From TikTok to Telegram: Cross-Platform Efficacy and User Acceptance of Erroneous and Flawless Misinformation Interventions

Katrin Hartwig
Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt
Darmstadt, Germany
hartwig@peasec.tu-darmstadt.de

Tom Biselli
Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt
Darmstadt, Germany
biselli@peasec.tu-darmstadt.de

Franziska Schneider
Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt
Darmstadt, Germany
franziska.schneider@tu-darmstadt.de

Immanuel Lamp
Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt
Darmstadt, Germany
immanuel.lamp@stud.tu-darmstadt.de

Christian Reuter
Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt
Darmstadt, Germany
reuter@peasec.tu-darmstadt.de

Abstract

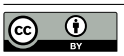
Misinformation interventions are often evaluated under ideal conditions, yet real-world systems are rarely flawless. We report on an online experiment ($N = 1,004$) comparing five state-of-the-art interventions – inoculation, accuracy prompt, community note, fact-check, and indicators – across TikTok, Telegram, and X. We examined efficacy and user perceptions under flawless and erroneous implementations. Misinformation accompanied by fact-checks and indicators was rated as significantly less accurate, while community notes showed weaker effects. Modality did not significantly influence intervention efficacy and had only minor effects on user acceptance. Community notes, fact-checks, and indicators were rated as more helpful but also more annoying than the less informative accuracy prompts. Notably, the efficacy of interventions disappeared under erroneous conditions. This highlights the crucial role of intervention quality in fostering trust and acceptance. Our findings provide (1) a cross-platform evaluation of interventions and (2) empirical evidence that accuracy and reliability are crucial in complex social media environments.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Social media**; *Empirical studies in collaborative and social computing*.

Keywords

misinformation, disinformation, fake news, intervention, nudge, indicator



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790630

ACM Reference Format:
Katrin Hartwig, Tom Biselli, Franziska Schneider, Immanuel Lamp, and Christian Reuter. 2026. From TikTok to Telegram: Cross-Platform Efficacy and User Acceptance of Erroneous and Flawless Misinformation Interventions. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3772318.3790630

1 Introduction

Most misinformation interventions are evaluated under ideal conditions. But real-world systems are rarely perfect: Community-driven interventions may be biased by the bubbles and communities of users on specific social media platforms (e.g., users on X or in Telegram channels), and also automatic credibility assessments (e.g., using machine learning) can incorrectly flag misinformation as accurate or vice versa. In a large-scale online experiment, we compare the efficacy and user perception of five erroneous versus flawless misinformation interventions across three major social media platforms, featuring short-videos on TikTok, voice messages on Telegram, and text/image posts on X.

Social media platforms play an essential role in information exchange, allowing for fast, individualized, and multimodal content dissemination. However, users on these platforms are also confronted with misleading and dangerous content. This includes deliberately misleading disinformation and false but inadvertently created and disseminated misinformation — both of which can cause significant harm, even death [40]. Despite their differences, we use the term *'misinformation'* as an umbrella term for both phenomena, following prior research to enhance readability [3, 17, 73].

Interdisciplinary research has been looking for effective and user-centered countermeasures to tackle misinformation's dangerous impact for several years. These efforts include, for example, media literacy training at schools, critical journalism, and efforts for automatic detection of harmful content. A significant body of

research in human-computer interaction (HCI) has proposed digital countermeasures that directly impact users via information presentation or withholding and often occur after a manual or automatic pre-filtering of misinformation [31]. These *user-centered misinformation interventions* come in many forms, such as warnings, corrections, or nudges. The vast majority of interventions are designed for and evaluated with text-based content, and only a few studies specifically delve into deceptive multimodal content on TikTok [30, 34], voice messages in messengers [23, 32], or assessed interventions' robustness across misinformation modalities. Research has systematized knowledge and clustered interventions and has started to derive implications on efficacy and user acceptance [31, 38]. Findings reveal that we are still far from a decisive solution, as many studies report only marginal effect sizes of interventions or even adverse effects such as spill-overs on accurate content or over-reliance [29, 30, 41, 50]. In addition, most studies evaluate interventions under idealized conditions, for instance, where warnings are only displayed if the content really is misinformation. However, systems operating in complex social media platform environments will sometimes be erroneous, for instance, showing a warning although the content is accurate (false positive) or missing a warning on misinformation (false negative). This is both the case for interventions that rely on automatic detection [57, 77] (e.g., because there is not enough information for an algorithm to decide) and manual credibility assessments [14] (e.g., due to a biased community). In light of recent shifts from professional fact-checks to community-driven approaches and in times of social media platforms contradicting common definitions of misinformation, it is even more essential for users to critically reflect not only on the content itself, but also on implemented countermeasures.

Conducting a large-scale online experiment ($N = 1,004$), our paper advances misinformation research from an HCI perspective by focusing on the efficacy and user acceptance of erroneous versus flawless misinformation interventions across modalities. This builds on related work that delved into adverse effects in the context of misinformation and research on automation biases in other contexts with decisions based on artificial intelligence (AI) [70]. We model imperfection by evaluating five state-of-the-art interventions (community notes, professional fact-checks, indicators, accuracy prompts, and inoculation) *with and without errors*. We compare the interventions across *three distinct social media platforms* in a simulated experimental environment that entails different modalities of misinformation, allowing for an assessment of the interventions' robustness. In accordance with other HCI research [30], we investigate both *efficacy* (e.g., accuracy assessment and sharing intentions) and *user perceptions* (e.g., perceived helpfulness and annoyance).

Our core contributions (C) and findings (F) are:

(C1) A quantitative evaluation of the efficacy of five misinformation interventions. We found that **(F1)** the accuracy assessment was significantly improved by professional fact-checks and indicators, and to a lesser extent by community notes, but **(F2)** all investigated interventions did not significantly reduce the sharing intentions of misinformation compared to the control groups.

(C2) A comparison of erroneous and flawless interventions. We found that **(F3)** for efficacy, it is decisive whether an intervention is erroneous, as significant changes in accuracy assessments were only found for the flawless interventions.

(C3) An investigation of perceived helpfulness and annoyance. **(F4)** community notes, professional fact-checks, and indicators were perceived as significantly more helpful but more annoying than accuracy prompts, which disrupt the natural flow of social media use by appearing, e.g., as additional pop-ups rather than being integrated into the interface. Further, **(F5)** interventions were perceived as more helpful and less annoying when they were flawless in contrast to erroneous interventions.

(C4) A cross-platform comparison of TikTok videos, voice messages on Telegram, and text-image combinations on X. We found that **(F6)** interventions' efficacy and user perception was robust regarding the modality and social media platform of misinformation content.

2 Related Work

We discuss related work on misinformation across different modalities, emphasizing the relevance of multimodal content (see Section 2.1). We further delve into digital misinformation interventions as one of many pieces of the puzzle to tackle the phenomenon (see Section 2.2), shedding light on the wide research landscape and state-of-the-art countermeasures. Then, we take a closer look at literature discussing beneficial and adverse effects of misinformation interventions (see Section 2.3). Finally, we outline the resulting research gaps and research questions that are addressed in our paper (see Section 2.4).

2.1 Multimodal Misinformation on Social Media

Misinformation and how it is presented varies significantly depending on the modality of content or social media platform on which it appears. The heterogeneity of social media platforms, characterized by a multitude of features and functionalities, fosters a substantial potential for misleading content.

Text-based misinformation remains the most extensively studied, especially on social media platforms like *X (formerly Twitter)*, historically known for its accessible research data [31]. X is widely used both by private individuals and by officials (e.g., for crisis communication), despite its current controversy and concerns regarding content moderation and political impact. Research has explored textual features for misinformation detection [64, 76]. Indeed, despite its text-heavy nature, X also incorporates multimodal elements (e.g., images or videos), though to a lesser extent than social media platforms like TikTok or Instagram. Posts on X include multiple layers that can either mislead or serve as cues for credibility assessment. For instance, posts provide information about the author [20] via a suspicious profile image or name, the number of followers, or other social media posts by the author. Further, interactions with the posts, such as likes and comments, or the content itself [21] with rhetorical strategies or links to questionable external content can be considered insightful characteristics [33]. Some characteristics of text-based misinformation are transferable to other modalities (e.g., loaded language) [30].

In contrast, contemporary social media platforms like *TikTok*, a leading video-sharing platform (VSPs), embed misinformation in richly multimodal forms. TikTok videos combine multiple content layers that may be susceptible to misdirection, including the video itself, audio material, interactions with the video, and textual

descriptions [47]. VSPs have rapidly grown in popularity, making TikTok one of the most successful social media platforms, especially for younger people, and the most downloaded app worldwide. While TikTok has been associated with positive effects like connecting communities [45, 61] or fostering creativity [48], recent events have demonstrated its vulnerability to misinformation and other harmful content [4, 9, 39, 47]. Studies have investigated creators' exposure to hate and harassment [68], how young users perceive digital safety on TikTok, including misinformation and deep fakes [25], who creates misinformation on the social media platform, and how users act on them [4]. Recent work highlights the multimodal layers of TikTok videos with their potential to mislead but also to give cues on the credibility of the content. That includes, for example, the comment section with user discussions [34, 47] or other forms of interaction such as likes, shares, or stitches, emotion-evoking characteristics (e.g., via music in the background, facial expressions, or wordings), identifiable features of AI-generated or manipulated content (e.g., additional fingers, filters), or a generally attention-grabbing layout [30].

Misinformation in voice messages — prevalent in messaging platforms such as *Telegram* and *WhatsApp* — remains underexplored, despite its growing societal impact and salient concerns, both in daily life and during times of crisis, when individuals seek information in public messenger groups [46]. Since messaging apps began supporting voice messages (e.g., in 2013 on WhatsApp [74]), they have become widely adopted [23, 75]. The ability to disseminate messages across multiple messenger groups has been shown to accelerate the spread of information [19, 58]. Speech can convey emotions that may unintentionally prompt users to spread misinformation. Research suggests that voice messages come with a unique misleading potential due to their emotion-evoking attributes in the audio material [23]. Initial insights have started looking into audio tracks of videos for the detection of misinformation [63]. Others delve deeper into voice messages by analyzing characteristics of misinformation within [15, 23, 32, 43]. For instance, they found that voice messages containing misinformation were longer on average and tend to contain more negative emotions [43]. Speech rate and sound volume have been shown to influence credibility assessment of audio material [23], though users tend to focus on the content itself (e.g., claims to be an expert or calls for action) rather than tonality or other voice-specific cues [32].

While prior work explores misinformation within individual social media platforms and modalities, few studies have examined whether interventions remain effective across them. Given the lack of research on the robustness of interventions across social media platforms and modalities, our work aims to provide initial quantitative insights into this gap.

2.2 Misinformation Interventions

Misinformation can be addressed through a range of approaches, including critical journalism, media literacy training in educational institutions, automated detection technologies [35], and user-centered interventions that exert a direct impact on users through the presentation or withholding of information [31]. Among these, digital user-centered interventions have become a central focus in HCI research [31, 38], aiming to reduce misinformation spread, alter sharing intentions, and foster critical thinking within human-centered and technology-driven frameworks. Interventions take various forms, including corrections (offered by private users, officials, or algorithms) [5, 12], as well as informational labels used to flag misinformation [8]. Research suggests that users favor interventions that offer a certain level of comprehensibility or transparency, in contrast to more opaque approaches such as binary flagging without context [27, 37].

Recent studies have begun to systematize the diverse landscape of interventions, developing taxonomies that cluster them based on design elements, timing, and the cognitive or behavioral processes they intend to influence [31, 38]. The intervention research landscape is notably heterogeneous, with most studies centered on text-based content. However, more recent work has expanded to visual and multimodal formats, including images [60], videos [30], and graphs [13].

In a large-scale study, Kozyreva et al. [38] present a comprehensive digital toolbox of state-of-the-art interventions, organized into categories such as accuracy prompts to shift users' attention to the concept of accuracy, debunking to counter false beliefs while offering corrections, inoculation to expose users with manipulation strategies before exposure to misinformation, media literacy tips, social norms like peer influence to reduce sharing intentions, source credibility labels such as ratings from 0 to 100, and warnings or fact-checking labels among other [38]. Our study builds directly on these systematic insights to select state-of-the-art interventions. We clustered and combined these categories and included those that are most commonly applied in the real world. Additionally, we ensured the inclusion of interventions that align with user preferences for clarity and transparency [27, 37], as well as complementary interventions designed to influence behavior on a more implicit or subconscious level.

2.3 Beneficial and Adverse Effects of Misinformation Interventions

Misinformation interventions vary in their approaches to influencing user behavior or content dissemination. Some aim to reduce overall misinformation spread, while others focus on enhancing users' ability to distinguish credible from false content. Some also aim to decrease the intention to share misinformation while increasing the sharing of accurate information. Related research has assessed intervention efficacy for specific contexts, including social media platforms, user demographics, and modalities [31]. Often, that involves comparisons with control groups without an intervention or state-of-the-art interventions applied on specific social media platforms. Complementary qualitative studies provide deeper insight into how users perceive and respond to interventions, examining perceptions of trust, psychological reactance, and comprehensibility [31].

Quantitatively, interventions tend to have small to medium effect sizes [e.g., 7, 71, 72]. Yet cross-study comparisons remain difficult due to substantial variations in settings (e.g., different social media platforms, modalities, participants), and ways to measure efficacy (e.g., credibility ratings vs. sharing intentions) [28, 31]. Calls for standardized efficacy metrics represent a step towards more successful interventions [28], with recent meta-analyses allowing for

initial insights, for example, comparing corrections in the context of scientific misinformation where corrections were more successful when detailed [16]. Similarly, Johansson et al. [36] systematically summarized evidence on the efficacy of various interventions. However, current studies evaluate the efficacy within homogeneous settings, often overlooking the variety of social media platform characteristics and their potential impact on interventions' efficacy and acceptance. Our study addresses this gap by comparing multiple widely used misinformation interventions across distinct content.

Importantly, interventions may have unintended consequences [31] like over-correction and other spill-over effects on accurate information [24, 29], over-reliance on interventions/automation bias [41, 65] (especially because they are technology-driven [30]), increased belief in misinformation when repeatedly exposed to the content [6, 55, 66], or priming of general mistrust in accurate content due to warnings [69]. Given the complexity of human decision-making, it is not surprising that there are many psychological effects and biases to be considered when designing misinformation interventions. Among these are backfire effects when users feel patronized [34] and the 'Implied Truth Effect' where warnings attached to misinformation increase the perceived accuracy of content without warnings [50] — even if it is misinformation.

Often, interventions depend on a prior misinformation detection — automated or manual — to trigger corrections or warnings. However, even with promising progress in (often AI-based) detection approaches [57, 77], it remains challenging and flawed (i.e., there are false positive and false negative outputs), and often lacks transparent and user-centered explanations. Related work on the 'automation bias' has investigated how people consider advice from AI algorithms in comparison to human advice, revealing a tendency for people to over-rely on automation. The bias has been explored for specific sensitive contexts such as medical decisions and diagnoses [26]. Considering the recent novelties in AI for misinformation detection, it is crucial to investigate that context as well. First glimpses have been proposed in qualitative research that explores how teenagers reflect on the credibility assessment of a simulated detection of misinformation on TikTok [30], requiring further investigations building on it. While the automation bias and biases in training data are relevant for misinformation interventions that include an automatic pre-filtering approach, interventions that rely on human credibility assessments entail different challenges. Recent controversial discussions have criticized social media platform operators for shifting from professional fact-checks to community-driven approaches[1]. For instance, X applies community notes that rely on other users' feedback and their helpfulness ratings instead of expert assessments[2]. Other interventions are designed to be displayed by default, independent of any previous detection. For example, this applies to many accuracy prompts that nudge users to critically reflect on the accuracy of content with targeted questions [54].

In our study, we include various types of interventions that entail different biases and errors, facilitating a more critical view of the beneficial and adverse effects of misinformation interventions.

