Hladík and Štětka 2017). However, synthetic and altered visuals, such as deep and shallow synthetics, undermine trust in the media. Such fabricated visuals can easily convey misleading narratives and provoke strong emotional reactions (Allcott and Gentzkow 2017). When these exposures are repeated several times, they can create a lasting mental impression even after being debunked or corrected publicly (Teubner-Rhodes et al. 2017) and can wrongly shape public perceptions. These fabricated narratives, designed to mimic authenticity, facilitate misleading content and deliberate deception (Maras and Alexandrou 2019; Shu and et al. 2017). Social media algorithms, focused on maximizing engagement, inadvertently amplify the spread of such misleading content (Vosoughi, Roy, and Aral 2018).

Prior work found that people most often struggle to identify fabricated or altered images (Moravec, Dennis, and Minas 2020b). People's inherent trust in visual information makes them more susceptible to believing and sharing altered media (Messaris and Abraham 2001; O'Brien and Lauer 2018). To overcome this challenge, there has been ongoing efforts to develop advanced algorithms that can detect even sophisticated fabricated images. Social platforms use such algorithms to tag synthetic content to warn their audience. However, tagging synthetic and altered images alone is often not enough to combat inaccurate claims, as scrolling through social media is a relaxing, pleasure-seeking task and people can easily miss subtle, non-contextual tags or flags in such relaxed state of mind (Moravec, Dennis, and Minas 2020a).

It is critical to understand how social platforms raise awareness about synthetic or fabricated images so they are noticeable and help users grasp the underlying agenda behind their posting. We hypothesize that understanding this hidden agenda—rather than simply knowing an image is synthetic—can motivate individuals to overcome misperceptions. This hypothesis draws on Brandolini's Law, which states that refuting misinformation requires significantly more effort than producing it (Brandolini et al. 2020). Effective debunking, therefore, demands explanations that not only clarify inaccuracies but also provide context and expose manipulative intent (Chan et al. 2017). In this paper, we explore contextual explanations for synthetic and altered images, which are more complex to generate than simple tags or flags conveying only one bit of information. We envision that such explanations can reveal the underlying agendas behind synthetic content. Specifically, we asked the following research questions.

- **RQ1** How can we generate explanations for synthetic and altered multimedia content on social platforms that can create a contextual understanding of the content among the audience?

Generating contextual and informative explanations for synthetic images is not an objective task. It requires understanding the sensitivity of the topic as well as the impact of narratives on people's perception. Media personalities and journalists do this job on a regular basis as part of their profession need. However, because of the sheer volume of the synthetic images on social platforms, it is unrealistic to assume that media personnel alone could generate all such explanations. Possible and more scalable solutions will be to employ either crowd workers (human-generated) or LLM engines (machine-generated). It is still not well-explored how effectively AI-agents can generate contextual explanations for synthetic images compared to human beings. So we asked:

- **RQ2** How well can off-the-shelf LLM-based agents generate contextually relevant explanations for synthetic and altered content compared to human beings of different levels of skillset?

To address RQ1, we designed two types of contextual explanations (Contrastive and Teleological) for synthetic images, drawing inspiration from the explainable AI literature (Miller 2019). A user study revealed that contextual explanations not only clarified how images were altered to achieve malicious objectives but also encouraged reflection on the broader intent behind their dissemination. For RQ2, we explored three sources for generating contextual explanations: crowd workers, journalism experts, and off-the-shelf large language models (LLMs). Experts rated LLM-generated explanations higher in quality; however, their complexity and readability may pose challenges for public adoption. We conclude by addressing the design implications and ethical considerations of this work.

Our work is well-aligned with the goals of the AIES community, as it sits at the intersection of AI, ethics, and societal impact. By investigating how AI-generated contextual explanations influence people's perception of synthetic and altered media, the study directly addresses ethical challenges related to misleading content, algorithmic transparency, and responsible AI deployment in sociopolitical contexts—core themes within the AIES community. In summary, this work contributes the following to the AIES community:

- An in-depth, theory-based empirical investigation of how to design explainable, human interpretable, and contextually relevant explanations for synthetic images
- An empirical, comparative exploration to understand the capacity of LLM-based agents to generate contextual explanations for synthetic images
- Directions for generating of future contextual and relevant explanations for synthetic images to support conscious and informed perception building of the audience of social platforms

## Related Works

### Impact of altered and synthetic multimedia content

Misleading content is pervasive online, particularly on social media, and spans domains such as public health (Malki, Patel, and Singh 2024), finance (Lodge et al. 2022), language communities (Fawzi and Magdy 2024), and belief systems (Khan, Ghafourian, and Patil 2024). Much of this content is multimedia-based and involves both public figures and private individuals. In 2016, 62% of U.S. adults obtained news from social networks (Shu and et al. 2017), with the number still high at 54% in 2024 (Pew 2024). Such exposure escalates susceptibility to ideologically motivated inaccurate claims that can polarize discourse (Olan et al. 2024), including both unintentional and deliberate forms of deception (Melchior and Oliveira 2024).

The electoral impact of misleading content is well-documented. During the 2016 the U.S. presidential election, false content appeared to disproportionately benefit one candidate (Allcott and Gentzkow 2017), raising concerns about its influence on public opinion (Franklin et al. 2017; Wilder and Vorobeychik 2019). Engagement-driven algorithms further amplify such content, reinforcing echo chambers and social divides (Binns 2020; Bessi et al. 2016). Within this ecosystem, visuals—such as manipulated images or graphs—can mislead due to their perceived objectivity (Hawkins 2018; Meyer and Schwabe 2020; Friedman 2014; O'Brien and Lauer 2018; Vargas-Restrepo et al. 2019).

Deepfakes compound these issues, raising concerns over consent, privacy, and informational integrity (Heidari et al. 2024; Maras and Alexandrou 2019; Debroy and Hemmige 2024; Cai and et al. 2022). Their harms are both individual and societal, with documented emotional distress in cases of unauthorized image use, including incidents involving school children (Khalifa, Anjum, and Qu 2024). While political implications draw significant attention, deepfakes disproportionately target women via non-consensual explicit content (Dunn 2021; Maddocks 2020), such as in the case of journalist Rana Ayyub, who was driven offline due to targeted harassment (Ayyub 2018). Once circulated, such content is hard to remove and causes lasting harm (Chesney and Citron 2019; Henry et al. 2020), underscoring the need for interventions like contextual explanations.

Research on synthetic and altered imagery has primarily focused on algorithmic detection—identifying features that distinguish real from synthetic images while optimizing models for speed and precision (Mutlu, Yousefi, and Ozmen Garibay 2022; Saini and Prasad 2022). However, detection alone is insufficient (Leibowicz, McGregor, and Ovadya 2021; Bird, Ungless, and Kasirzadeh 2023). Simply tagging content as "synthetic" often fails to engage users or prompt critical evaluation (Walker et al. 2023; Lyu et al. 2022). Without clear explanatory context, users may dismiss these labels or misunderstand their relevance (Bertrand et al. 2022; Fabbri 2023). Our work addresses this gap by investigating how explanatory strategies can strengthen understanding and encourage more informed judgments of manipulated media.

### Explanations Can Bring Changes

Individuals often encounter unfamiliar topics online, where anchoring bias—the tendency to rely heavily on initial information regardless of its accuracy—can strongly influence interpretation and decision-making (Furnham and Boo 2011). This cognitive bias affects various domains, includ-

ing purchasing (Simonson and Drolet 2004), judicial reasoning (Hunt 1941), market behavior (Beggs and Graddy 2009), and estimations (Kaustia, Alho, and Puttonen 2008), and also shapes engagement with altered multimedia content. Brandolini's Law underscores the asymmetry between the ease of spreading falsehoods and the difficulty of correcting them (Brandolini et al. 2020), a challenge amplified by rapid digital information flows (Vosoughi, Roy, and Aral 2018). Misleading content can persist in memory, especially in emotionally charged or socially relevant areas such as public health and civic discourse (Lewandowsky, Ecker, and Cook 2012; Schwarz et al. 2016). Social media dynamics and confirmation biases further reinforce such beliefs, fueling interest in explainable AI (XAI) and personalized methods to counter inaccurate claims (Chou, Gaysynsky, and Vanderpool 2018; Maas and Liao 2019).