[1]https://www.dw.com/en/fact-check-are-xs-community-notes-fixing-or-fueling-misinformation/a-73315972

[2]https://help.x.com/en/using-x/community-notes

2.4 Research Gap

This study advances HCI research on misinformation interventions with a critical perspective on the efficacy and perceptions of erroneous versus flawless interventions across modalities. Despite the growing body of research on misinformation interventions, several key gaps remain underexplored:

1st gap: Robustness across modalities and social media platforms. While numerous studies have explored the efficacy of interventions in platform-specific settings, these investigations often occur in isolation, mostly focus on text-based content, and lack cross-contextual comparisons. It remains unclear how interventions perform across different social media platform environments and if specific content modalities necessitate specific intervention types.

2nd gap: Erroneous interventions. Existing research often evaluate interventions under ideal deployment conditions, where detection systems (manual or automatic) are flawless and interventions are applied appropriately. However, real-world systems are fallible, and detection errors can undermine efficacy and trust. Few studies have accounted for such erroneous intervention behavior, despite its essential implications for users.

This study addresses these research gaps by conducting a comparative evaluation of misinformation interventions. Our overarching goal is to answer the following research questions:

RQ1: *How does the efficacy and user perception of state-of-the-art misinformation interventions vary across different content modalities?*

RQ2: *How does the presence of errors in misinformation interventions impact efficacy and user perception?*

3 Methodology

To answer our research questions, we conducted an online experiment with $N = 1,004$ English-speaking participants from the USA, simulating the environments of X, TikTok, and Telegram. The experiment implementation (i.e., stimuli, interventions, and questionnaires) has been optimized for both mobile and desktop participation. In the following, we provide details on study procedure, participants, stimuli selection, misinformation interventions, and statistical analysis.

3.1 Study Procedure and Ethics

The study procedure was inspired by the work of Pennycook et al. [51], who provide practical guidelines for misinformation research. Please see Figure 1 for a visualization of the study procedure. Our study design was approved in advance by the university's ethics committee (IRB Number EK 05/25). The data was secured and processed in accordance with the data protection provisions of the GDPR. We collected limited personal information (age, gender, education, political preferences) and did not gather additional sensitive data (e.g., ethnicity, religion, health). The experimental data was stored on servers compliant with GDPR regulations.

Participants were randomly assigned to one of 30 groups: either (a) to one of three *control groups* without misinformation intervention (one per social media platform), (b) to one of five interventions with a *flawless performance* (i.e., interventions always have a correct output), or (c) to one of four misinformation interventions with a

erroneous performance (i.e., interventions occasionally being erroneous; inoculations were only included in the flawless performance condition). Thus, each participant interacted with only one social media platform — X, TikTok, or Telegram — and, if assigned, was exposed to only one intervention type.

After providing informed consent and socio-demographic information, participants were shown a two-minute introductory video explaining the study procedure and social media platform simulations. The video demonstrated interactive elements of the respective social media platform simulation (i.e., how to open the comment section and profile description) and, if applicable, guided users to explore the intervention (e.g., by hovering over *see more* buttons).

Participants viewed 18 social media posts in randomized order (nine containing misinformation and nine containing accurate information). After each post, they responded to four questions on a 6-point Likert scale assessing (1) sharing intention, (2) content familiarity, (3) perceived importance, and (4) perceived accuracy (see Table 2, Appendix, for the full questionnaires). Following this task, participants completed a short survey on political preferences and their usage of X, TikTok, and Telegram. They also indicated whether they had responded randomly at any point during the study. In addition, we included two attention-check items within the experiment. Participants who failed were excluded from further participation.

Participants assigned to an intervention completed an additional questionnaire evaluating their experience. They were asked (1) whether they noticed any additional interface features or elements different from the typical social media platform, (2) whether they noticed the intervention (showing a screenshot), (3) how annoying or helpful they found the intervention if noticed, (4) whether they noticed the intervention being erroneous, and (5) what they liked or disliked about the intervention (in a open-text format).

For all participants, the experiment concluded with a debriefing outlining the study goals, emphasizing that the task involved both misinformation and accurate information, as well as (erroneous) interventions. We provided a downloadable PDF that explicitly corrected the misinformation posts and intervention errors, with links to professional fact-checks. Participants were thanked for their valuable contribution and redirected to the panel provider to receive their compensation.

3.2 Participants and Recruitment

Because standardized power analysis methods for cumulative link mixed models (CLMMs) are not available, we conducted a conventional a priori power analysis based on a one-way ANOVA with 30 groups ($\alpha = 0.05$, $1 - \beta = 0.80$). This analysis suggested that a sample of between $N = 410$ and $N = 2,444$ participants would be required to detect small to medium effects. To enable fine-grained comparisons across experimental conditions, and in light of the ordinal outcome structure and inclusion of random effects in the CLMM, we ultimately recruited $N = 1,004$ (approximately 33 participants per group, see Tab. 5, Appendix), which provided sufficient statistical power for the planned analyses. Participants were recruited via *Prolific*[3], a panel provider platform specifically designed to provide samples for scientific studies and whose reliability has been confirmed by several studies [2, 49]. Misinformation research is always connected to the contextual environment it is located in, for example, regarding the choice of stimuli that should be relevant to the audience. As for our statistical analysis a substantial sample size was essential, we focused on individuals with a residency in the U.S. as the largest represented sample on Prolific. We sought diverse perspectives by selecting a representative sample regarding sex, age, and political affiliation based on the USA census data, and included diverse educational backgrounds. The participants were 52% female, 47% male, and 1% diverse. Their educational levels ranged from less than a high school degree to a doctoral degree (see Tab. 4, Appendix). Participants received $4,50 as compensation for an average duration of 30 minutes.

3.3 Stimuli

Our goal was to use realistic, ecologically valid content based on real-world social media examples. To identify suitable misinformation posts, we (1) *consulted official fact-checking websites* (e.g., PolitiFact) to identify current and officially debunked topics that span a wide range of misinformation and political perspectives. These debunking articles often link directly to original social media posts, which was the case for all selected original X posts and TikTok videos. For Telegram, we manually searched public channels for voice messages related to the identified misinformation topics.

We also (2) included nine social media posts with *accurate information or neutral personal opinions*. Following guidelines on misinformation research [52], we refrained from using only verified-true content from fact-checking websites and instead included a variety of creators as sources. As with the misinformation content, we selected a mix of topics that were neither outdated nor likely to be overly familiar to average participants.

For both misinformation and accurate content, we (3) *selected six original posts* per social media platform (X, TikTok, Telegram). As we aimed to confront all participants with the same misinformation and accurate information, despite the different assigned social media platforms, each participant saw six posts original to the assigned social media platform, and the remaining posts were converted to the social media platform modality. Figure 11 (Appendix) demonstrates an exemplary conversion flow. Content was carefully reformatted to match the requirements and conventions of the target modality. Text and audio were relatively straightforward to convert from TikTok to X and Telegram, while converting X posts or Telegram voice messages to TikTok videos required greater effort. We used *CapCut*[4] to produce TikTok-style videos, using either existing footage from fact-check sources or suitable public TikTok content. Audio was created via a researcher-recorded voiceover reading the original post text, and videos included typical on-screen captions based on the transcribed audio. Message framing, length, and tone were preserved during reformatting. We argue that this kind of cross-modal content adaptation mirrors naturally occurring social media practices, as sharing and re-purposing content across social media platforms is common in everyday use.

We (4) enriched all posts with *contextual cues*, including the number of likes, shares (not for Telegram, as not applicable there),

[3] https://prolific.co/

[4] https://capcut.com

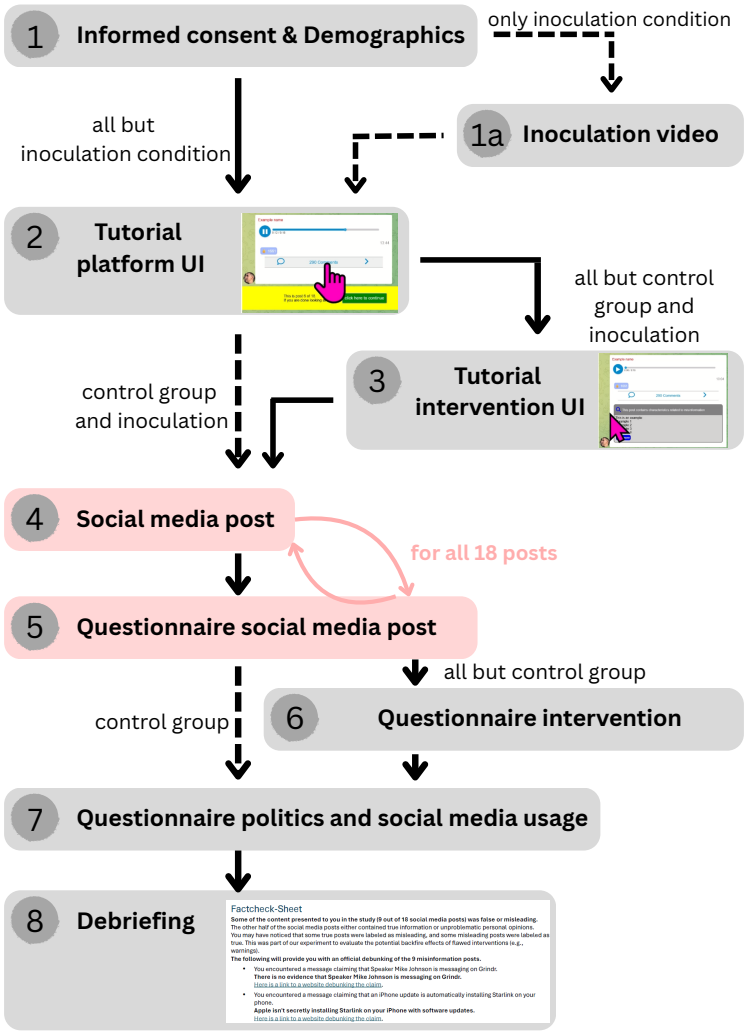


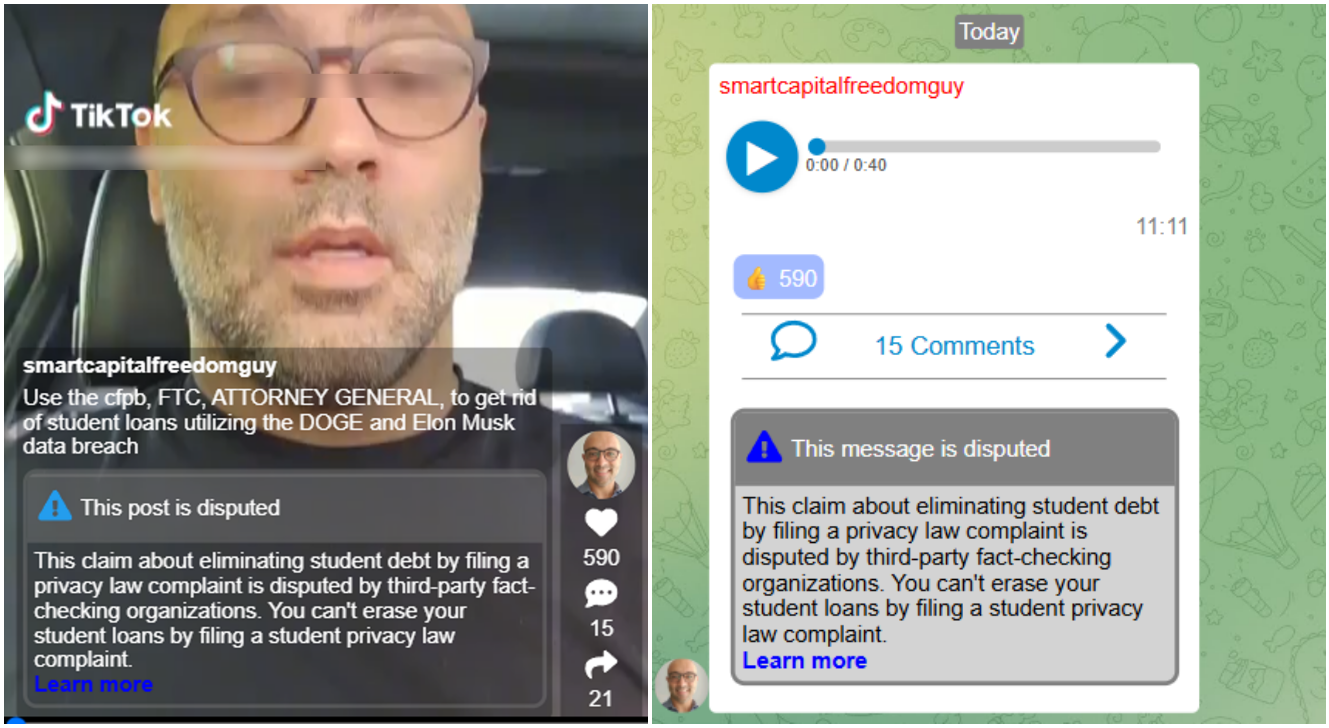
Figure 1: Flow of the study procedure.

and comments drawn from the original source. Participants could view the first five comments and access creator profiles, including bio texts. Usernames were pseudonymized, except in the case of verified official accounts (e.g., those with a blue check mark). See Table 3 in the Appendix for detailed information on the selected stimuli.

3.4 Erroneous and Flawless Misinformation Interventions

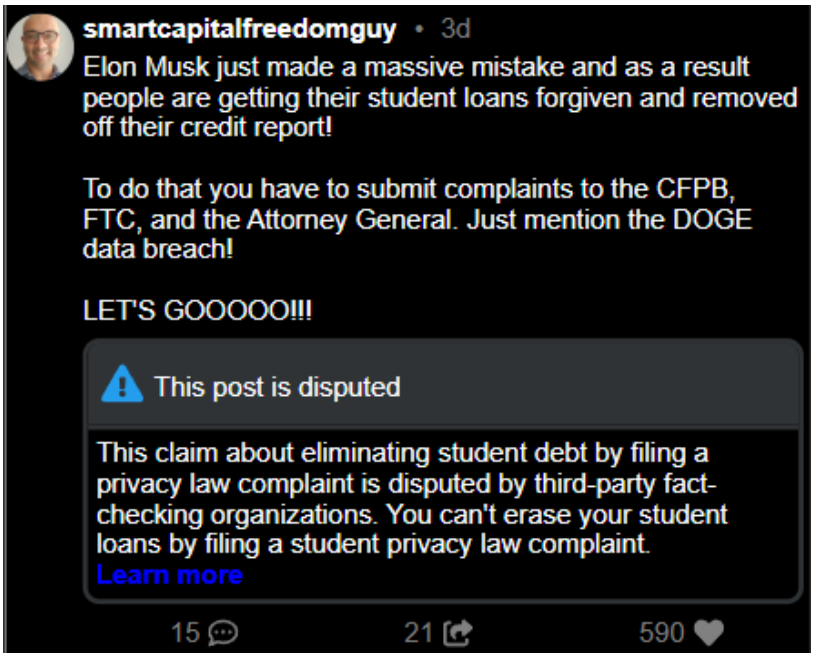
Given the numerous types of interventions explored in the academic research, the question arises as to which specific ones should be investigated. In selecting the interventions for this study, we aimed to cover a broad variety, guided by the umbrella categories (e.g., refutation strategies, nudges, boosts) proposed by Kozyreva et al. [38]. Within each category, we prioritized interventions either already used in practice or those that are often implemented in research. The selected interventions were: (1) professional fact-checks, (2) community notes, (3) indicators, (4) accuracy prompts, and (5) inoculation (see Figures 2, 3, 4, 5, 6, and 7). The interventions were styled consistently across the social media platforms. Consequently, all user interface elements described in Section 3.4.1 are applied to X, TikTok, and Telegram. For instance, a contextual note by other users is inserted below the post (i.e., below the X post, below the TikTok video, or below the voice message on Telegram) in the community notes. The key elements of all interventions can be found in Figures 2, 3, 4, 5, 6, and 7.

3.4.1 Intervention Descriptions. In light of the recent shifts from *professional fact-checks* to *community notes* on several social media



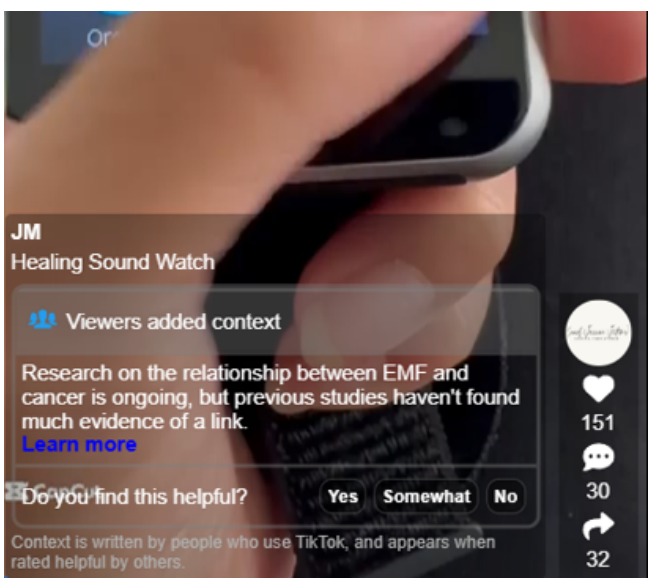
(a) Professional fact-check on TikTok

(b) Professional fact-check on Telegram

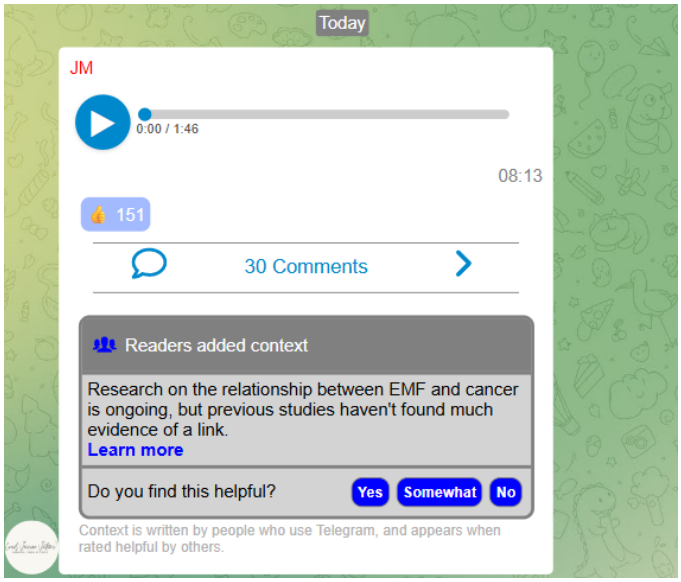


(c) Professional fact-check on X

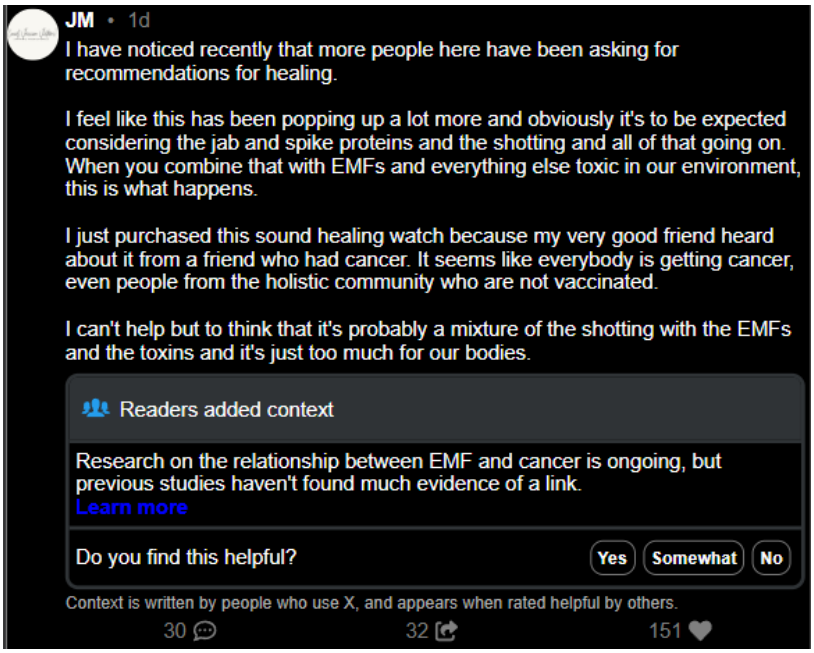
Figure 2: Overview of professional fact-checks across social media platforms.



(a) Community note on TikTok



(b) Community note on Telegram

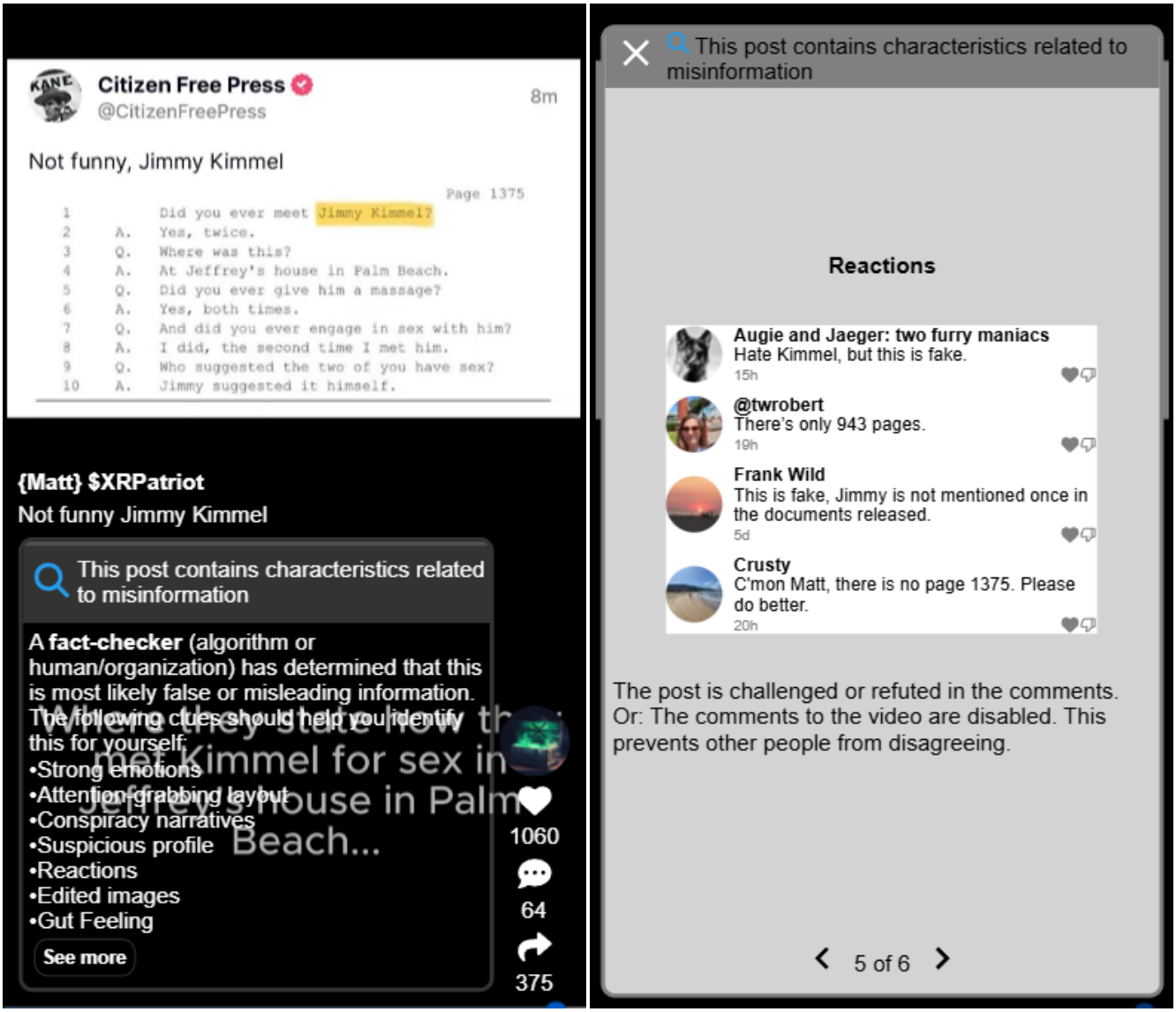


(c) Community note on X

Figure 3: Overview of community notes across social media platforms.

platforms, and their key role in practice, we included both interventions (see Figures 2 and 3). Their designs were based on historical and current interventions on X. In the community notes intervention, a contextual note contributed by other users is shown below the post, along with the option to rate the helpfulness of the information. In contrast, professional fact-checks use a similarly styled note, but written and verified by professional fact-checkers.

As a similarly transparent intervention, we included *indicators* for misinformation (see Figures 4, 5, 6), which have been qualitatively studied across modalities and are likely to address users' needs for comprehensibility and transparency [30, 37]. Unlike fact-checks or community notes, indicator-based interventions (though not always) rely on an automated pre-screening, followed by highlighting content features that signal unreliability (e.g., dramatic



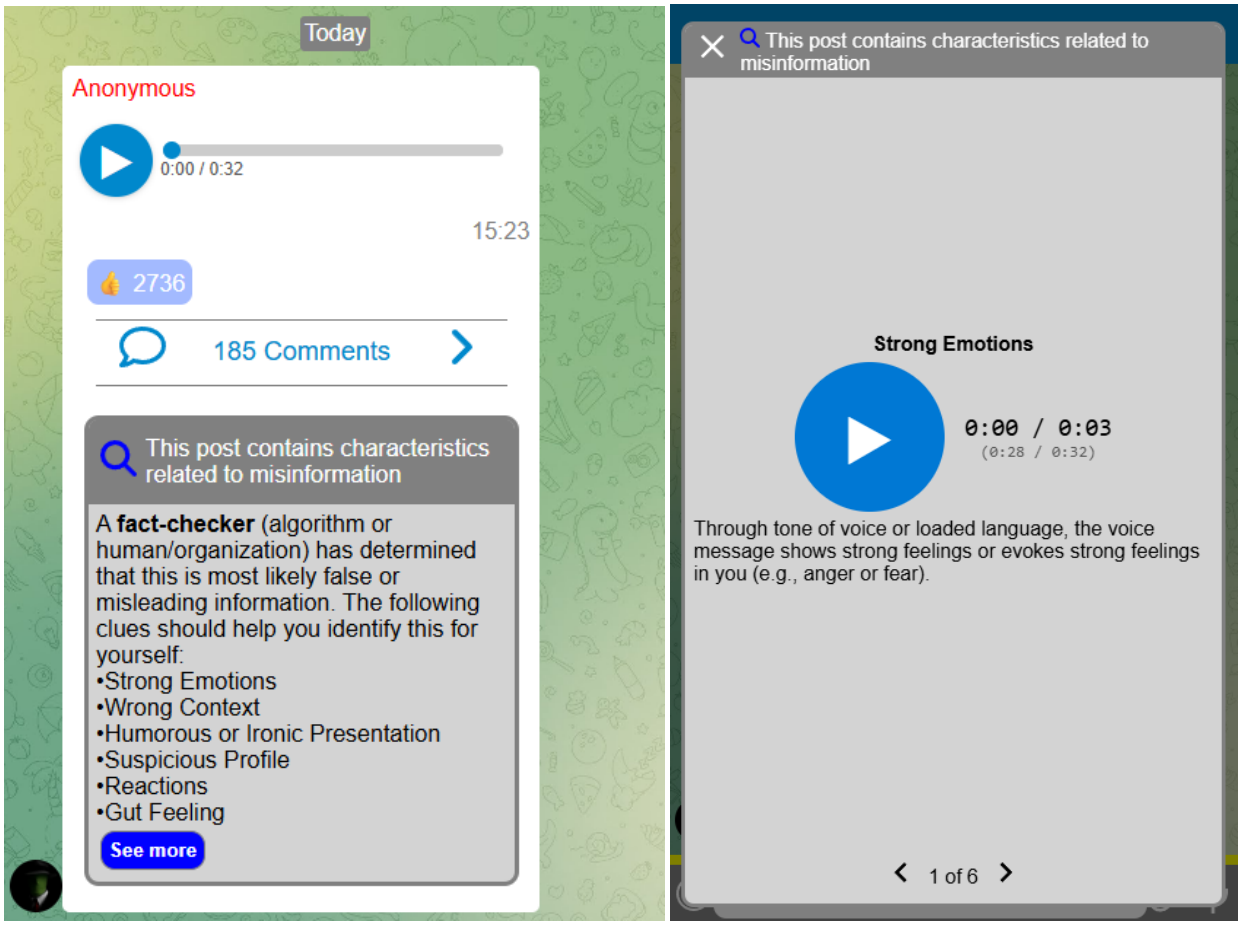
(a) Indicators on TikTok **(b) Indicators on TikTok (detailed view)**

Figure 4: Overview of indicators on TikTok with main list (left) and detailed view on one specific indicator (right).

music playing in the background, hashtags with conspiratorial keywords, attention-grabbing layout). In our version, a text indicates that an algorithm or a human evaluator has determined that this is most likely false and presents a list of clues to support this judgment. Participants can swipe through the intervention to explore each indicator and see how it maps to components of the post.

Representing the nudge category, we included *accuracy prompts* (see Figure 7). While not currently deployed by social media platforms in practice (to our knowledge), their efficacy is well supported by academic research [53, 54]. Unlike the other four interventions, accuracy prompts operate at a more subconscious level. The prompt appears after each misinformation post, once participants finish engaging with it and attempt to proceed. It asks participants to rate the accuracy of the post using a 4-point Likert scale, with the goal of nudging cognitive reflection. Responding is mandatory to ensure engagement, although responses are not saved or analyzed in this study.

To represent the boost category Kozyreva et al. [38], we included *inoculation* (see Figure 7d). Despite strong empirical support, inoculation techniques are not yet widely implemented on social media platforms. Unlike the other interventions, which appear during or after each misinformation post, inoculation was presented once before participants viewed any stimuli. We adopted the video-based approach developed by Roozenbeek et al. [59], originally consisting



(a) Indicators on Telegram **(b) Indicators on Telegram (detailed view)**

Figure 5: Overview of indicators on Telegram with main list (left) and detailed view on one specific indicator (right).

of five short videos[5], each targeting a manipulation technique. Due to time constraints, we included the three videos most relevant to our stimuli and edited them into a single compilation. These addressed emotionally manipulative language (e.g., words that evoke fear or outrage), scapegoating (i.e., singling out a person or group for a particular problem), and ad hominem attacks (i.e., attacking the person making an argument and not addressing the argument itself to redirect away from the subject and towards an individual) [59].