Refutation, which not only labels content as false but also explains the evidence behind its inaccuracy, is among the most effective corrective strategies. It helps bridge knowledge gaps and supports longer-term memory integration (Tippett 2010; Sanderson and Ecker 2020). Refutation has proven successful in political discourse and science education (Guzzetti et al. 1993; Kowalski and Taylor 2009), and prior work has used it to highlight deceptive elements in synthetic images (Ruffin, Wang, and Levchenko 2023). Meta-analyses indicate that refutations outperform simple retractions, particularly when explanations are detailed and context-sensitive (Chan et al. 2017; Ecker et al. 2020; Walter and Murphy 2018; Kendeou et al. 2014). Complementary approaches such as behavioral nudges (Jahanbakhsh et al. 2021), community moderation (Jahanbakhsh, Zhang, and Karger 2022; Chuai et al. 2024), and platform alerts (e.g., credibility labels or context tags) also show promise in fostering critical engagement, though their effects remain nuanced (Cook, Ecker, and Lewandowsky 2017; Ecker, Lewandowsky, and Tang 2010; Blair et al. 2017; Lyons 2017).

Recently, AI-generated explanations have emerged as a promising tool against misleading content through multimedia. These systems generate contextual, free-text explanations that improve comprehension and trust, especially for users lacking technical verification expertise (Schmitt et al. 2024). Their social integration and diversity of perspectives strengthen perceived credibility (Mittelstadt, Russell, and Wachter 2019), and recent studies highlight their potential for scalable fact-checking despite resource constraints (Wolfe and Mitra 2024). Building on this, our work examines how AI-generated explanations can bolster users' contextual understanding and resilience when encountering misleading or fabricated multimedia content. This work extends explainable automated fact-checking into the less-explored domain of images, addressing a gap in prior text-focused research (Atanasova 2024; Augenstein et al. 2024).

## Study 1: Designing Explanations for Synthetic Images

Study 1 aims to explore how to design contextually relevant explanations for synthetic and altered multimedia content, and to assess how these explanations help audiences understand the narrative changes intended by synthetic images. This study specifically addresses RQ1. The methodology flowchart for Study 1 is shown in Figure 1, with the following subsections providing a concise overview of the methodology and key findings.

### Methodology: Study 1

**Dataset for Synthetic and Altered Images** Synthetic and edited images of political and social events often gain traction on social media, even when intended as satire (NBC 2024). These visuals can influence democratic discourse and public perception. This study examines altered images related to political figures, events, and movements, focusing on 15 selected from an initial pool of 37 for their relevance to a U.S. audience. Six were drawn from a prior dataset (Ruffin, Wang, and Levchenko 2023), and nine were sourced from platforms such as Reddit, X (formerly Twitter), Snopes, and Google. Nine images include side-by-side comparisons with originals, while six are AI-generated (Cheetham and Joshua 2023; Wendling 2024). Repeated appearances include Donald Trump (3), Barack Obama (2), and Joe Biden (2), with no individual featured more than three times to minimize unbalanced perception. Images lacking political context or narrative alignment were excluded since they fell outside of our research scope.

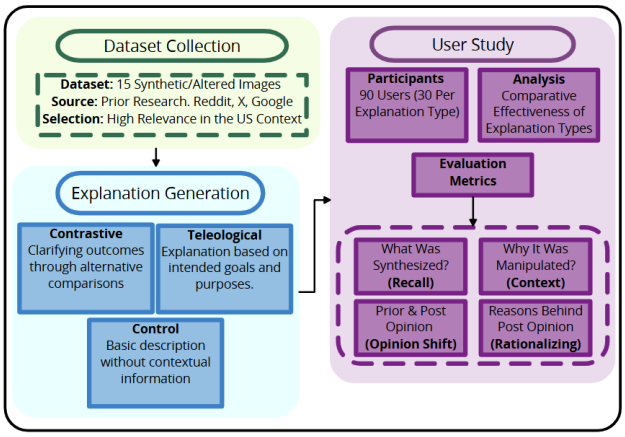


Figure 1: The flow diagram for study 1, which aims to understand how contextually relevant explanations for synthetic and altered multimedia content can be designed and whether and how it can help the audience understand the context, in which the synthetic image tried to change the narrative.

**Designing Explanations for Synthetic Images** To design contextually relevant and meaningful explanations, we grounded our approach in the social science literature on AI explainability (Miller 2019; Mittelstadt, Russell, and Wachter 2019). Explanations of behavior often take the form of reasons, causal history, and enabling factors, each reflecting distinct intentional and contextual dimensions (Malle 2006). Causal interpretations can also vary across different

structural forms—temporal, coincidental, unfolding, opportunity, and preemptive chains—that influence how events are cognitively linked (Hilton and John 2007). We prioritized explanation formats that are logically coherent and accessible to a broad audience. Explanations may clarify why one event occurred instead of another (Lipton 1990), emphasize the underlying purposes or goals that motivate an outcome (Miller 2019), or highlight deviations from normative expectations to account for salience or surprise (Hilton and Slugoski 1986). While many AI explanations are tailored for expert users familiar with internal model mechanisms, such framework does not translate meaningfully to the domain of synthetic images. Instead, we adopted two explanation types, guided by the theoretical motivations discussed above, that are specifically suited to altered or synthetic imagery related to social and political events. This also ensures the resulting explanations remain efficient and accessible for lay audiences. The two explanation types are briefly discussed here.

**Contrastive Explanation** This explanation type reflects how individuals understand events as dependent on specific conditions and often interpret outcomes by comparing them to plausible alternatives. It incorporates the idea that an outcome is meaningful when considered in light of what could have happened instead, offering insight not only into what occurred, but why it occurred in that particular way (Hume 2007; Lewis 2000; Hilton 1990; Lipton 1990). Contrastive explanations align with human expectations for intuitive and context-sensitive reasoning, particularly in socially and politically charged scenarios where understanding causal distinctions is essential.

**Teleological Explanation** This explanation type characterizes events in terms of their intended goals and perceived deviations from expected norms. It reflects the human tendency to interpret actions by referencing their overarching purpose or intended outcome (Hankinson 2001), offering insight into why an event was aimed at a particular end rather than how it was brought about. Teleological explanations are especially salient when events appear atypical or surprising, as they help observers make sense of such deviations by situating them within a framework of intentionality and socially recognized aims (Hesslow 1988; Hilton 1996; Hilton and Slugoski 1986).

**Generating Explanations for Synthetic Images** The first two authors individually created two types of explanation for each image (as per the definitions) and met daily to discuss and update their perspectives. Weekly, all authors of the paper discussed together to established uniformity in generated explanations and addressed inconsistencies. After several iterations, the explanations were finalized. While **contrastive** explanations addressed the comparative rationale behind the alteration, **teleological** explanations offered a larger context. On the other hand, a **control** explanation describes the edited part without context, as demonstrated in prior work (Ruffin, Wang, and Levchenko 2023). Three explanations for Figure 2 are presented in Table 1 for reference.

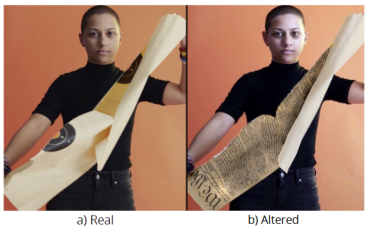


Figure 2: The real and the altered version of the image of Emma González

**Study Procedure** Once the contextual explanations for synthetic images were generated, our goal was to assess how well these explanations helped participants understand the alteration and recall the images for future reference. We aimed to determine whether the structure of the explanations impacted recall factor compared to the control explanation, which lacked context. To investigate this, we conducted a user study. The recruitment materials, study protocol, data collection, and risk factors for user study were reviewed and approved by a full IRB review board at UMBC.

We recruited Prolific crowd workers (Pro 2025) to use a web platform resembling a social media page. Participants reviewed 45 news posts: 15 with synthetic/altered images and 30 real political or social event posts. Real posts included images with descriptions; synthetic/altered posts showed either the original–altered image pair or only the synthetic image, each with one explanation (contrastive or teleological), which can be found in https://tinyurl.com/2p3up232. Each participant saw a single explanation type for all synthetic images. We recruited 30 participants per type (N = 90), meeting the central limit theorem threshold and power analysis for large effects. Post order was randomized to reduce bias (Figure 3).

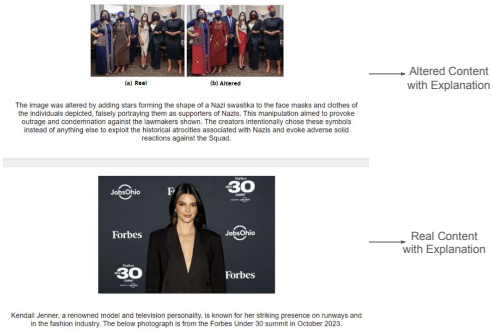


Figure 3: The view with altered and real multimedia contents with explanations

Participants could take as much time as needed to examine the posts. Afterward, they completed a survey when they could not access the posts, relying on memory recall. The survey included questions about 15 posts: 10 synthetic images (from the 15 synthetic image posts) and 5 real news posts (from the 30 real posts). These posts were randomly

| Explanation Type | Example |
|---|---|
| **Contrastive Explanation** | The synthetic image shows Emma Gonzalez, who was found tearing The US Constitution, which was changed from shooting target paper. The community against gun control altered the image to portray her as opposing the US Constitution and citizens' gun rights. They used the Constitution instead of a regular paper or shooting practice paper to impact people emotionally, suggesting that those against gun violence oppose the Constitution. |
| **Teleological Explanation** | The synthetic image shows Emma Gonzalez, who was found tearing The U.S. Constitution, which was changed from shooting target paper. The anti-gun control community altered Emma González's image to counter her strong support against gun violence. She turned her grief into activism, founded "Never Again MSD," and organized the "2018 March for Our Lives" in D.C., where her impactful speech gained media attention, likely prompting the image alteration. |
| **Control Condition** | This image is synthetic, where the shooting target paper found in the hand of Emma Gonzalez is changed to the US Constitution. |

Table 1: Examples of different explanation types for the altered image in Figure 2.

selected for each participant. For each synthetic image, participants were asked what part of the image was altered, why it was altered, their initial opinion of the event or political figure, whether their opinion changed after seeing the explanation, and why. For real posts, participants were asked to recall the content. However, responses to real posts were not evaluated. The survey included brief descriptions of the image topic to pinpoint the corresponding image, but no images were shown.

Survey questions for Figure 2 are, **Q1:** What part of the Emma Gonzalez image was synthesized/altered in which she was shown tearing a large piece of paper? **Q2:** Why was the picture of Emma Gonzalez tearing a large piece of paper being synthesized? **Q3:** What was your opinion on stricter gun control laws in the United States? **Q4:** What is your current opinion on gun control after seeing the explanation of the synthetic image of Emma Gonzalez? (more favorable/no change/less favorable) **Q5:** Why did your opinion become more/less favorable on gun control after seeing the explanation?/Why did your opinion not change about gun control after seeing the explanation?

At the end, participants completed a demographic survey and received a debriefing about the study's true objective—examining the impact of contextually aware explanations of synthetic images on social media. On average, participants spent 40.94 minutes (SD = 12.22) to complete the study and were paid $10 each. To ensure quality control, we discarded 5. 6% responses from participants who completed the study in less than 10 minutes or failed attention check questions. Additionally, we excluded responses that were not meaningful or relevant to the study.

**User Study Participants** We recruited 90 participants from Prolific from 27 states of the USA. Of these, 60% (N = 54) were female, 38.88% (N = 35) were male, and 1.11% (N = 1) preferred not to disclose their gender. Ethnically, 50% (N = 45) identified as White, 15.56% (N = 14) as Black, 11.11% (N = 10) as Asian, 14.44% (N = 13) as Mixed, and 8.89% (N = 8) did not disclose their ethnicity. Employment status was as follows: 36.67% (N = 33) worked full-time, 8.89% (N = 8) part-time, 12.22% (N = 11) were unemployed, and 42.13% (N = 38) did not disclose their status. Participants' ages ranged from 19 to 66 years, with 20 participants aged 19-25, 40 aged 26-35, 21 aged 36-45, and 9 aged over 45.

## Results: Study 1

Our objective in study 1 was to understand the impact of contextually aware explanations of synthetic/altered images on the broader populations. To this end, we analyzed the following factors: 1) whether contextual explanations helped people recall the topic of the synthetic image better than the non-contextual, control explanation, 2) how well the contextual explanations helped participants examine why the image was altered, 3) whether the contextual explanations motivated the audience to change their prior opinion on the topic or the personalities shown in the image, and 4) why/why not participants decided to change their prior opinion on the topic or the personlaities shown in the synthetic image. We answered these questions in this section.

**Recalling The Synthetic/Altered Images** We hypothesized that contextually relevant explanations would strengthen recall of synthetic/altered images compared to the basic control explanation. To test this, three authors independently coded 20% of the Q1 responses, created initial code lists, and resolved disagreements through discussion to form a combined list. No new codes emerged, supporting the reliability of the original scheme. The authors then independently coded the remaining 80%, achieving a Cohen's Kappa of 0.84, indicating substantial agreement. Each response was reviewed to determine whether it successfully described the synthesized/altered section of the image, marking it as 1 if correct and 0 if not. This allowed for quantitative analysis.

A Chi-square test of independence was conducted to assess if the type of explanation affected recall of synthetic images. The independent variable was the type of explanations (contrastive, teleological, control), and the dependent variable was recall. All expected cell frequencies were greater

than five, with the minimum expected count being 51.33, indicating sufficient data for the Chi-square test.

Out of 900 instances, participants did not recall the synthetic images in 154 cases (17.1%) but recalled them in 746 cases (82.9%). The control condition had the highest recall failure rate at 45.5%, while Contrastive and teleological had 22.7% and 31.8%, respectively. Table 2 presents the breakdown of these counts. A Chi-square test showed a significant association between the type of explanation and memory recall, $\chi^2(2) = 14.59, p < .001$, with a moderate effect size (Cramer's V = 0.327). This suggests that the type of explanation (Contrastive, teleological, or Control) influenced participants' ability to recall the context of synthetic/altered images.

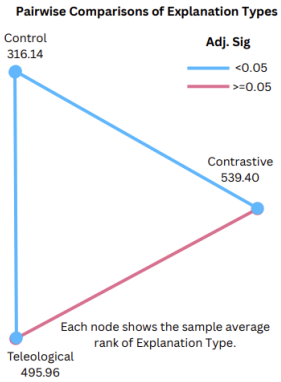


Figure 4: Pairwise Comparisons of Explanation Type in Creating Context

| Explanation Type | Recall | Not Recall |
|---|---|---|
| Contrastive | 260 | 40 |
| Teleological | 250 | 50 |
| Control | 225 | 75 |

Table 2: Recall of Synthetic Multimedia Content by Type of Explanation

With an adjusted residual of 3.1, more participants than expected remembered the altered part of the images after viewing the contrastive explanations, indicating they were effective for recall. The teleological explanations had a neutral effect (adjusted residual 0.4), while the control explanation had a negative residual of -3.5, showing it was least effective in aiding memory recall.