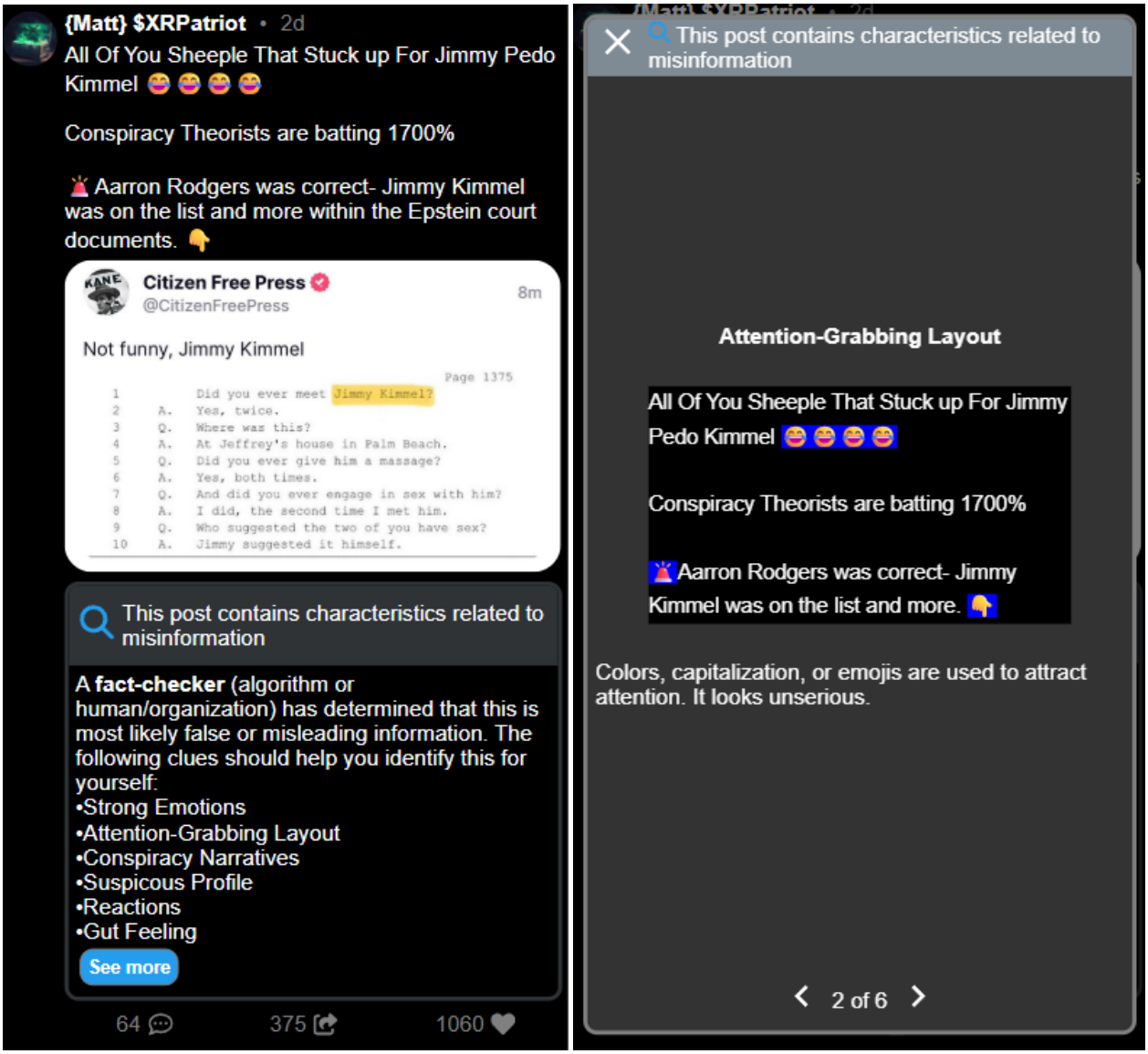
3.4.2 Erroneous Conditions. While some experimental groups were assigned to perfectly functioning interventions, others experienced an erroneous condition. We assume that all selected interventions — except inoculation — can naturally produce errors in real-world implementations.

These flaws manifest as *false positives*, where manual or automatic pre-screening incorrectly classifies accurate content as misinformation. As a result, users may encounter fact-checks, community notes, accuracy prompts, and indicators attached to posts that are, in fact, accurate.

Conversely, *false negatives* occur when misinformation is not detected and is thus presented without any corrective intervention. In this case, fact-checks, community notes, indicators, and accuracy prompts are absent, even though the content is misleading. Following prior work on indicator-based interventions, we included a generic message in such cases, suggesting that an algorithm (or human) determined the content is likely accurate, even though it is actually misinformation.

Based on the assumption that false negatives are slightly more probable than false positives, the erroneous condition included three false negatives and two false positives, in addition to thirteen correctly applied interventions. This distribution was designed to simulate the imperfect performance of (automated and human) misinformation detection systems. The error rates, however, are higher than in recent promising detection efforts [1] to allow for a noticeable amount within the restricted sample of 18 stimuli. Thus,

[5]https://inoculation.science/inoculation-videos/



(a) Indicators on X **(b) Indicators on X (detailed view)**

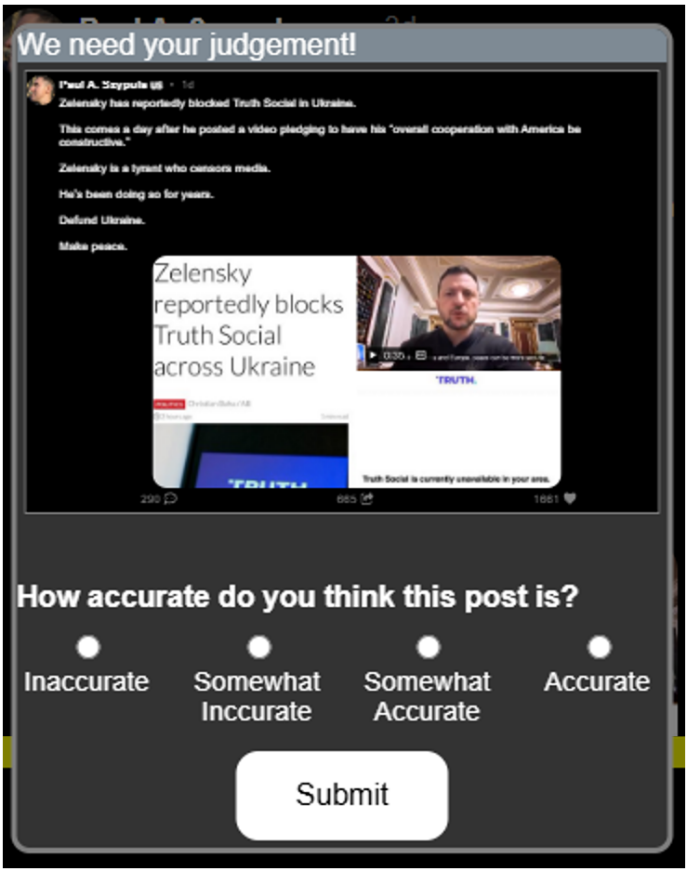
Figure 6: Overview of indicators on X with main list (left) and detailed view on one specific indicator (right).

the distribution in our experiment is a simplification. It is noteworthy that particularly professional fact-checking interventions are likely to produce significantly fewer errors in real-world scenarios and may actively seek to correct them once identified. Future work may complement our simplified assumption with a more realistic, nuanced perspective on error rates.

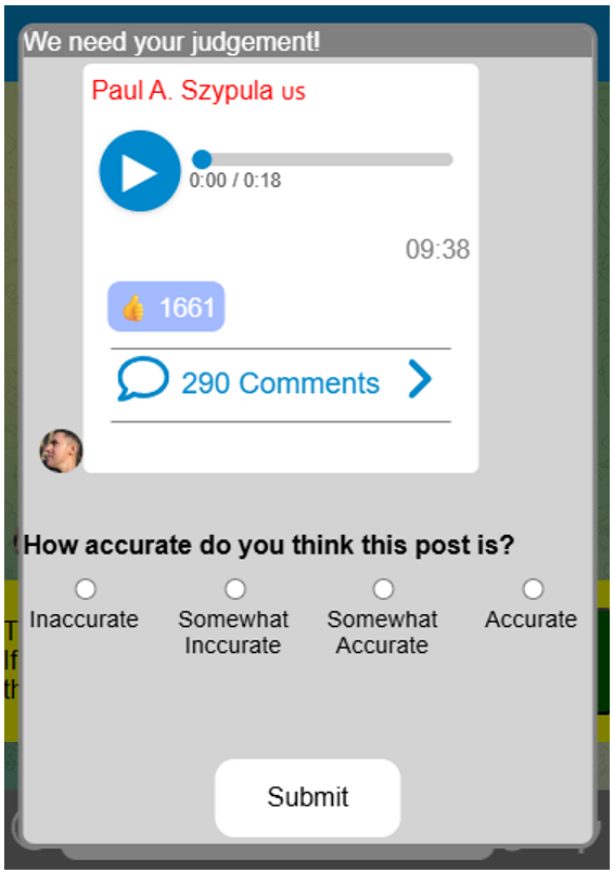
3.5 Analysis

The main objective in the analysis was to compare our 30 groups (control groups, interventions with flawless performance, intervention types with erroneous performance, three social media platforms) regarding accuracy ratings, sharing intentions, and user acceptance of interventions. To detect group-level differences, we used CLMMs as an extension of the cumulative link models (CLM) [67] using the R 'ordinal' package[6], allowing for a precise analysis that incorporates each individual response separately. The approach facilitates accounting for the non-independence of observations, as

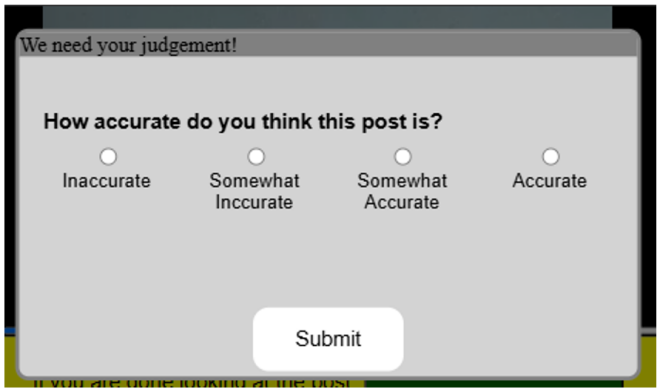
[6]https://cran.r-project.org/web/packages/ordinal/index.html



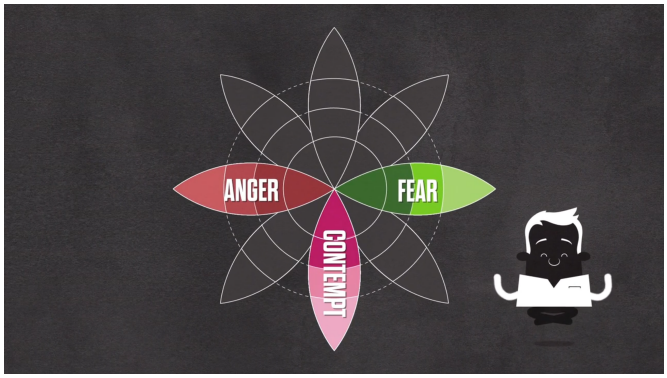
(a) Accuracy prompt on X



(b) Accuracy prompt on Telegram



(c) Accuracy prompt on TikTok



(d) Screenshot from inoculation video

Figure 7: Overview of accuracy nudges on all three social media platforms and screenshot of the inoculation video.

multiple participants responded to the same posts and individual participants responded to multiple posts. We modeled the posts and participants as random effects and the differences between the interventions and experimental groups as fixed effects using a logic link function. Subsequently, we conducted pairwise comparisons using the 'emmeans' package[7]. This resulted in estimated marginal means (EMMs), commonly known as least-squares means, enabling us to adjust for multiple comparisons and focus on the direct comparisons of interest within our model's framework. Additionally, we fitted ordinal logistic regression models to examine differences in

[7]https://cran.r-project.org/web/packages/emmeans/index.html

perceived helpfulness and annoyance of interventions as measures of user acceptance. To account for multiple comparisons during the statistical analysis, the Benjamini-Hochberg method was applied to correct alpha errors [11]. In addition, we separately analyzed the subset of participants who were assigned to erroneous interventions to examine effects on user acceptance and whether they noticed the errors. Therefore, we conducted a χ^2 analysis looking for group differences and enriched them with anecdotal qualitative insights from a free-text question. All analyses were conducted in RStudio version 4.2.2[8].

4 Findings

The main goal of our analysis was to understand differences in accuracy ratings and sharing tendencies for misinformation based on intervention type (*inoculation*, *accuracy prompt*, *community note*, *fact-check*, *indicators*), whether the intervention was erroneous or not (intervention quality *high* and *low*), and the assigned social media platform/modality (TikTok, X, Telegram). For misinformation posts, a low accuracy rating is desirable, while for accurate posts, a high accuracy rating is desirable. Figure 8 shows the mean accuracy ratings of misinformation (6-point Likert scale from 'extremely inaccurate' to 'extremely accurate') by intervention type, intervention quality, and social media platform. Across all social media platforms, the control group without an intervention consistently produced the highest mean accuracy ratings of misinformation (i.e., rating misinformation as more accurate). With flawless interventions, misinformation was generally rated less accurate than in the control group or erroneous interventions. With erroneous interventions, misinformation was often rated less accurate than in the control group but more accurate than with flawless interventions – particularly regarding *fact-check* and *indicators*. Interestingly, in some instances, like *community note*, the flawless implementation did not consistently lead to misinformation being rated less accurate compared to erroneous implementation. The error bars, however, indicate some overlap between conditions within social media platforms, suggesting that while trends are visible, differences in means should be interpreted alongside the inferential results in Section 4.1. In the following, we report on statistically significant effects on accuracy ratings and user acceptance, indicating differences between the various conditions.

4.1 Efficacy of Interventions

To assess statistical differences between experimental conditions and social media platforms, we fitted a CLMM with accuracy ratings for misinformation as the outcome, and study group and social media platform (and their interaction) as fixed effects. To control for systematic effects of the stimuli (which were the same across participants) and participants (who viewed multiple stimuli), we incorporated these as random effects. The results in this regard indicated substantial between-user variance and smaller between-stimulus variance.

Overall, the results (see Table 1) showed how interventions varied in their effect on accuracy ratings of misinformation compared to the control group without intervention. Specifically, *fact-check* ($estimate = -1.16, p = .009$) and *indicators* ($estimate = -1.13, p = .01$) significantly improved (i.e., reduced) accuracy ratings of misinformation. In contrast, *inoculation*, *accuracy prompt*, and *community note* did not significantly affect accuracy ratings according to this complex model. Interestingly, we found no significant main effects for social media platform differences. Furthermore, no interactions between intervention and social media platform were observed, indicating consistent effects of interventions across modalities and social media platforms. We then conducted post-hoc pairwise comparisons between all study groups using estimated marginal means (EMMs) to gain a more detailed understanding of the results (see Table 6). Specifically, *fact-check* and *indicator* interventions significantly reduced the perceived accuracy of misinformation, compared to both the *control group* and the *inoculation* group (for both *fact-check* and *indicators*), as well as compared to the accuracy prompt (*fact-check*). Interestingly, when considering EMMs averaged across social media platforms, *community note* also significantly reduced accuracy ratings of misinformation compared to the control group. Thus, the results confirm the potential of *fact-check*, *indicators*, and *community notes* for more accurately discerning misinformation, with stronger evidence for the former two and weaker evidence for the latter.

Importantly, these effects were only observed when the analysis was restricted to interventions that were implemented flawlessly. When erroneous variants were included in the CLMM, none of the interventions significantly altered the perceived accuracy of misinformation ratings (all $p > .05$), although the direction of the effects remained consistent with the analysis of flawless interventions only (see Table 7, Appendix). Similarly, when modeling sharing intentions as the dependent variable, no significant effects were found, even for flawless interventions. This indicates that improved accuracy assessments did not directly result in changes to sharing behavior (see Table 8, Appendix).

4.2 User Acceptance of Interventions

In addition to effects of interventions on sharing intentions and accuracy ratings, two measures of user acceptance were collected: perceived helpfulness and perceived annoyance of the interventions. Spearman's rank correlation and Cronbach's alpha as measures of association between the two suggest that they do not form a coherent scale (both $< .50$), which is why we analyzed the two separately. We first evaluated user acceptance of the interventions by combining the erroneous and flawless variants to obtain more robust insights into how the interventions would be perceived in a realistic scenario.