In summary, contrastive explanations were strongly associated with recalling the altered parts of the synthetic images. The control explanation was least effective, while the teleological explanation had a neutral effect on recall.

**Context Associated With Types of Explanations** We aimed to assess how well the types of explanations helped participants understand the context of synthetic images and their potential motivations (Q2). We compared participants' responses with the explanations and calculated cosine similarity to evaluate how well participants could explain the context. Normality tests (Shapiro-Wilk) revealed non-normal distributions for all explanation types ($p < .001$). As a result, we used the Kruskal-Wallis H test to examine the association between explanation types and context understanding. The Kruskal-Wallis test showed a significant difference in cosine similarity across explanation types ($p < .001$), indicating that explanation type significantly impacted participants' understanding of the image context.

We conducted pairwise comparisons using the Kruskal-Wallis H Test. The comparison between control and teleological yielded a significant result (test statistic = 8.480, $p < .001$). Similarly, the comparison between control and contrastive also showed a significant difference (test statistic = 10.528, $p < .001$). However, the comparison between teleological and contrastive yielded a test statistic of 2.048 (p = .041), but after applying the Bonferroni correction, the result was not significant ($p = .122$). Overall, significant differences were observed between the control group and both the teleological and contrastive explanations, though the latter two were not significantly different after adjustment, as shown in Figure 4.

**Change of Opinion** We asked participants about their prior opinions on the topics or personalities shown in the synthetic images and whether the explanations influenced any change in their opinion. We hypothesized that the control explanation would lead to more opinion changes, while contrastive and teleological explanations would encourage more objective thinking and thus discourage opinion changes based on unreliable information.

We compared responses from Q3 and Q4 to identify how many participants changed their opinions. A Chi-square test of independence showed that of the 159 opinion changes, 26.4% (N = 42) were exposed to contrastive explanations, 29.4% (N = 46) to teleological, and 42.1% (N = 67) to control explanations.

There was a significant association between explanation type and opinion change, $\chi^2(2) = 7.471, p < .001$, with moderate strength (Cramer's $V = 0.391$). Contrastive explanations led to fewer opinion changes than expected (adjusted residual = -2.0), suggesting reduced influence. Teleological explanations had a neutral effect (adjusted residual = -0.6), while control explanations resulted in significantly more opinion changes (adjusted residual = 2.6), indicating the strongest impact. We controlled for potential confounding by recruiting equal numbers of Democratic and Republican participants using Prolific filters. All participants viewed the same set of synthetic images featuring both Democratic and Republican figures. Although Democratic-leaning participants showed a slightly higher likelihood of changing their opinions, the difference was not statistically significant.

**Reasons for Not Changing/Changing Opinions** Finally, participant explained their point-of-views for changing or not changing their opinions due to the explanations. The first two authors analyzed all the responses provided by the participants and identified the primary reasons for changing or not changing the existing opinions. In this section, we briefly discuss these reasons.

**Knowledge of the images being synthetic (reason for not changing opinions):** Many participants (N = 741) didn't change their opinion after recognizing the image was altered. The contextual explanations helped them understand the motive behind the alteration and process the information more objectively. Those exposed to contrastive explanations found them insightful, realizing how easily images could be altered and the harm it could cause. Some (N = 37) noted the newness of contextual knowledge on altered content, and the explanations drew their attention to the synthetic images, even without specific instructions. In This scenario the consumer encountered with the synthetic content for the first time. But with proper context they juxtaposed the context and image together, with no need of changing opinion.

**Indifference due to minimal prior knowledge (reason for not changing opinions):** Some participants (N = 78) lacked prior knowledge about topics or individuals in the synthetic images, such as Antifa and Sarah Palin, and remained neutral in their opinions. Even after seeing the explanations, they showed indifference towards topics they didn't care about, struggling to fully grasp the contextual explanations due to a lack of background knowledge. As P3 noted:

> *"To be honest, I have no opinion because I have no information or knowledge about Antifa in general, nonetheless Antifa protests." [P3]*

In this case the people did not change there opinion as they did not have any prior opinion or knowledge about the topic portrayed in the synthetic image and did not encounter the synthetic content before the study. That is why, when they got to see the synthetic image with explanations they did not care to have any opinion.

**Political affiliations (reason for not changing opinions):** A significant number of participants (N = 268) maintained their prior opinions due to strong and deeply held political beliefs. Contrastive and teleological explanations helped them recognize that online communities might deliberately work to influence their views, encouraging more critical engagement with social media. While these contextual explanations clarified broader objectives behind image alteration, the information was too brief to shift entrenched political positions. Instead, participants felt reaffirmed in their beliefs, though they expressed interest in learning more about synthetic and altered political content on social media, indicating strong political affiliation remained a significant barrier to opinion change.

**Discovering the Truth/Clarification of Misconception (reason for changing opinions):** A key reason many participants (N = 81) changed their prior opinions was that they had previously encountered synthetic images on social media and believed them to be true. These images had shaped misconceptions about political events and figures. Our study helped them recognize these images as synthetic, and the contextual explanations enabled them to overcome their long-held misconceptions. As a result, they changed their opinions, especially after realizing they had lacked accurate baseline information. As one participant mentioned about the synthetic image of a tampered voting ballot:

> *"I remember how shitty that whole thing was, the aftermath of the election, the country became even more divisive. I truly believed that voter fraud happened during the presidential elections last time. Now I seriously started doubting that belief. I have a very unfavorable opinion of tampering with ballots, and even more so having seen the explanation, it just makes me more disgusted with all of that." [P28]*

In this case the participant encountered the synthetic image before taking part in the study and changed their opinion. But when they took part in the study and understood the context to changed their opinion in the right way neglecting the effect of the creation of the synthetic content.

### Summary: Study 1

Contextual explanations outperformed control explanations in recall, context construction, and opinion stability. As hypothesized, participants exposed to control explanations were more likely to change their opinions. These results underscore the potential of contextual explanations to enhance users' understanding and judgment when evaluating altered or synthetic images. However, realizing their full potential and broader impact requires the ability to scale such interventions to large platforms, such as social media. Study 2, detailed in the following section, addresses this challenge by assessing the feasibility of deploying contextual explanations at scale through collaboration between large language models (LLMs) and human contributors.

## Study 2: Comparison of Explanations Generated by Crowd Workers, Experts, and Large-Language Models

In the second phase, we examined how well experts, crowd workers, and Large Language Models (LLMs) can generate contextually relevant explanations with minimal instructions and training. This part of the research, illustrated as Study 2 in Figure 5, is designed to address RQ2. The following subsections briefly present the methodology and key findings.

### Methodology: Study 2

The second phase aimed to explore scalable methods for generating contextual explanations for synthetic and altered images. We considered three options: 1) crowd workers, 2) experts, and 3) large language models. Crowd workers were recruited from Prolific, experts were graduate journalism students (trained in fact-checking and argument presentation), and large language models were used to generate explanations for all images in our dataset.

Crowd workers (N = 40) received two example synthetic images with one type of contextual explanation (either contrastive or teleological). They also received the definition of their assigned explanation type. Each worker generated one type of explanation for the remaining 13 images accompanied with basic descriptions (control explanation). On average, crowd workers spent 24.80 minutes (SD = 4.31) and were compensated $6. Like phase 1, in phase 2 we performed quality control checks on all responses of crowd

worker before analysis. Overall, 4.8% of responses were rejected for quality concerns.

Experts (N = 2) received two examples with both types of explanation and then generated both types for the remaining 13 images. They were also provided with basic explanations (control) for context. Experts took around 90 minutes to complete the task and received $50.

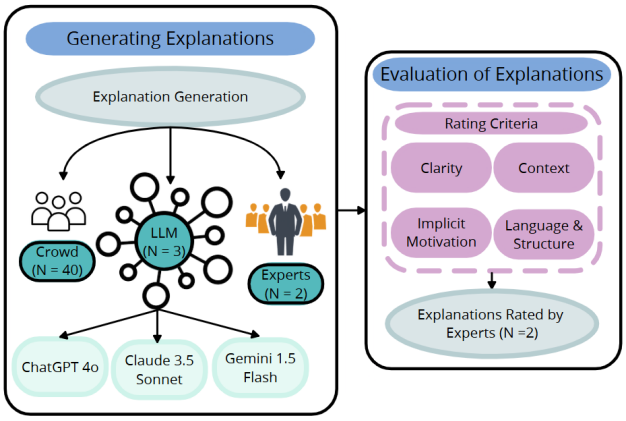


Figure 5: The flow diagram for study 2, which aims to examine how contextual explanations can be generated for synthetic and altered images at a large scale

In the final step, we used large language models (LLMs) to generate both types of explanations for all 13 synthetic/altered images. Similar to the experts and crowd workers, LLMs received two example images with both types of explanations as prompts. We used GPT-4, Claude 3.5 Sonnet, and Gemini 1.5 Flash for explanation generation. The prompt to generate Contrastive explanation can be found in https://tinyurl.com/2p3up232.

**Crowd Workers' Demographic Status** In the second phase, we recruited 40 crowd workers from 17 states in the USA. Of these, 55% (N = 22) were female and 45% (N = 18) male. Ethnically, 47.5% (N = 19) were White, 20% (N = 8) Black, 10% (N = 4) Asian, 12.5% (N = 5) mixed race, and 10% (N = 4) did not disclose their ethnicity.. 30% (N = 12) worked full-time, 12.5% (N = 5) part-time, 10% (N = 4) were unemployed, and 37.5% (N = 15) did not disclose. Ages ranged from 19 to 58, with 10 participants aged 19-25, 19 aged 26-35, 6 aged 36-45, and 5 over 45.

**Measurements for rating Explanations** To evaluate the quality of explanations generated by crowd workers, experts, and LLMs, we recruited two graduate journalism students as blind raters, ensuring they were unaware of the explanation sources to prevent bias. The rating process began with an in-person meeting where the raters met the research team and each other. The team outlined the task objectives and collaboratively reviewed a set of 50 sample explanations from all sources. Together, they established grading criteria focused on clarity, contextualization, implicit motivation (why the image was synthesized), and language and structure (fluency, complexity, coherence). The raters then independently evaluated all explanations and later participated in an exit interview to reflect on their experience and challenges. Each rater was compensated $100 for their work.

## Results: Study 2

Study 2 of the project aimed to determine how to allocate resources for generating contextually relevant explanations for synthetic content on social media. Explanations were created by three groups: 1) crowd workers, 2) experts, and 3) large language models (LLMs). Table 3 presents the average readability scores and word counts of the explanations. Crowd explanations were the easiest to read, while LLM explanations were the hardest. Additionally, crowd explanations were the shortest, and LLM explanations were the longest. The distribution of readability scores and word counts across the groups is shown in Figure 6.

| **Word Count** | | | |
|---|---|---|---|
| **Source** | **Mean Score** | **Std Dev** | **Range** |
| Crowd | 46.58 | 24.33 | 7 - 134 |
| Expert | 84.69 | 38.40 | 39 - 243 |
| LLM | 116.73 | 43.00 | 55 - 225 |
| **Readability** | | | |
| **Source** | **Mean Score** | **Std Dev** | **Range** |
| Crowd | 10.54 | 3.53 | 3.3 - 19.6 |
| Expert | 13.65 | 2.42 | 7.8 - 21.0 |
| LLM | 14.93 | 1.95 | 9.6 - 19.9 |

Table 3: The table shows word counts and readability scores for explanations by crowd workers, experts, and LLMs. Crowd explanations were shortest and most readable, suitable for a 10th-grade level, while expert and LLM explanations were longer and required college-level comprehension—indicating crowd explanations are easier to read.

All generated explanations were rated by expert raters to evaluate if one group's explanations were significantly better than the others. Explanations were rated across four criteria: 1) clarity, 2) context, 3) implicit motivation and 4) language and structure. Ratings were done on a 5-point Likert scale (1 = "extremely poor," 5 = "extremely good"). Cohen's Kappa was used to measure inter-rater agreement, and a good agreement was found between the two expert raters, with $\kappa = 0.693$ (95% CI, 0.31 to 0.64), $p < 0.05$.

Next, We conducted a one-way MANOVA to assess which group generated better explanations based on expert raters' evaluations. The Shapiro-Wilk test indicated significant deviations from normality ($p < 0.05$), allowing us to reject the null hypothesis. Wilks' Lambda score showed a significant effect of the explanation type ($p < 0.001$, partial eta-squared values ranging from 0.228 to 0.454), indicating that the explanation type influenced the results.

In terms of creating context crowd explanations had significantly lower ratings than expert explanations (mean difference = -1.156, $p < .001$) and LLM explanations (mean difference = -1.708, $p < .001$). Expert explanations were

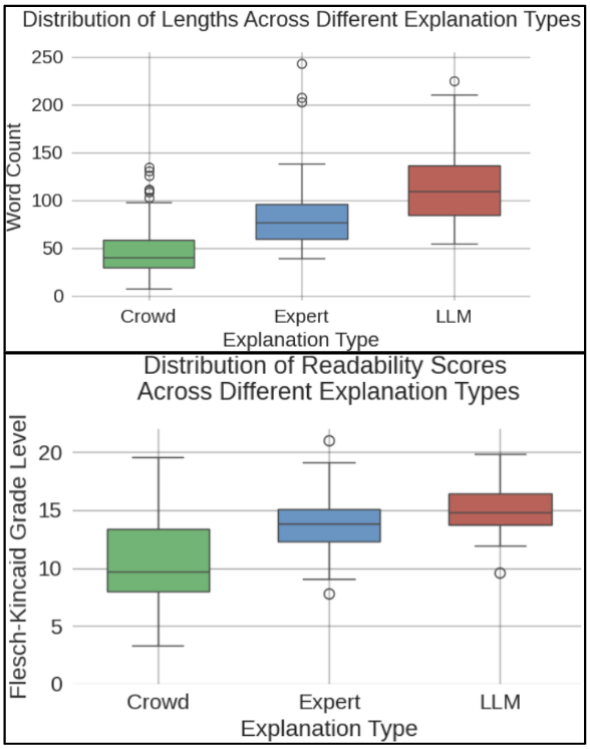


Figure 6: Distribution of the average Word Length (Top) & average Readability Score (Bottom) Across Different Explanation Types

rated higher than LLMs, though the difference was smaller (mean difference = -0.551, $p = .007$). Then in terms of implicit motivation crowd explanations had significantly lower ratings than expert explanations (mean difference = -1.156, $p < .001$) and LLM explanations (mean difference = -1.708, $p < .001$). Expert explanations were rated higher than LLMs, though the difference was smaller (mean difference = -0.551, $p = .007$).

After that in terms of clarity crowd explanations were rated significantly lower than expert explanations (mean difference = -0.908, $p < .001$) and LLM explanations (mean difference = -1.301, $p < .001$). There was no significant difference between expert and LLM explanations ($p = 0.080$), indicating similar clarity scores. Then in terms of language & structure crowd explanations were rated lower than expert (mean difference = -1.091, $p < .001$) and LLM explanations (mean difference = -1.542, $p < .001$). Expert explanations were rated lower than LLMs (mean difference = -0.451, $p = .025$). The detailed results can be found in https://tinyurl.com/2p3up232.