An ordinal logistic regression was conducted to evaluate potential interactions between study groups and social media platforms, examining the effects of intervention type and social media platform on participants' ratings of helpfulness. Compared to the *accuracy prompt* with the lowest helpfulness ratings, the model revealed a significant difference for *community note* and *fact-check* (see Table 9, Appendix). Interaction terms suggested that on X, the relative advantage of *community notes* and *indicators* over *accuracy prompts* regarding perceived helpfulness is significantly larger than on TikTok, where additional information partly overlapped the video content. Direct comparisons between interventions using estimated marginal means (EMMs) showed significantly higher

[8]http://www.rstudio.com/

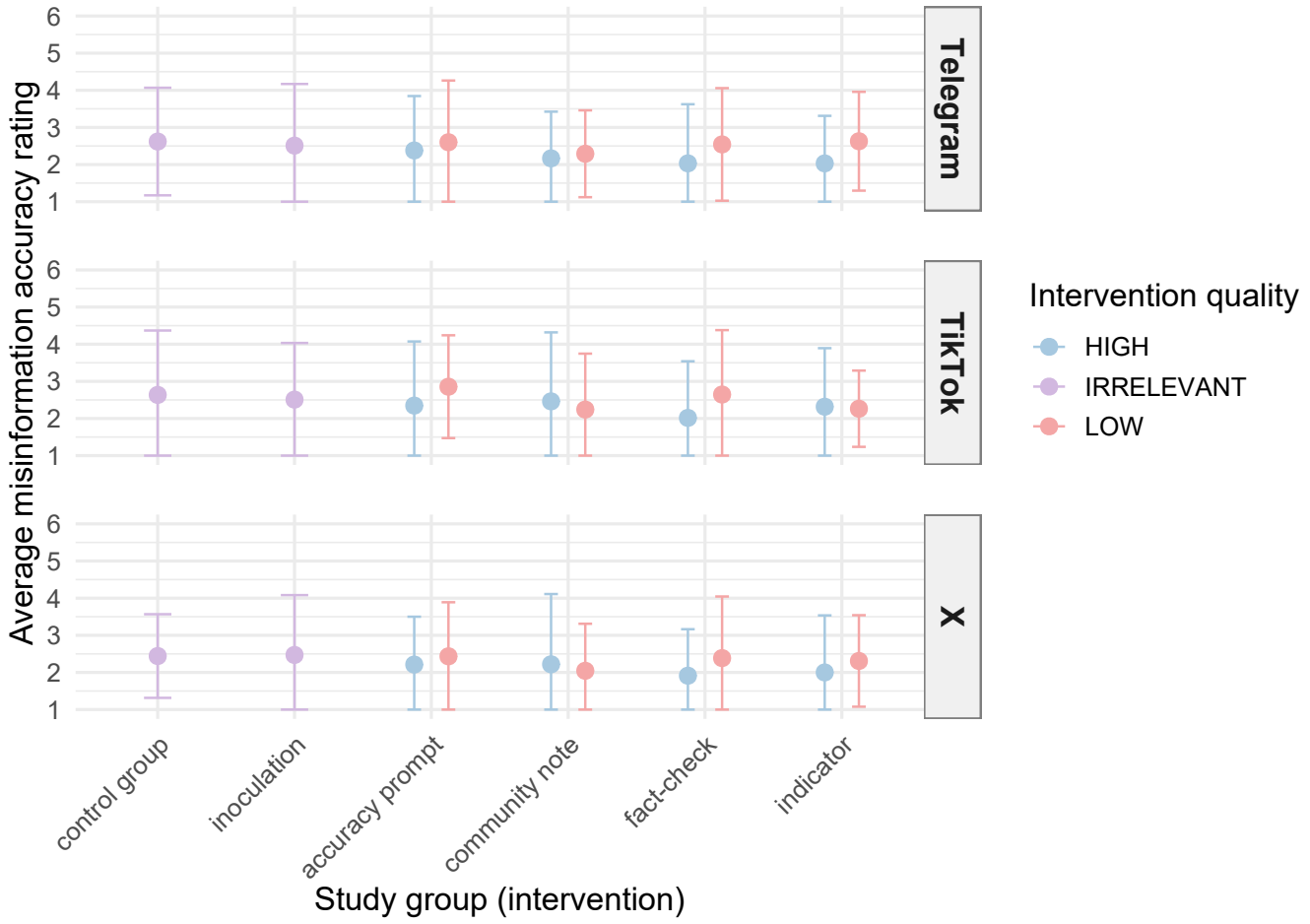


Figure 8: Mean accuracy ratings of misinformation across social media platforms and intervention types. For these corresponding misinformation posts, a low accuracy rating is desirable in contrast to a high accuracy rating.

helpfulness ratings for *fact-check* and *community note* compared to the other interventions, but also for *indicators* to a smaller extent (see Table 11 (Appendix) and Figure 9a). Importantly, estimates here are based on the latent scale of the ordinal logistic regression and do not correspond to raw means. Overall, these results suggest that certain interventions, particularly *fact-check* and *community notes*, were perceived as more helpful by participants across social media platforms. This indicates a preference for interventions that do not disrupt the natural flow of social media use by being integrated into the interface instead of appearing, for instance, as additional pop-ups.

To evaluate effects of intervention type and social media platform on participants' ratings of annoyance, another ordinal logistic regression was conducted. Compared to the *accuracy prompt* with the lowest annoyance ratings, the model revealed significant differences for *fact-check* and *indicators*, among others (see Table 10, Appendix). Interaction terms suggested that on X, the perceived annoyance of *community notes* and *indicators* relative to *accuracy prompts* is significantly lower than on TikTok. More detailed direct comparisons between interventions using EMMs (see Table 12 (Appendix) and Figure 9b) showed the lowest annoyance rating for *accuracy prompt* and the highest annoyance rating for *inoculation*. *Fact-check* and *indicators* rank in the middle, and are still perceived as significantly more annoying than *accuracy prompt*.

Separately analyzing the subset of participants that were assigned to erroneous interventions (i.e., all but inoculation), we found that 35% stated not to have noticed the interventions' errors, 45% were not sure, and 8% did not notice the intervention overall. In contrast, 11% stated that they had noticed the assigned intervention's errors. This largely differs among intervention types. In particular, only two participants noticed the accuracy prompt being erroneous (i.e., it was sometimes missing when the content was misinformation and was sometimes displayed, although the content was correct). As the accuracy prompt does not include a statement about an external accuracy assessment of other parties and only aims to nudge a critical reflection of the individually targeted user,

Table 1: Cumulative Link Mixed Model predicting perceived accuracy ratings (flawless interventions only).

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Fixed Effects:				
inoculation	-0.159	0.424	-0.375	0.708
accuracy prompt	-0.445	0.419	-1.060	0.289
community note	-0.629	0.438	-1.438	0.150
fact-check	-1.155	0.440	-2.628	0.009
indicator	-1.128	0.436	-2.586	0.010
TikTok	-0.007	0.422	-0.017	0.986
X	-0.239	0.415	-0.576	0.564
inoculation × TikTok	-0.149	0.599	-0.248	0.804
accuracy prompt × TikTok	-0.017	0.600	-0.029	0.977
community note × TikTok	0.187	0.609	0.307	0.759
fact-check × TikTok	-0.146	0.610	-0.239	0.811
indicator × TikTok	0.618	0.603	1.025	0.305
inoculation × X	0.059	0.591	0.099	0.921
accuracy prompt × X	-0.044	0.591	-0.074	0.941
community note × X	-0.082	0.607	-0.135	0.892
fact-check × X	0.124	0.597	0.208	0.835
indicator × X	0.252	0.609	0.414	0.679
Random Effects:	**Variance**	**Std. Dev.**		
Participant (Intercept)	5.469	2.339		
Post (Intercept)	0.052	0.229		
Thresholds:	**Estimate**	**Std. Error**	**z value**	
1|2	-0.942	0.337	-2.792	
2|3	0.266	0.337	0.790	
3|4	1.041	0.338	3.083	
4|5	2.466	0.340	7.262	
5|6	4.371	0.350	12.481	

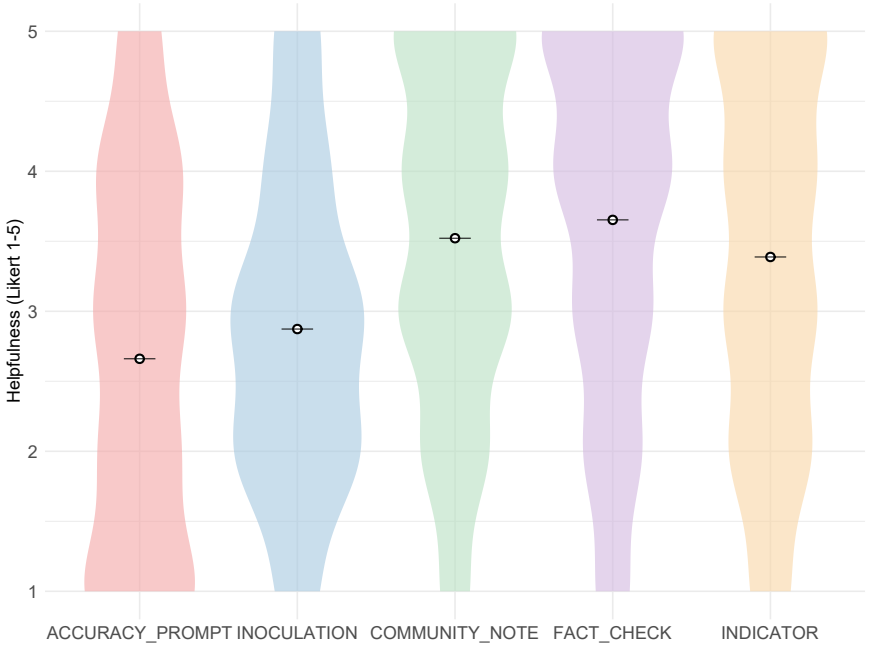
Note: Reference categories used in the model are 'No intervention' and 'Social Media Platform: Telegram'. A Likelihood Ratio Test indicates a significant improvement over a null model ($\chi^2 = 34.2(17), p = 0.007$).

patterns of 'flaws' are hard to detect, and it is even arguable if the intervention can actually be noticeably erroneous. Therefore, we excluded the erroneous accuracy prompt from the following statistical analyses. This allowed us to conduct a χ^2 analysis, as all remaining groups are large enough for statistical comparison.

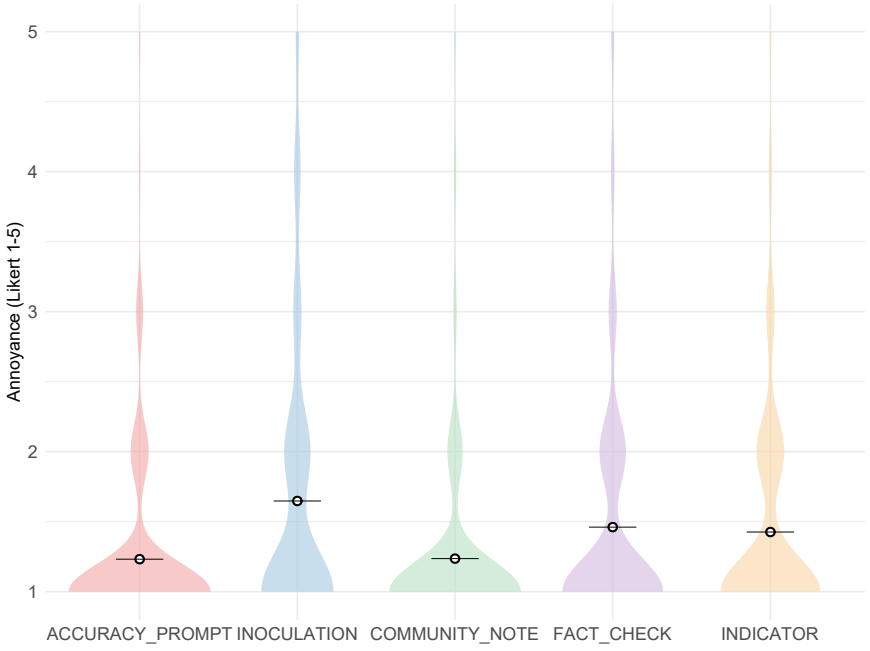
Confronted with erroneous community notes, $N = 41$ participants did not notice any errors, and $N = 6$ stated that they had noticed the errors. This is similar for professional fact-checks ($N = 42$ and $N = 9$). In contrast, errors of indicators were not noticed by $N = 24$ participants, while $N = 18$ stated that they had noticed the indicator intervention's errors. Conducting a χ^2 test, we found significant statistical differences with a small effect size between the different erroneous interventions regarding whether the errors were noticed or not ($\chi^2(2, N = 302) = 9.22$, $p = .010$, Cramer's $V = 0.17$).

We expected that noticing the errors of interventions impacts the participants' perceived helpfulness and annoyance. A Wilcoxon rank-sum test revealed statistically significant differences (see Figure 10): Interventions' helpfulness was rated lower when participants noticed errors ($W = 2262.5, p = .004$), and the annoyance was rated significantly higher ($W = 1337, p = .013$). Both observations confirmed our hypothesis.

We aimed to delve deeper into participants' reflections on the erroneous interventions by considering the free-text question that asked what they liked or disliked about the interventions. We did not systematically analyze these textual answers, but provided some anecdotal insights to enrich our quantitative analysis. We were particularly interested in the subgroups that stated to have noticed the errors (community notes: $N = 6$ out of 101, professional fact-checks: $N = 9$ out of 105, indicators: $N = 18$ out of 96). For the *community notes*, some participants explicitly stated that they did not like the intervention because of its errors (i.e., claiming the content is accurate although it is false or the other way around). It was further criticized that community notes are solely based on other users' assessments and sometimes *"lack clear sources, which may reduce trust in the intervention"* (#P155). Among those six who noticed the errors, only two participants positively commented on the intervention. For the *professional fact-checks*, some participants showed general skepticism (*"I don't believe fact checkers"* (#P324)) or raised concerns of oversimplification and censorship (*"[...] I'm concerned it may oversimplify complex issues or limit open discourse if not carefully implemented"* (#P641)). For the *indicators*, out of the 18 participants who noticed the intervention being erroneous, twelve still highlighted positive aspects of the intervention. They

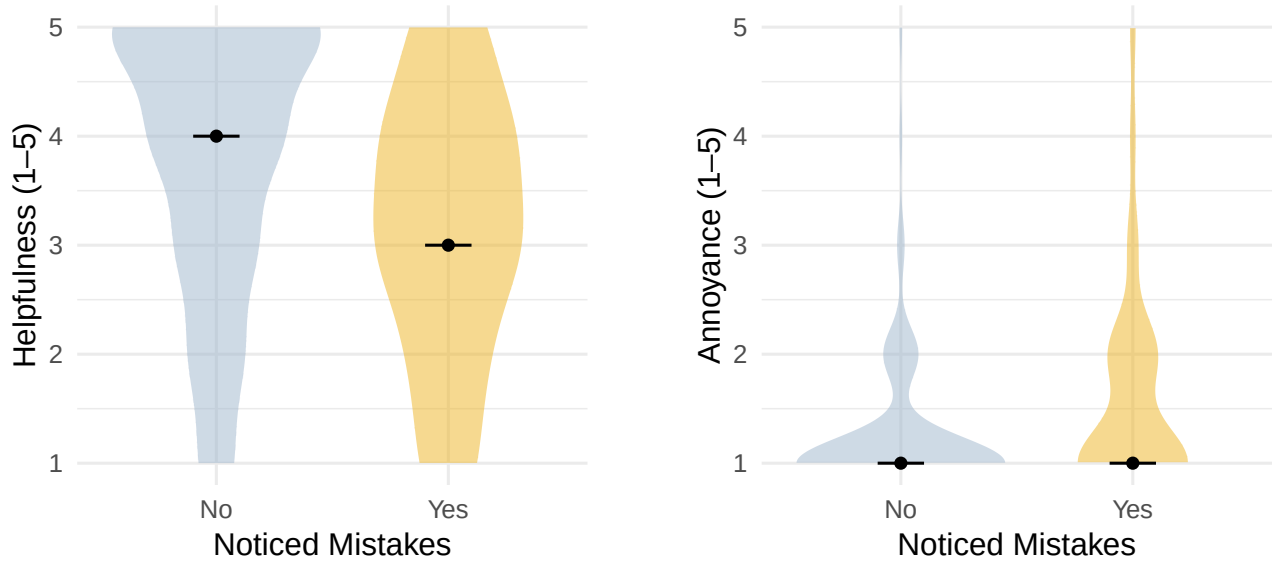


(a) Helpfulness ratings



(b) Annoyance ratings

Figure 9: Distributions of participants' ratings on helpfulness and annoyance of the interventions across social media platforms.