### Summary: Study 2

Study 2 found that crowd-generated explanations consistently underperformed compared to those from experts and LLMs across all evaluation criteria: clarity, contextual relevance, implicit motivation, and language & structure. LLMs outperformed crowd workers by a notable margin and were comparable to expert-written explanations in terms of clarity. However, experts maintained a slight advantage in providing richer context and more nuanced language use. These findings suggest that while crowd explanations are simpler and more readable, expert and LLM explanations are perceived as more comprehensive and better structured.

## Discussion

### Comparison Between Types of explanations

Our analysis showed that contrastive explanations were significantly more effective than teleological ones in helping participants understand the context of synthetic images. This may be because contrastive explanations explicitly identify why a particular situation is misleading or altered, whereas teleological explanations rely on broader background knowledge about the characters or scenario. These findings align with prior research emphasizing the role of clarity and framing in addressing misleading content in generative media (Xu, Fan, and Kankanhalli 2023; Weikmann and Lecheler 2023), as well as with work on counterfactual reasoning as a tool for reflective engagement (Dai et al. 2022).

These results also suggest that explanation types should be tailored to audience needs. Contrastive explanations are well suited for general users who may lack prior knowledge or interest in the topic, while teleological explanations may be more engaging for users seeking deeper insight. To accommodate diverse audiences, platform designers could implement a layered explanation structure—presenting contrastive explanations by default, with teleological elaborations as optional, on-demand content.

### Ineffectiveness of Control Type Explanation

Our study found that control explanations were largely ineffective in conveying the context behind synthetic images or the motives for their manipulation. Participants with pre-existing misconceptions were rarely persuaded, and many disregarded the control explanations altogether, remaining influenced by the fabricated content. This supports the view that not all explanations are equally effective in mitigating the impact of synthetic media. From a platform design perspective, control explanations function much like a minimal "synthetic content" tag, which prior work has shown to be frequently ignored (Moravec, Dennis, and Minas 2020a). Given their low production effort and lack of contextual grounding, such explanations may be perceived as untrustworthy or even adversarial, especially by skeptical users. Overuse of ineffective explanations could also reduce user engagement by contributing to information fatigue and eroding perceived credibility (Bourdieu 1986).

### Long-Term Effect of Explanations

This study evaluated explanation effectiveness within a single-session design; longitudinal effects were beyond its scope. In real-world settings, however, repeated exposure to contextual explanations may influence users' vigilance toward synthetic content. Importantly, the impact of such explanations may vary depending on broader sociopolitical

contexts—such as elections or international conflicts—that shape public interpretation. As a result, platforms and policymakers must carefully consider how and when to deploy contextual explanations, given their potential to both inform and inadvertently destabilize public discourse.

### The Implications of Using LLMs for Generating Explanations

Our study highlights the potential of off-the-shelf LLMs to generate contextual explanations for synthetic images. As AI-generated images (Yang, Singh, and Menczer 2024) increasingly challenge social platforms in the fight against inaccurate claims, LLM agents may offer a scalable solution. While LLM-generated explanations underperformed expert ones in areas such as contextual nuance and implicit motivation, these limitations could be mitigated through prompt tuning, adjusted loss functions, or integration of custom RAG modules. However, LLM-generated content can exhibit skewness, potentially favoring certain communities and producing controversial explanations. We therefore caution against relying solely on LLMs. Instead, we recommend a human-AI collaboration model, where an expert audit board routinely reviews LLM outputs, ensuring accountability and strengthening the trustworthiness of explanations.

### Trusting AI-agents on Socially Sensitive Issues

Lastly, public opinion remains divided on the use of AI agents in socially sensitive contexts (Binns 2018). One group views AI agents as untrustworthy and opaque due to their black-box nature and lack of human oversight, while others argue that AI, free from human bias, can offer greater neutrality. As researchers, we are not positioned to fully endorse either view. Further investigation is needed to identify factors influencing trust and perceived reliability of AI agents, particularly in sensitive domains. Social platforms may benefit from partnering with academic researchers to pursue deeper, evidence-based insights in this area.

Prior work has highlighted synthetic content through labeling (Ruffin, Wang, and Levchenko 2023). Building on this, we introduce contextual explanations aimed at more effectively refuting synthetic images. For each manipulated image, we generated contrastive and teleological explanations, along with a control condition lacking contextual detail. Contextual explanations help individuals construct conceptual frameworks that enhance comprehension and memory. According to Bartlett's schema theory (Bartlett 1932), people integrate new information into schemas based on prior knowledge. Contextual cues facilitate this process (Bransford and Johnson 1972), consistent with Mayer's multimedia learning theory (Mayer 2002), which emphasizes linking unfamiliar material to known concepts. Such context also supports recall and deepens engagement by making content more vivid and relatable (Mayer 2002; Schwarz and Tormala 2006; Kendeou and van den Broek 2007; Bransford and Schwartz 2000).

### Contextual Explanation Escalate Recall

In our study, contrastive and teleological explanations served as contextual forms, while the control condition lacked necessary context. Contrastive explanations significantly improved recall of synthetic images, whereas control explanations were associated with poor retention. Teleological explanations had a neutral effect, aligning with prior work suggesting that contextual framing supports memory by situating information within meaningful narratives. These results suggest that emphasizing the topics involved in synthetic content can strengthen viewers' understanding of both the context and purpose behind alterations.

This finding is consistent with previous research showing that contextual explanations strengthen recall by fostering mental models that aid knowledge organization and retrieval (Gentner and Stevens 1983; Mayer 2002; Bain 2000; Carroll 1997). However, while teleological explanations may resonate with well-informed users, they posed challenges for crowd workers, who struggled to learn and remember their content. This highlights the importance of tailoring explanation design to the audience's level of expertise. It also points to the need for personalized explanatory strategies, rather than relying on a one-size-fits-all approach.

## Limitations

Our study has limitations, particularly in terms of participants' exposure to synthetic images during the user study. This may lead to long-term memories that are difficult to disprove. Second, detailed contextual explanations may unintentionally shape participants' perspectives, distorting their understanding of similar content found online. This raises concerns about the long-term consequences, especially for those who want to avoid misleading content. While participants were warned about potentially violent content in the study, this notice may not have been enough to alleviate emotional distress, as individual reactions can vary greatly. Finally, though synthetic multimedia content circulates in form of images and videos; for the scope of this study we solely focused on inaccurate claims spreading through synthetic images. We will address the misleading content circulating through videos in future work. Future work will explore contextual analysis methods to better capture semantic similarity.

## Conclusion

In today's digital landscape, the prevalence of multimedia content—both real and synthetic—complicates users' ability to discern misleading content. Once misleading content circulates widely, it becomes difficult to retract. To address this, we developed contextual explanations for synthetic images and evaluated their effectiveness through a user study. Results show that contextual explanations consistently outperformed non-contextual ones. We further compared explanations generated by crowd workers, journalism experts, and large language models (LLMs), finding that experts and LLMs produced higher-quality explanations. These findings highlight the potential for contextual explanations to serve as scalable, platform-level intervention against misleading content.

## Ethical Statement

Our dataset included synthetic and real images representing a range of sociopolitical themes, including movements such as Antifa and Black Lives Matter, as well as political personalities. Care was taken to avoid including graphic, excessively violent, or gory imagery. All images were curated to minimize potential psychological harm while still representing content typical of altered and synthetic media in online environments.

Participants were fully informed that the study involved synthetic media and were explicitly told that their participation was voluntary. Prior to beginning the task, they were provided with a clear consent form outlining the purpose, procedures, and potential risks of the study. They were also debriefed that the study was for research purposes only and they had the right to withdraw at any time without penalty. Additionally, participants were given the option to contact the research team should they have any concerns or questions during or after their participation.