(a) Helpfulness ratings **(b) Annoyance ratings**

Figure 10: Distributions of participants' ratings on helpfulness and annoyance of the erroneous interventions, separated by whether they noticed errors.

emphasized its overall goal to *"add a bit more information to help understand what the message is saying"* (#P806), or *"help [...] to better identify ways in which emotion is evoked"* (#P813). However, it was also criticized for its inaccuracy: *"It could help people distinguish between true and not true, but it is wrong and untrustworthy"* (#P360). Indeed, some participants underscored the importance of the intervention being always correct: *"I think it's good to have the intervention in there, but the information must be correct. If it's not, it doesn't help at all"* (#P348).

From those who did not notice the errors, there were mostly positive comments on all three interventions. Twenty-nine participants confronted with erroneous community notes gave exclusively positive feedback like: *"I like it a lot. It combats misinformation by communicating the truth"* (#P45). Some were critical about the community being the source of the correction: *"I like the concept, but I think it could further misinformation because its just from a random source"* (#P116), while others explicitly liked that: *"It was helpful to see what other people think about the post"* (#P1210). Interestingly, one participant explicitly emphasized they did not expect an intervention to always be accurate and still liked it: *"[...] even if the intervention isn't necessarily 100% correct, it helps you to think a little and think twice about what you're reading"* (#P1242). Twenty-seven participants assigned to the erroneous professional fact-checks still only made positive comments on the intervention (e.g., *"What I like about the intervention is that it encourages people to pause and think before sharing something. It's a simple way to remind us to check if the information is actually true. In a world where things spread so fast online, that little moment of reflection can make a big difference in stopping misinformation."* (#P616)), while five exclusively gave negative feedback or raised concerns (e.g., *"Fact checking isn't always accurate either. So, relying on the intervention isn't a good idea. I don't like fact checkers. I'll do my own fact-checking, thank you very much. The internet is about 80 percent lies anyway."* (#P1264) or: *"Who's checking the fact-checkers"* (#P430)). Regarding indicators, the erroneous intervention received exclusively positive feedback from twenty participants, for instance: *"The intervention provided some important information needed when it comes to certain content that is on Telegram. Big help shuffling, which is factual from what is a lie"* (#P267), or *"I like how it combats misinformation and how deep and detailed the information it goes through to show you. I thought it was interesting that it also looks at the comments/replies to the post in its fight against misinformation"* (#P425). Others highlighted that the indicators facilitate a shift from gut feelings to rationality.

5 Discussion

This study evaluates the efficacy and user acceptance of five types of misinformation intervention (inoculation, accuracy prompt, community note, fact-check and indicators) across three social media platforms (TikTok, X and Telegram) in a large online experiment. The results show that the fact-check and indicators interventions, and to a lesser extent community notes, significantly reduced the perceived accuracy of misinformation. In contrast, inoculation and accuracy prompts did not influence accuracy ratings. These results were robust across social media platforms and modalities. The study also compared the effects of interventions that worked perfectly with those that were erroneous, resulting in (a) corrections or warnings being displayed on accurate information and (b) corrections or warnings missing on misinformation. The results show that significant positive effects can only be observed in flawless conditions, thus strengthening the need for accurate (manual or automatic) misinformation detection. The interventions were generally well

received in terms of user acceptance, particularly the community note, fact-check and indicators that provide additional information, whereas the accuracy prompt as less transparent approach was not considered very helpful. Conversely, accuracy prompts were rated as particularly less annoying, whereas the fact-check and indicators were rated as significantly more annoying. Some nuanced differences regarding social media platform and user perception also emerged, with community notes and indicators rated as relatively more helpful and less annoying on X compared to TikTok, highlighting that social media platform context can shape how interventions are received. This particularly reflects the limits of intervention integration within existing social media platform user interfaces that result in annoying content overlaps on TikTok, which might be solved with improved collapsible elements.

5.1 RQ1: How does the efficacy and user perception of state-of-the-art misinformation interventions vary across different content modalities?

5.1.1 Effective interventions. Fact-checking and indicators were consistently effective in reducing the perceived accuracy of misinformation on all social media platforms. The results regarding fact-checks are consistent with previous studies that have suggested their effectiveness for general claims and headlines independent from a specific social media platform or modality [44, 56, 62]. Our study further expands our understanding of indicator-based approaches to help users navigate misinformation. While previous studies have mainly provided qualitative insights into how such indicators are perceived by users [30, 33], the results of this study demonstrate their practical quantitative potential and suggest that they can effectively assist users in informed credibility assessments.

Community notes have previously produced mixed results: for example, one study suggests that they do not reduce engagement with misinformation [18], while another indicates that users perceive them as trustworthy [22]. In our experiment, community notes were found to have a significant, albeit less pronounced effects than fact-checks and indicators. As fact-checks are increasingly being replaced with community notes on Instagram, Facebook, and X, our results suggest that while community notes have potential, the value of fact-checks should not be overlooked. This is particularly relevant given that fact-checks are often cited within community notes, highlighting the close connection between these types of interventions.

5.1.2 Non-effective interventions. Our findings raise the question of why inoculation and accuracy prompts did not significantly affect the accuracy perception and sharing intention of misinformation. Previous studies have shown that accuracy prompts can generally significantly reduce the sharing of misinformation on social media [53, 54]. They are often shown after users indicate an intention to share a post containing misinformation. In our study, however, the intervention was presented regularly when the content contained misinformation, regardless of users' intent to share, in order to evaluate the effect of nudging people to reflect on accuracy independent from sharing intentions [53]. However, this frequent exposure to the intervention may have felt overwhelming or redundant, preventing it from producing a focused effect. Furthermore, our qualitative insights confirm a user preference for interventions providing comprehensible and transparent additional information, as previously shown in related work [27, 37], which was provided in all interventions but the accuracy prompt.

Similarly, inoculation has been shown to be effective in helping people to distinguish true information from misleading information, and in reducing their intention to share misinformation in related studies [10, 42, 59]. The design of the specific inoculation can be considered decisive, as it varies widely in terms of its modality, duration, and content. The inoculation used in this study was based on previous research [59] and adapted and shortened due to time constraints. However, the design may not have resonated with everyone: some participants described the inoculation videos in open comments as "childish", which undermined their credibility as a serious source of information. Others explicitly emphasized the entertainment value of the inoculation video. Lower levels of user acceptance, compared to other interventions, may have been due to the additional effort required to process the information presented in the inoculation intervention, as well as increased annoyance caused by it. It is therefore important to strike a balance between the lightness and the utility of an intervention. Indeed, those interventions embedded in the natural social media platform interface while providing additional comprehensible information (community notes, fact-checks, and indicators) were perceived as particularly helpful [22, 30]. This underscores how crucial the specific design of an intervention is, and how contradictory study results can often be explained by inconsistencies in its concrete implementation.

5.1.3 Sharing intentions. Although the interventions improved accuracy ratings for misinformation, it is important to note that they did not significantly reduce sharing intentions. Therefore, an improved ability to discern accuracy does not necessarily lead to behavioral change. This finding is consistent with previous research indicating that influencing actual behavior is a complex endeavor that encompasses more than just attitudinal and accuracy domains.

Overall, these findings suggest that efficacy and acceptance are relatively robust across content modalities, but the type of intervention still matters. These are promising findings, as most prior research on user-centered countermeasures against misinformation has focused on text-based misinformation [31]. One central contribution of this work, therefore, lies in showing that interventions such as fact-checks, indicators, and community notes are robust across modalities. While interventions are usually evaluated in only one modality, our findings provide crucial insights and give hope that lessons learned in certain domains can, to some degree, be transferred to others.

5.2 RQ2: How does the presence of errors in misinformation interventions impact efficacy and user perception?

5.2.1 Efficacy of erroneous interventions. The results of our study highlight that the presence of errors substantially undermines the efficacy of misinformation interventions. Interventions that operated

without any errors (i.e., only targeting misinformation correctly) significantly improved accuracy ratings, while erroneous versions of the same interventions failed to produce similar significant beneficial effects. It is interesting to observe that, on the one hand, users often tend to over-rely on AI-based tools – commonly referred to as automation bias [41, 65] – particularly because such tools are perceived as technologically advanced [30]. On the other hand, our study demonstrates that users are highly sensitive to errors in the system outputs of our interventions that partly encompass an underlying automated detection mechanism.

5.2.2 User perception of erroneous interventions. From a user perspective, interventions were consistently rated as more helpful and less annoying when simulated under perfect conditions than when errors occurred. This suggests that perceived reliability is a key driver of acceptance: if users cannot trust the accuracy of an intervention, they are less likely to value it or consider it when evaluating potential misinformation, regardless of its format. In addition to our quantitative results, this was also supported by the anecdotal qualitative insights reported as quotes of an open-text item. Participants who did not notice errors of the interventions often gave exclusively positive feedback in contrast to participants that noticed the errors. These findings emphasize the importance of designing interventions that minimize errors and clearly communicate the foundation of their judgments.

At the same time, our results highlight the fundamental dilemma of social media environments and technical support mechanisms being inherently imperfect. Therefore, future designs should not only aim to improve accuracy but also consider how to mitigate the consequences of inevitable errors. For example, this might be achieved through transparent communication of uncertainty or mechanisms for correcting errors over time.

5.3 Limitations and Future Work

While our study advances the understanding of misinformation interventions under flawless and erroneous conditions, several limitations need to be considered. *First*, our operationalization of erroneous interventions combined both types of errors – false positives and false negatives – into a single category. Future work should further disentangle the consequences of each variant of erroneous interventions for a more nuanced picture.

Second, our simulation encompassed error rates that are higher than expected from state-of-the-art detection approaches [1], as we aimed to include a noticeable amount of errors within our sample of 18 stimuli. To address the higher probability of false negatives, these occurred slightly more often than false positives in our simulation. It is a limitation of our study that the error rate was simulated as consistent across all erroneous intervention types, not considering expected nuances. For instance, manual professional fact-checks may make fewer mistakes than automated or crowd-based interventions and may encompass different error correction mechanisms. Future work should model these differences more realistically, while also considering correction mechanisms that professional fact-checkers and crowd approaches may employ.

Third, while our experiment simulated erroneous interventions to get a more realistic understanding of the potential of interventions, actual real-world environments are more complex. Evaluating erroneous interventions in field studies or in collaboration with social media platforms, therefore, would provide an even stronger 'reality check' of how errors manifest in practice and how users might adapt over time, or how many false positives and false negatives would be accepted by users within both manual (crowd-based or professional) and automated detection approaches.

Fourth, we aimed to compare modalities and social media platforms consistently for the same misinformation content, which led to content conversion from one social media platform to another. While we conducted the conversion process thoroughly and systematically, there are still some simplifications involved that do not represent perfect real-world conditions. For example, we included the same quantitative interaction characteristics (i.e., number of likes, shares, and comments) for all social media platforms, while these would normally differ depending on the corresponding social media platform. Future research might aim to find more realistic stimuli selection processes for cross-platform comparisons.

Fifth, participants' answers to the intervention 'accuracy prompt' were not saved or analyzed. In future studies, collecting these responses could provide additional valuable insights into users' thought processes.

Sixth, while we gained qualitative insights into the overall perception of the intervention types (e.g., their perceived annoyance and helpfulness), adding the dimension of trust would be an interesting additional consideration for future work. Particularly in lights of current discussions on community notes and professional fact-checks, these insights could provide valuable implications for intervention design.

6 Conclusion

Through a large-scale online experiment across three social media platforms (TikTok, X, and Telegram), we evaluated the efficacy and user acceptance of five prominent misinformation interventions: inoculation, accuracy prompts, community notes, fact-checks and indicators. By systematically comparing interventions under both flawless and erroneous conditions, our study contributes to a nuanced understanding of how interventions function in realistic environments across modalities and where errors are inevitable.

Our core contributions and findings are first, conducting a quantitative evaluation of the efficacy of five misinformation interventions, showing that *participants were more likely to recognize misinformation as inaccurate* when accompanied by fact-checks, indicators, and to a lesser extent by community notes, while *people were still just as likely to state they would share the content* compared to a control group. Second, we contrast erroneous and flawless interventions, showing that intervention errors critically determine efficacy: *only flawless versions* led to significant improvements in accuracy assessments. Third, we examine user perceptions of helpfulness and annoyance of the interventions, finding that *community notes, fact-checks, and indicators* were rated as *significantly more helpful but more annoying* than accuracy prompts. Moreover, *across all types*, interventions were perceived *more positively when operating without errors*. Fourth, we conduct a cross-platform comparison spanning TikTok videos, Telegram voice messages, and text-image combinations on X to illustrate how *both efficacy and user acceptance were robust across modalities* and social media platforms. Together,

these contributions advance a more realistic and nuanced understanding of how interventions are perceived under ideal and more realistic conditions, underscoring the importance of both effective and trusted approaches to mitigate the spread of misinformation online.

Acknowledgments

This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) in the project NEBULA (13N16361) and in the project CYLENCE (13N16636), and by the German Federal Ministry of Research, Technology and Space and the Hessian Ministry of Science and Research, Arts and Culture within their joint support of the National Research Center for Applied Cybersecurity ATHENE. We sincerely thank the participants whose participation played a crucial role in conducting this research.

References

[1] Sara Abdali, Sina Shaham, and Bhaskar Krishnamachari. 2025. Multi-Modal Misinformation Detection: Approaches, Challenges and Opportunities. *Comput. Surveys* 57, 3 (2025), 1–29. https://dl.acm.org/doi/10.1145/3697349

[2] Troy Adams, Yuanxia Li, and Hao Liu. 2020. A Replication of Beyond the Turk Alternative Platforms for Crowdsourcing Behavioral Research – Sometimes Preferable to Student Groups. *Transactions on Replication Research* 6 (2020), 1–22.

[3] Malik Almaliki. 2019. Online Misinformation Spread: A Systematic Literature Map. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining (ICISDM 2019)*. Association for Computing Machinery, New York, NY, USA, 171–178. https://doi.org/10.1145/3325917.3325938

[4] Nadia Alonso-López. 2021. Beyond Challenges and Viral Dance Moves: TikTok as a Vehicle for Disinformation and Fact-Checking in Spain, Portugal, Brazil, and the USA. *Analisi: Quaderns de Comunicacio i Cultura* 64, 1 (2021), 65–84.

[5] Alberto Ardevol-Abreu, Patricia Delponti, and Carmen Rodriguez-Wanguemert. 2020. Intentional or Inadvertent Fake News Sharing? Fact-checking Warnings and Users' Interaction with Social Media Content. *Profesional de la Informacion* 29, 5 (2020), 1–13.

[6] Kevin Autry and Shea Duarte. 2021. Correcting the Unknown: Negated Corrections May Increase Belief in Misinformation. *Applied Cognitive Psychology* 35, 4 (2021), 960–975.

[7] Carl-Anton Werner Axelsson, Mona Guath, and Thomas Nygren. 2021. Learning How to Separate Fake from Real News: Scalable Digital Tutorials Promoting Students' Civic Online Reasoning. *Future Internet* 13, 3 (2021), 60.

[8] Ranojoy Barua, Rajdeep Maity, Dipankar Minj, Tarang Barua, and Ashish Kumar Layek. 2019. F-NAD: An Application for Fake News Article Detection Using Machine Learning Techniques. In *2019 IEEE Bombay Section Signature Conference (IBSSC)*. Institute of Electrical and Electronics Engineers, Mumbai, 1–6.

[9] Corey H. Basch, Grace C. Hillyer, and Christie Jaime. 2022. COVID-19 on TikTok: Harnessing an Emerging Social Media Platform to Convey Important Public Health Messages. *International Journal of Adolescent Medicine and Health* 34, 5 (2022), 367–369.

[10] Melisa Basol, Jon Roozenbeek, Manon Berriche, Fatih Uenal, William P. McClanahan, and Sander van der Linden. 2021. Towards Psychological Herd Immunity: Cross-cultural Evidence for Two Prebunking Interventions against COVID-19 Misinformation. *Big Data & Society* 8, 1 (2021), 20539517211013868. https://doi.org/10.1177/20539517211013868

[11] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x

[12] Puneet Bhargava, Katie MacDonald, Christie Newton, Hause Lin, and Gordon Pennycook. 2023. How Effective Are TikTok Misinformation Debunking Videos? *Harvard Kennedy School Misinformation Review* 4, 2 (2023), 1–17.

[13] Tom Biselli, Katrin Hartwig, Niklas Kneissl, Louis Pouliot, and Christian Reuter. 2025. ChartChecker: A User-Centred Approach to Support the Understanding of Misleading Charts. In *Proceedings of the ACM Designing Interactive Systems Conference (DIS)*. ACM, Madeira, Portugal, 2075–2102.

[14] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve This Process—A Case Study of COVID-19. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 155 (2023), 33 pages.

[15] Judee K. Burgoon, J. P. Blair, Tiantian Qin, and Jay F. Nunamaker. 2003. Detecting Deception through Linguistic Analysis. In *Proceedings of Intelligence and Security Informatics (Lecture Notes in Computer Science)*, Hsinchun Chen, Richard Miranda, Daniel D. Zeng, Chris Demchak, Jenny Schroeder, and Therani Madhusudan (Eds.). Springer, Tucson, AZ, USA, 91–101.

[16] Man-pui Sally Chan and Dolores Albarracin. 2023. A Meta-Analysis of Correction Effects in Science-Relevant Misinformation. *Nature Human Behaviour* 7, 9 (2023), 1514–1525.

[17] Sijing Chen, Lu Xiao, and Akit Kumar. 2022. Spread of Misinformation on Social Media: What Contributes to It and How to Combat It. *Computers in Human Behavior* 141 (2022), 107643. https://www.sciencedirect.com/science/article/pii/S0747563222004630

[18] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 2024. Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (2024), 428:1–428:52. https://dl.acm.org/doi/10.1145/3686967

[19] William Davies. 2020. What's Wrong with WhatsApp. https://williamdavies.blog/2020/07/02/whats-wrong-with-whatsapp/

[20] Giandomenico Di Domenico, Daniel Nunan, and Valentina Pitardi. 2022. Marketplaces of Misinformation: A Study of How Vaccine Misinformation Is Legitimized on Social Media. *Journal of Public Policy & Marketing* 41, 4 (2022), 319–335.

[21] Carlos Diaz Ruiz and Tomas Nilsson. 2023. Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies. *Journal of Public Policy & Marketing* 42, 1 (2023), 18–35.

[22] Chiara Patricia Drolsbach, Kirill Solovev, and Nicolas Pröllochs. 2024. Community Notes Increase Trust in Fact-Checking on Social Media. *PNAS Nexus* 3, 7 (2024), pgae217. https://doi.org/10.1093/pnasnexus/pgae217

[23] Azza El-Masri, Martin J. Riedl, and Samuel Woolley. 2022. Audio Misinformation on WhatsApp: A Case Study from Lebanon. *Harvard Kennedy School Misinformation Review* 3, 4 (2022), 1–13.

[24] K. J. Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X. Zhang. 2023. Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 270 (2023), 42 pages.

[25] Diana Freed, Natalie N. Bazarova, Sunny Consolvo, Eunice J Han, Patrick Gage Kelley, Kurt Thomas, and Dan Cosley. 2023. Understanding Digital-Safety Experiences of Youth in the U.S.. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15.

[26] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

[27] Sukeshini Grandhi, Linda Plotnick, and Starr Roxanne Hiltz. 2021. By the Crowd and for the Crowd: Perceived Utility and Willingness to Contribute to Trustworthiness Indicators on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, GROUP (2021), 1–24.

[28] Brian Guay, Adam J. Berinsky, Gordon Pennycook, and David Rand. 2023. How to Think about Whether Misinformation Interventions Work. *Nature Human Behaviour* 7, 8 (2023), 1231–1233.

[29] Michael Hameleers and Toni van der Meer. 2023. Striking the Balance between Fake and Real: Under What Conditions Can Media Literacy Messages That Warn about Misinformation Maintain Trust in Accurate Information? *Behaviour & Information Technology* (2023), 1–14.

[30] Katrin Hartwig, Tom Biselli, Franziska Schneider, and Christian Reuter. 2024. From Adolescents' Eyes: Assessing an Indicator-Based Intervention to Combat Misinformation on TikTok. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–20.

[31] Katrin Hartwig, Frederic Doell, and Christian Reuter. 2024. The Landscape of User-centered Misinformation Interventions – A Systematic Literature Review. *ACM Computing Surveys (CSUR)* 56, 11 (2024), 1–36.

[32] Katrin Hartwig, Ruslan Sandler, and Christian Reuter. 2024. Navigating Misinformation in Voice Messages: Identification of User-Centered Features for Digital Interventions. *Risk, Hazards & Crisis in Public Policy* 15, 2 (2024), 203–235.

[33] Katrin Hartwig, Stefka Schmid, Tom Biselli, Helene Pleil, and Christian Reuter. 2024. Misleading Information in Crises: Exploring Content-Specific Indicators on Twitter from a User Perspective. *Behaviour & Information Technology* 0, 0 (2024), 1–34.

[34] Amelia Hassoun, Ian Beacock, Sunny Consolvo, Beth Goldberg, Patrick Gage Kelley, and Daniel M. Russell. 2023. Practicing Information Sensibility: How Gen Z Engages with Online Information. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17.

[35] Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep Learning for Misinformation Detection on Online Social Networks: A Survey and New Perspectives. *Social Network Analysis and Mining* 10, 1 (2020), 82.

[36] Pica Johansson, Florence Enock, Scott Hale, Bertie Vidgen, Cassidy Bereskin, Helen Margetts, and Jonathan Bright. 2022. How Can We Combat Online Misinformation? A Systematic Overview of Current Interventions and Their Efficacy. http://arxiv.org/abs/2212.11864

[37] Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of*

the ACM: Human Computer Interaction (PACM): Computer-Supported Cooperative Work and Social Computing 4, CSCW2 (2020), 1–27.

[38] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K. H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam J. Berinsky, Cornelia Betsch, John Cook, Lisa K. Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark Smith, Briony Swire-Thompson, Paula Szewach, Sander van der Linden, and Sam Wineburg. 2024. Toolbox of Individual-Level Interventions against Online Misinformation. *Nature Human Behaviour* (2024), 1–9. https://www.nature.com/articles/s41562-024-01881-0

[39] Chen Ling, Krishna P. Gummadi, and Savvas Zannettou. 2023. "Learn the Facts about COVID-19": Analyzing the Use of Warning Labels on TikTok Videos. *Proceedings of the International AAAI Conference on Web and Social Media* 17 (2023), 554–565.

[40] Jennifer S. Love, Adam Blumenberg, and Zane Horowitz. 2020. The Parallel Pandemic: Medical Misinformation and COVID-19. *Journal of General Internal Medicine* 35, 8 (2020), 2435–2436.

[41] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.

[42] Rakoen Maertens, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. 2021. Long-Term Effectiveness of Inoculation against Misinformation: Three Longitudinal Experiments. *Journal of Experimental Psychology: Applied* 27, 1 (2021), 1–16.

[43] Alexandre Maros, Anastasia Giachanou, Viktoria Spaiser, Francesca Spezzano, Anna George, Alexandra Pavliuc, Alexandre Bright, Jonathan, Jussara M. Almeida, and Marisa Vasconcelos. 2021. A Study of Misinformation in Audio Messages Shared in WhatsApp Groups. In *Disinformation in Open Online Media*, Vol. 12887. Springer International Publishing, Cham, 85–100.

[44] Cameron Martel and David G. Rand. 2024. Fact-Checker Warning Labels Are Effective Even for Those Who Distrust Fact-Checkers. *Nature Human Behaviour* 8, 10 (2024), 1957–1967. https://www.nature.com/articles/s41562-024-01973-x

[45] Ashlee Milton, Leah Ajmani, Michael Ann DeVito, and Stevie Chancellor. 2023. "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17.

[46] Lynnette Hui Xian Ng and Jia Yuan Loke. 2021. Analyzing Public Opinion and Misinformation in a COVID-19 Telegram Group Chat. *IEEE Internet Computing* 25, 2 (2021), 84–91.

[47] Shuo Niu, Zhicong Lu, Amy X. Zhang, Jie Cai, Carla F. Griggio, and Hendrik Heuer. 2023. Building Credibility, Trust, and Safety on Video-Sharing Platforms. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7.

[48] Katherine O'Toole. 2023. Collaborative Creativity in TikTok Music Duets. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16.

[49] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research. *Journal of Experimental Social Psychology* 70 (2017), 153–163. https://www.sciencedirect.com/science/article/pii/S0022103116303201

[50] G Pennycook, A Bear, ET Collins, and DG Rand. 2020. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *MANAGEMENT SCIENCE* 66, 11 (2020), 4944–4957.

[51] Gordon Pennycook, Jabin Binnendyk, Christie Newton, and David G. Rand. 2021. A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra: Psychology* 7, 1 (2021), 25293.

[52] Gordon Pennycook, Jabin Binnendyk, Christie Newton, and David G. Rand. 2021. A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra: Psychology* 7, 1 (2021), 1–13. https://online.ucpress.edu/collabra/article/7/1/25293/117809/A-Practical-Guide-to-Doing-Behavioral-Research-on

[53] 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. doi:10.1038/s41586-021-03344-2

[54] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson J. Lu, and David G. Rand. 2020. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science* 31, 7 (2020), 770–780.

[55] Sara Pluviano, Caroline Watt, and Sergio Della Sala. 2017. Misinformation Lingers in Memory: Failure of Three pro-Vaccination Strategies. *PLOS ONE* 12, 7 (2017), e0181640.

[56] Ethan Porter, Yamil Velez, and Thomas J. Wood. 2023. Correcting COVID-19 Vaccine Misinformation in 10 Countries. *Royal Society Open Science* 10, 3 (2023), 221097. https://royalsocietypublishing.org/doi/10.1098/rsos.221097

[57] Umair Qudus, Michael Röder, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. 2025. Fact Checking Knowledge Graphs – A Survey. *Comput. Surveys* (2025), 3749838. https://dl.acm.org/doi/10.1145/3749838

[58] Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *The World Wide Web Conference*. ACM, San Francisco CA USA, 818–828.

[59] Jon Roozenbeek, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. Psychological Inoculation Improves Resilience against Misinformation on Social Media. *Science Advances* 8, 34 (2022), eabo6254. https://www.science.org/doi/10.1126/sciadv.abo6254

[60] Margie Ruffin, Gang Wang, and Kirill Levchenko. 2022. Explaining Why Fake Photos Are Fake: Does It Work? *Proc. ACM Hum.-Comput. Interact.* 7, GROUP, Article 8 (2022), 22 pages.

[61] Anastasia Schaadhardt, Yue Fu, Cory Gennari Pratt, and Wanda Pratt. 2023. "Laughing so I Don't Cry": How TikTok Users Employ Humor and Compassion to Connect around Psychiatric Hospitalization. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[62] Philipp Schmid and Cornelia Betsch. 2022. Benefits and Pitfalls of Debunking Interventions to Counter mRNA Vaccination Misinformation During the COVID-19 Pandemic. *Science Communication* 44, 5 (2022), 531–558. https://journals.sagepub.com/doi/10.1177/10755470221129608

[63] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, Online, 899–908.

[64] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 15.

[65] Stefan Strauß. 2021. Deep Automation Bias: How to Tackle a Wicked Problem of AI? *Big Data and Cognitive Computing* 5, 2 (2021), 18. Issue 2. https://www.mdpi.com/2504-2289/5/2/18

[66] Yuko Tanaka and Rumi Hirayama. 2019. Exposure to Countering Messages Online: Alleviating or Strengthening False Belief? *Cyberpsychology, Behavior, and Social Networking* 22, 11 (2019), 742–746.

[67] Jack E. Taylor, Guillaume A. Rousselet, Christoph Scheepers, and Sara C. Sereno. 2022. Rating Norms Should Be Calculated from Cumulative Link Mixed Effects Models. *Behavior Research Methods* 55, 5 (2022), 2175–2196. https://link.springer.com/10.3758/s13428-022-01814-7

[68] Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. "It's Common and a Part of Being a Content Creator": Understanding How Creators Experience and Cope with Hate and Harassment Online. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15.

[69] Toni G. L. A. van der Meer, Michael Hameleers, and Jakob Ohme. 2023. Can Fighting Misinformation Have a Negative Spillover Effect? How Warnings for the Threat of Misinformation Can Decrease General News Credibility. *Journalism Studies* 24, 6 (2023), 803–823.

[70] Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. 2022. Do Humans Trust Advice More If It Comes from AI? An Analysis of Human-AI Interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 763–777. https://dl.acm.org/doi/10.1145/3514094.3534150

[71] Emily K. Vraga, Leticia Bode, and Melissa Tully. 2021. The Effects of a News Literacy Video and Real-Time Corrections to Video Misinformation Related to Sunscreen and Skin Cancer. *Health Communication* 37, 13 (2021), 1622–1630.

[72] Emily K. Vraga, Melissa Tully, and Leticia Bode. 2021. Assessing the Relative Merits of News Literacy and Corrections in Responding to Misinformation on Twitter. *New Media & Society* 24, 10 (2021), 2354–2371.

[73] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine* 240 (2019), 112552. https://linkinghub.elsevier.com/retrieve/pii/S0277953619305465

[74] WhatsApp. 2013. Introducing Voice Messages. https://blog.whatsapp.com/introducing-voice-messages

[75] WhatsApp. 2022. We're Making Voice Messages Even Better. https://blog.whatsapp.com/making-voice-messages-better?lang=af

[76] Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, Marina Del Rey CA USA, 637–645.

[77] Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated Fact-Checking: A Survey. *Language and Linguistics Compass* 15, 10 (2021), e12438. https://onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12438

A Appendix

Table 2: Questionnaires before, during, and after the online experiment.

Item	Answer Options
Before the experiment	
What is your age group?	[18-24; 25-34; 35-44; 45-54; 55-64; 65 or older]
Which most accurately describe(s) you?	[Woman; Man; Non-Binary; Prefer to self-describe (free-text field); Prefer not to answer]
What is the highest level of education you have completed?	[Less than high school; High school diploma or equivalent; Some college or vocational training but no degree; Bachelor's degree; Master's degree; Doctoral degree; Prefer not to say]
During the experiment (repeated for all 18 stimuli)	
If you were to see the prior content on social media, how likely would you be to share it?	[1. Extremely unlikely; 2. Moderately unlikely; 3. Slightly unlikely; 4. Slightly likely; 5. Moderately likely; 6. Extremely likely]
Are you familiar with the prior content (have you seen or heard about it before)?	[1. Not at all; 2. Slightly; 3. Somewhat; 4. Moderately; 5. Very much; 6. Extremely]
Assuming the prior content is entirely accurate, how important would the news be?	[1. Extremely unimportant; 2. Moderately unimportant; 3. Slightly unimportant; 4. Slightly important; 5. Moderately important; 6. Extremely important]
To the best of your knowledge, are the claims in the prior content accurate?	[1. Extremely inaccurate; 2. Moderately inaccurate; 3. Slightly inaccurate; 4. Slightly accurate; 5. Moderately accurate; 6. Extremely accurate]
After the experiment	
Which of the following best describes your political preference?	[Strongly Democratic; Democratic; Lean Democratic; Lean Republican; Republican; Strongly Republican; Independent; Other (specify in free-text format)]
How frequently do you watch or create content on the social media platform X or previously Twitter?	[Never; Once a month; Once a week; 2-6 times a week; Daily]
How frequently do you watch or create content on the social media platform TikTok?	[Never; Once a month; Once a week; 2-6 times a week; Daily]
How frequently do you listen to or send voice messages, for example on WhatsApp, Telegram or Signal?	[Never; Once a month; Once a week; 2-6 times a week; Daily]
Did you respond randomly at any point during the study? Note: Please be honest! You will get your payment regardless of your response.	[Yes; No]
During the experiment, did you notice any additional features or elements on the posts that seemed different from a typical [X-post/TikTok video/voice message on Telegram]?	[Yes; No; Not sure]
During the experiment, additional elements and features were provided as an intervention for some of the posts. Here is an example of the intervention that was presented: [include picture] Did you notice this intervention?	[Yes; No; Not sure]
How annoying did you find the above intervention during the experiment? (If you did not notice any, please select "Not applicable")	[Not at all annoying; Slightly annoying; Moderately annoying; Very annoying; Extremely annoying; Not applicable/I did not notice any intervention]
How helpful did you find the intervention for combating misinformation? (If you did not notice any, please select "Not applicable")	[Not at all helpful; Slightly helpful; Moderately helpful; Very helpful; Extremely helpful; Not applicable/I did not notice any intervention]
Did you notice the intervention making any mistakes? (If you did not notice any intervention, please select "Not applicable")	[Yes; No; Not sure; Not applicable/I did not notice any intervention]
Please shortly specify: What do you like or not like about the intervention as a measure to combat misinformation?	[free-text format]

Table 3: Short descriptions of the included nine misinformation posts and accurate information posts, specifying which three posts were erroneous as false negatives and which two posts were erroneous as false positives.