All study procedures were approved by the authors' Institutional Review Board (IRB) committee and were conducted in accordance with established guidelines for human subjects research.

As researchers situated at the intersection of human-computer interaction, AI ethics, and social media studies, we acknowledge the influence of our disciplinary training, geographic location, and sociocultural context on the framing and interpretation of this work. This study was conducted from an academic perspective based in the Global North, with access to institutional support and computational resources that may not be universally available. While we aim to improve the interpretability and trustworthiness of AI-generated explanations in combating synthetic and altered media, we recognize that our understanding of "effective" or "accessible" explanations is shaped by Western epistemologies and educational norms.

We designed the study with a commitment to minimize harm, promote information equity, and empower users through transparent and explainable AI systems. However, we acknowledge the limits of our perspective, particularly in accounting for how explanation strategies can resonate differently across cultures, languages, or socioeconomic contexts. We encourage further research that centers marginalized voices and incorporates participatory approaches to better contextualize the implications of AI in diverse settings.

Our work aims to improve public understanding of synthetic and altered media through AI-generated explanations. We understand that the findings of this work may also introduce unintended consequences. First, contextual explanations, if interpreted wrongly or poorly designed, could reinforce existing perceptual slants or create false confidence in misleading content. Additionally, reliance on large language models raises concerns about skew in training data, which may lead to explanations that unintentionally favor certain sociopolitical perspectives. Large language Models can also hallucinate while generating desired explanations. In addition, there is also a risk that adversarial actors could mimic these explanatory strategies to lend false legitimacy to deceptive media. To mitigate such risks, we advocate for human-AI collaboration, regular auditing, and transparency in explanation generation pipelines.

## References


2024. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

2024. https://www.nbcnews.com/tech/misinformation/kamala-harris-deepfake-shared-musk-sparks-free-speech-debate-rcna164119. NBCNews.com.

2025. Prolific. https://www.prolific.com.

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2): 211–236.

Atanasova, P. 2024. *Accountable and explainable methods for complex reasoning over text*. Springer.

Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G. L.; Corney, D.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.; et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8): 852–863.

Ayyub, R. 2018. I was the victim of a deepfake porn plot intended to silence me.

Bain, R. B. 2000. Into the breach: Using research and theory to shape history instruction. In Stearns, P.; Seixas, P.; and Wineburg, S., eds., *Knowing, teaching, and learning history*, 331–352. New York University Press.

Bartlett, F. C. 1932. Remembering: A Study in Experimental and Social Psychology.

Beggs, A.; and Graddy, K. 2009. Anchoring effects: Evidence from art auctions. *American Economic Review*, 99(3): 1027–1039.

Bertrand, A.; Belloum, R.; Eagan, J. R.; and Maxwell, W. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 78–91. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Bessi, A.; et al. 2016. Socio-technical interactions in online social networks: The case of fake news. *PloS One*, 11(4).

Binns, R. 2018. On the Meaning of 'Trust' in AI and the Impact of AI on Trust. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. New York, NY, USA: ACM.

Binns, R. 2020. Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 149–158.

Bird, C.; Ungless, E.; and Kasirzadeh, A. 2023. Typology of Risks of Generative Text-to-Image Models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 396–410. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Blair, S.; Ecker, U. K. H.; Cook, J.; and Lewandowsky, S. 2017. Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. *Psychological Science in the Public Interest*, 17(3): 107–116.

Bourdieu, P. 1986. *The Forms of Capital*. New York: Cambridge University Press. Originally published in French as "Les formes de capital" in 1983.

Brandolini, A.; et al. 2020. The Impossibility of Truth: Brandolini's Law in Social Media. *International Journal of Media Management*, 22(1): 1–10.

Bransford, J. D.; and Johnson, M. K. 1972. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11: 717–726.

Bransford, J. D.; and Schwartz, D. L. 2000. Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 25: 61–100.

Cai, Y.; and et al. 2022. The impact of deep fakes on trust and privacy: Implications for policy. *Cybersecurity*, 1(1): 1–10.

Carroll, J. M. 1997. *Minimalism beyond the Nurnberg funnel*. MIT Press.

Chan, M.-p. S.; Jones, C. R.; Hall Jamieson, K.; and Albarracín, D. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, 28(11): 1531–1546.

Cheetham, K. D.; and Joshua. 2023. Fake trump arrest photos: How to spot an AI-generated image.

Chesney, B.; and Citron, D. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107: 1753.

Chou, H.-T. G.; Gaysynsky, A.; and Vanderpool, R. C. 2018. The fake news effect: Experiments on motivated reasoning, credibility assessment, and misinformation in social media. *Health Communication*, 33: 1075–1083.

Chuai, Y.; Tian, H.; Pröllochs, N.; and Lenzini, G. 2024. Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Cook, J.; Ecker, U. K. H.; and Lewandowsky, S. 2017. Misinformation and How to Correct It: The Strategic Case for Stealthy Corrections. *Psychological Science in the Public Interest*, 18(3): 163–169.

Dai, S.-C.; Hsu, Y.-L.; Xiong, A.; and Ku, L.-W. 2022. Ask to know more: Generating counterfactual explanations for fake claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2800–2810.

Debroy, O.; and Hemmige, D. 2024. Psycho-social Impact of Deepfake Content in Entertainment Media. 10: 2455–0620.

Dunn, S. 2021. Women, not politicians, are targeted most often by deepfake videos. *Centre for Internal Governance Innovation*.

Ecker, U. K.; O'Reilly, Z.; Reid, J. S.; and Chang, E. P. 2020. The effectiveness of short-format refutational fact-checks. *British journal of psychology*, 111(1): 36–54.

Ecker, U. K. H.; Lewandowsky, S.; and Tang, D. T. W. 2010. Explicit Warnings Reduce but Do Not Eliminate the Continued Influence of Misinformation. *Memory Cognition*, 38(8): 1087–1100.

Fabbri, M. 2023. Self-determination through explanation: an ethical perspective on the implementation of the transparency requirements for recommender systems set by the Digital Services Act of the European Union. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 653–661. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Fawzi, M.; and Magdy, W. 2024. "Pinocchio had a Nose, You have a Network!": On Characterizing Fake News Spreaders on Arabic Social Media. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Franklin, D.; et al. 2017. The fake news epidemic: What can be done? *Harvard International Review*, 39(1): 12–16.

Friedman, D. 2014. The visual influence of graphs: A review of the evidence and its implications for scientific communication. *Science Communication*, 36(1): 63–92.

Furnham, A.; and Boo, H. C. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1): 35–42.

Gentner, D.; and Stevens, A. L. 1983. *Mental models*. Erlbaum.

Guzzetti, B. J.; Snyder, T. E.; Glass, G. V.; and Gamas, W. S. 1993. Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly*, 117–159.

Hankinson, R. J. 2001. *Cause and explanation in ancient Greek thought*. Oxford University Press.

Hawkins, G. 2018. The role of visual representations in scientific communication: Implications for understanding and persuasion. *Visual Communication Quarterly*, 25(1): 45–56.

Heidari, A.; Jafari Navimipour, N.; Dag, H.; and Unal, M. 2024. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2): e1520.

Henry, N.; McGlynn, C.; Flynn, A.; Johnson, K.; Powell, A.; and Scott, A. J. 2020. *Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery*. Routledge.

Hesslow, G. 1988. The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, 11–32.

Hilton, D. J. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1): 65.

Hilton, D. J. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4): 273–308.

Hilton, D. J.; and John, L. M. 2007. The course of events: counterfactuals, causal sequences, and explanation. In *The psychology of counterfactual thinking*, 56–72. Routledge.

Hilton, D. J.; and Slugoski, B. R. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, 93(1): 75.

Hladík, R.; and Štětka, V. 2017. The powers that tweet: Social media as news sources in the Czech Republic. *Journalism studies*, 18(2): 154–174.