Stimulus Description	false positive or false negative
Misinformation	
Stimulus 1: Misinformation post claiming to show screenshots of Speaker Mike Johnson messaging on Grindr.	Yes
Stimulus 2: Misinformation post claiming that iPhone update secretly installs Starlink.	Yes
Stimulus 3: Misinformation post claims that people can get their student debt erased by filing Family Educational Rights.	No
Stimulus 4: Misinformation post claiming that Zelenskyy banned Truth Social in Ukraine.	No
Stimulus 5: Misinformation post claiming to show leaked audio of JD Vance criticizing Musk.	Yes
Stimulus 6: Misinformation post claiming to show Jimmy Kimmel in Epstein court documents.	No
Stimulus 7: Misinformation post claiming that the U.S. election in 2020 was stolen.	No
Stimulus 8: Misinformation post claiming that diseases can be healed by a healing sound watch.	No
Stimulus 9: Misinformation post claiming that poison is used before elections.	No
Accurate Information	
Stimulus 10: Accurate information about the production process of Parmesan cheese.	No
Stimulus 11: Accurate information about Lego donating Lego MRI scanners to hospitals.	Yes
Stimulus 12: Accurate information about AI training data in medical contexts is biased.	No
Stimulus 13: Accurate information about the U.S. being an outlier on paid parental leave.	No
Stimulus 14: Accurate information about not having to pay taxes on KitKat but on menstrual products.	Yes
Stimulus 15: Accurate information about photovoltaic power plants generating electricity.	No
Stimulus 16: Accurate information about growing potatoes without planting them in the ground.	No
Stimulus 17: Accurate information about Anitta being a significant part of Latin-American urban music.	No
Stimulus 18: Accurate information about growing different types of chickens.	No

Table 4: Demographic sample information

Age Group	N	%
18–24	112	11
25–34	185	18
35–44	170	17
45–54	160	16
55–64	231	23
65 or older	146	15
Gender		
Man	473	47
Woman	522	52
Non-Binary	1	1
Education		
Less than high school	8	1
High school diploma or equivalent	139	14
Some college or vocational training but no degree	273	27
Bachelor's degree	345	34
Master's degree	194	19
Doctoral degree	42	4
Prefer not to say	3	0
Total N	1004	

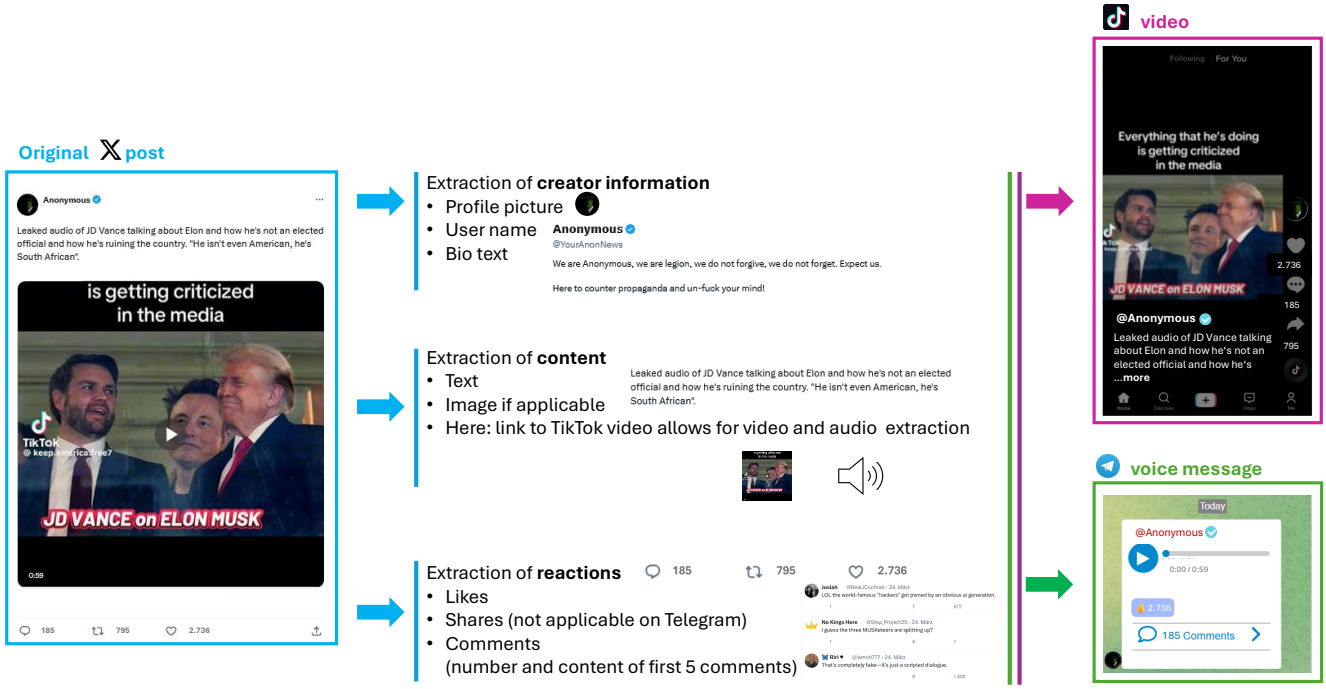


Figure 11: Exemplary flow of modality conversion from X to TikTok and Telegram.

Table 5: Sample sizes of the study groups of the experiment

	Telegram	TikTok	X
control group	33	35	37
inoculation (flawless)	34	33	34
accuracy prompt (flawless)	36	31	33
accuracy prompt (erroneous)	32	34	32
community note (flawless)	30	33	33
community note (erroneous)	36	30	35
fact check (flawless)	31	35	38
fact check (erroneous)	33	36	36
indicator (flawless)	32	35	31
indicator (erroneous)	34	30	32

Table 6: Contrasts using EMMs for accuracy ratings of misinformation (flawless interventions only)

contrast	estimate	SE	df	z-ratio	p-value
control group - inoculation	0.189	0.242	Inf	0.779	0.4669
control group - accuracy prompt	0.465	0.244	Inf	1.910	0.1053
control group - community note	0.594	0.247	Inf	2.410	0.0399
control group - fact-check	1.162	0.244	Inf	4.755	<.0001
control group - indicator	0.838	0.246	Inf	3.403	0.0033
inoculation - accuracy prompt	0.276	0.246	Inf	1.123	0.3270
inoculation - community note	0.406	0.249	Inf	1.627	0.1727
inoculation - fact-check	0.974	0.247	Inf	3.943	0.0006
inoculation - indicator	0.649	0.249	Inf	2.609	0.0272
accuracy prompt - community note	0.129	0.250	Inf	0.516	0.6058
accuracy prompt - fact-check	0.697	0.248	Inf	2.810	0.0186
accuracy prompt - indicator	0.372	0.250	Inf	1.491	0.2041
community note - fact-check	0.568	0.251	Inf	2.263	0.0507
community note - indicator	0.243	0.253	Inf	0.962	0.3878
fact-check - indicator	-0.325	0.250	Inf	-1.298	0.2651

Table 7: Cumulative Link Mixed Model predicting perceived accuracy ratings (including erroneous and flawless interventions).

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Fixed Effects:				
inoculation	-0.151	0.394	-0.383	0.702
accuracy prompt	-0.220	0.343	-0.642	0.521
community note	-0.532	0.344	-1.544	0.122
fact-check	-0.573	0.347	-1.650	0.099
indicator	-0.538	0.346	-1.558	0.119
TikTok	-0.001	0.392	-0.002	0.998
X	-0.230	0.386	-0.596	0.551
inoculation × TikTok	-0.140	0.557	-0.251	0.802
accuracy prompt × TikTok	0.195	0.482	0.404	0.686
community note × TikTok	-0.037	0.485	-0.075	0.940
fact-check × TikTok	0.024	0.482	0.049	0.961
indicator × TikTok	-0.010	0.485	-0.020	0.984
inoculation × X	0.061	0.550	0.110	0.912
accuracy prompt × X	-0.022	0.477	-0.047	0.963
community note × X	-0.191	0.478	-0.400	0.689
fact-check × X	-0.043	0.476	-0.090	0.928
indicator × X	-0.027	0.481	-0.056	0.955
Random Effects:	**Variance**	**Std. Dev.**		
Participant (Intercept)	2.566	1.602		
Post (Intercept)	0.198	0.445		
Thresholds:	**Estimate**	**Std. Error**	**z value**	
1|2	-0.871	0.321	-2.711	
2|3	0.285	0.321	0.886	
3|4	1.013	0.321	3.152	
4|5	2.315	0.323	7.176	
5|6	4.181	0.328	12.728	

Note: Reference categories used in the model are 'No intervention' and 'Social Media Platform: Telegram'. A Likelihood Ratio Test indicates a significant improvement over a null model ($\chi^2 = 28.0(17), p = 0.044$).

Table 8: Cumulative Link Mixed Model predicting sharing intentions for flawless condition.

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Fixed Effects:				
inoculation	-0.570	0.632	-0.902	0.367
accuracy prompt	-0.825	0.626	-1.317	0.188
community note	-1.285	0.663	-1.939	0.053
fact-check	-0.832	0.647	-1.287	0.198
indicator	-1.153	0.651	-1.771	0.077
X	-0.194	0.611	-0.317	0.751
TikTok	-0.264	0.623	-0.423	0.672
inoculation × X	0.292	0.878	0.333	0.739
accuracy prompt × X	0.078	0.880	0.088	0.930
community note × X	1.241	0.904	1.372	0.170
fact-check × X	0.119	0.879	0.135	0.893
indicator × X	0.307	0.911	0.337	0.736
inoculation × TikTok	0.707	0.885	0.800	0.424
accuracy prompt × TikTok	0.592	0.893	0.663	0.507
community note × TikTok	1.149	0.912	1.259	0.208
fact-check × TikTok	0.756	0.892	0.847	0.397
indicator × TikTok	0.950	0.896	1.060	0.289
Random Effects:	**Variance**	**Std. Dev.**		
Participant (Intercept)	5.469	2.339		
Post (Intercept)	0.052	0.229		
Thresholds:	**Estimate**	**Std. Error**	**z value**	
1|2	0.542	0.454	1.193	
2|3	1.515	0.455	3.333	
3|4	2.120	0.455	4.656	
4|5	3.180	0.457	6.957	
5|6	4.659	0.462	10.081	

Note: Reference categories used in the model are 'No intervention' and 'Social Media Platform: Telegram'. A Likelihood Ratio Test does not indicate a significant improvement over a null model ($\chi^2 = 11.5(17), p = 0.826$).

Table 9: Ordinal logistic regression results for helpfulness

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
inoculation	0.133	0.413	0.322	0.748
community note	0.826	0.335	2.467	0.014
fact-check	1.069	0.331	3.235	0.001
indicator	0.625	0.336	1.860	0.063
Telegram	-0.076	0.345	-0.221	0.825
X	-0.450	0.341	-1.319	0.187
inoculation × Telegram	-0.134	0.605	-0.221	0.825
community note × Telegram	0.049	0.475	0.104	0.917
fact-check × Telegram	0.232	0.470	0.493	0.622
indicator × Telegram	0.127	0.477	0.266	0.790
inoculation × X	0.616	0.585	1.053	0.292
community note × X	0.931	0.470	1.982	0.047
fact-check × X	0.710	0.467	1.520	0.128
indicator × X	0.975	0.474	2.057	0.040
Thresholds:	**Estimate**	**Std. Error**	**z value**	
1|2	-1.393	0.250	-5.583	
2|3	-0.195	0.240	-0.812	
3|4	0.827	0.241	3.436	
4|5	1.900	0.248	7.662	

Note: Reference categories used in the model are 'no intervention' and 'Social Media Platform: TikTok'. A Likelihood Ratio Test indicates a significant improvement over a null model ($\chi^2 = 75.2(14), p < 0.001$).

Table 10: Ordinal logistic regression results for annoyance.

Parameter	Estimate	Std. Error	z value	Pr(>|z|)
inoculation	0.859	0.573	1.499	0.134
community note	0.909	0.456	1.992	0.046
fact-check	1.354	0.440	3.074	0.002
indicator	1.252	0.445	2.818	0.005
Telegram	-0.241	0.541	-0.445	0.657
X	0.381	0.494	0.771	0.441
inoculation × Telegram	0.867	0.809	1.073	0.283
community note × Telegram	-1.360	0.767	-1.773	0.076
fact-check × Telegram	-0.611	0.664	-0.920	0.358
indicator × Telegram	-0.141	0.658	-0.214	0.831
inoculation × X	0.138	0.777	0.178	0.859
community note × X	-2.270	0.767	-2.958	0.003
fact-check × X	-1.174	0.623	-1.883	0.060
indicator × X	-1.468	0.648	-2.264	0.024
Thresholds:	**Estimate**	**Std. Error**	**z value**	
1|2	1.712	0.361	4.741	
2|3	3.018	0.377	8.013	
3|4	3.945	0.404	9.775	
4|5	4.910	0.467	10.509	

Note: Reference categories used in the model are 'no intervention' and 'Social Media Platform: TikTok'. A Likelihood Ratio Test indicates a significant improvement over a null model ($\chi^2 = 60.7(14), p < 0.001$).

Table 11: Contrasts using EMMs for helpfulness.

contrast	estimate	SE	df	z-ratio	p-value
accuracy prompt - inoculation	-0.294	0.245	Inf	-1.201	0.2553
accuracy prompt - community note	-1.153	0.195	Inf	-5.911	<.0001
accuracy prompt - fact-check	-1.383	0.196	Inf	-7.065	<.0001
accuracy prompt - indicator	-0.992	0.196	Inf	-5.068	<.0001
inoculation - community note	-0.860	0.240	Inf	-3.575	0.0007
inoculation - fact-check	-1.090	0.241	Inf	-4.521	<.0001
inoculation - indicator	-0.698	0.241	Inf	-2.897	0.0063
community note - fact-check	-0.230	0.185	Inf	-1.244	0.2553
community note - indicator	0.161	0.187	Inf	0.862	0.3884
fact-check - indicator	0.391	0.187	Inf	2.097	0.0514

Table 12: Contrasts using EMMs for annoyance.

contrast	estimate	SE	df	z-ratio	p-value
accuracy prompt - inoculation	-1.194	0.322	Inf	-3.712	0.0010
accuracy prompt - community note	0.301	0.328	Inf	0.916	0.3995
accuracy prompt - fact-check	-0.759	0.266	Inf	-2.855	0.0086
accuracy prompt - indicator	-0.716	0.270	Inf	-2.652	0.0133
inoculation - community note	1.495	0.348	Inf	4.300	0.0002
inoculation - fact-check	0.435	0.289	Inf	1.504	0.1659
inoculation - indicator	0.478	0.293	Inf	1.631	0.1469
community note - fact-check	-1.060	0.297	Inf	-3.571	0.0012
community note - indicator	-1.017	0.300	Inf	-3.384	0.0018
fact-check - indicator	0.043	0.231	Inf	0.187	0.8519