Hume, D. 2007. An enquiry concerning human understanding and other writings.

Hunt, W. A. 1941. Anchoring effects in judgment. *The American Journal of Psychology*, 54(3): 395–403.

Jahanbakhsh, F.; Zhang, A. X.; Berinsky, A. J.; Pennycook, G.; Rand, D. G.; and Karger, D. R. 2021. Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Jahanbakhsh, F.; Zhang, A. X.; and Karger, D. R. 2022. Leveraging structured trusted-peer assessments to combat misinformation. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW2): 1–40.

Kaustia, M.; Alho, E.; and Puttonen, V. 2008. How much does expertise reduce behavioral biases? The case of anchoring effects in stock return estimates. *Financial Management*, 37(3): 391–412.

Kendeou, P.; and van den Broek, P. 2007. The effects of text coherence on comprehension: Evidence from eye movements. *Discourse Processes*, 43(1): 19–36.

Kendeou, P.; Walsh, E. K.; Smith, E. R.; and O'Brien, E. J. 2014. Knowledge revision processes in refutation texts. *Discourse Processes*, 51(5-6): 374–397.

Khalifa, N.; Anjum, M.; and Qu, Z. J. 2024. The Harmful Impact of Fake Images in Local Societies: A Case Study and the Path to Regulation. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 5(1): 98–108.

Khan, S. O.; Ghafourian, T.; and Patil, S. 2024. Targets of Weaponized Islamophobia: The Impact of Misinformation on the Online Practices of Muslims in the United States. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Kowalski, P.; and Taylor, A. K. 2009. The effect of refuting misconceptions in the introductory psychology class. *Teaching of Psychology*, 36(3): 153–159.

Leibowicz, C. R.; McGregor, S.; and Ovadya, A. 2021. The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 736–744. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Lewandowsky, S.; Ecker, U. K. H.; and Cook, J. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3): 106–131.

Lewis, D. 2000. Causation as influence. *The Journal of Philosophy*, 97(4): 182–197.

Li, Y.; and Xie, Y. 2020. Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of marketing research*, 57(1): 1–19.

Lipton, P. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27: 247–266.

Lodge, T.; et al. 2022. The effects of fake news on stock prices: Evidence from social media. *Journal of Financial Economics*, 145(2): 497–513.

Lyons, K. 2017. Facebook is testing new ways to combat fake news and misinformation. *The Verge*.

Lyu, Y.; Liang, P. P.; Deng, Z.; Salakhutdinov, R.; and Morency, L.-P. 2022. DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 455–467. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Maas, M. M.; and Liao, S. 2019. AI and the ethics of misinformation: implications for individuals and society. *AI and Society*, 35: 727–737.

Maddocks, S. 2020. 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political'deep fakes. *Porn Studies*, 7(4): 415–423.

Malki, L. M.; Patel, D.; and Singh, A. 2024. "The Headline Was So Wild That I Had To Check": An Exploration of Women's Encounters With Health Misinformation on Social Media. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Malle, B. F. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

Maras, M.-H.; and Alexandrou, A. 2019. Determining authenticity in video forensics: The problem with deepfakes. *Computer Law Security Review*, 35(3): 238–244.

Mayer, R. E. 2002. Multimedia learning. In *Psychology of learning and motivation*, volume 41, 85–139. Elsevier.

Melchior, C.; and Oliveira, M. 2024. A systematic literature review of the motivations to share fake news on social media platforms and how to fight them. *new media & society*, 26(2): 1127–1150.

Messaris, P.; and Abraham, K. 2001. The Role of Images in the Communication of Information. *Journal of Communication*, 51(4): 479–497.

Meyer, M.; and Schwabe, L. 2020. How visualizations shape perceptions of credibility. *Cognitive Research: Principles and Implications*, 5(1): 23.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Mittelstadt, B.; Russell, C.; and Wachter, S. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 279–288. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Moravec, P.; Dennis, A.; and Minas, R. 2020a. Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense at All. *Proceedings of the 2020 International Conference on Information Systems (ICIS)*.

Moravec, P. A.; Dennis, A. R.; and Minas, R. K. 2020b. The Perception of Fake News on Social Media: The Role of Visual Literacy. *Journal of Information Technology*, 35: 201–215.

Mutlu, E. c.; Yousefi, N.; and Ozmen Garibay, O. 2022. Contrastive Counterfactual Fairness in Algorithmic Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 499–507. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

O'Brien, J.; and Lauer, J. 2018. Misleading Data Visualizations: How Graphical Distortion Impacts Perception. *International Journal of Data Visualization*, 3(1): 45–62.

Olan, F.; Jayawickrama, U.; Arakpogun, E. O.; Suklan, J.; and Liu, S. 2024. Fake news on social media: the impact on society. *Information Systems Frontiers*, 26(2): 443–458.

Ruffin, M.; Wang, G.; and Levchenko, K. 2023. Explaining why fake photos are fake: Does it work? *Proceedings of the ACM on Human-Computer Interaction*, 7(GROUP): 1–22.

Saini, A.; and Prasad, R. 2022. Select Wisely and Explain: Active Learning and Probabilistic Local Post-hoc Explainability. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 599–608. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Sanderson, J. A.; and Ecker, U. K. 2020. The challenge of misinformation and ways to reduce its impact. In *Handbook of learning from multiple representations and perspectives*, 461–476. Routledge.

Schmitt, V.; Villa-Arenas, L.-F.; Feldhus, N.; Meyer, J.; Spang, R. P.; and Möller, S. 2024. The Role of Explainability in Collaborative Human-AI Disinformation Detection. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2157–2174.

Schwarz, N.; Jalbert, M.; Ecker, U.; and Zimmerman, C. 2016. Worldwide propagation of anti-vaccine sentiments and their impact on global healthcare. *Journal of Health Communication*, 21: 17–23.

Schwarz, N.; and Tormala, Z. L. 2006. The constraining effects of context on interpretation. *Psychological Science*, 17(1): 33–38.

Shu, K.; and et al. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, 19(1): 22–36.

Simonson, I.; and Drolet, A. 2004. Anchoring effects on consumers' willingness-to-pay and willingness-to-accept. *Journal of consumer research*, 31(3): 681–690.

Teubner-Rhodes, S.; Vaden Jr, K. I.; Dubno, J. R.; and Eckert, M. A. 2017. Cognitive persistence: Development and validation of a novel measure from the Wisconsin Card Sorting Test. *Neuropsychologia*, 102: 95–108.

Tippett, C. D. 2010. Refutation text in science education: A review of two decades of research. *International journal of science and mathematics education*, 8: 951–970.

Vargas-Restrepo, M.; Yang, Y.; Stanley, G.; and Marsh, E. 2019. Data Visualization and Its Role in Misinformation: Misleading Graphs and Their Impact. *Journal of Experimental Psychology: Applied*, 25(2): 185–197.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.

Walker, J.; Thuermer, G.; Vicens, J.; and Simperl, E. 2023. AI Art and Misinformation: Approaches and Strategies for Media Literacy and Fact Checking. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 26–37. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Walter, N.; and Murphy, S. T. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication monographs*, 85(3): 423–441.

Weikmann, T.; and Lecheler, S. 2023. Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, 25(12): 3696–3713.

Wendling, M. 2024. Ai can be easily used to make fake election photos - report.

Wilder, B.; and Vorobeychik, Y. 2019. Defending elections against malicious spread of misinformation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 2213–2220.

Wolfe, R.; and Mitra, T. 2024. The Impact and Opportunities of Generative AI in Fact-Checking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1531–1543.

Xu, D.; Fan, S.; and Kankanhalli, M. 2023. Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9291–9298.

Yang, K.-C.; Singh, D.; and Menczer, F. 2024. Characteristics and prevalence of fake social media profiles with AI-generated faces. *arXiv preprint arXiv:2401.02627*.