"Helps me Take the Post With a Grain of Salt:" Soft Moderation Effects on Accuracy Perceptions and Sharing Intentions of Inauthentic Political Content on X

Filipo Sharevski
DePaul University

Verena Distler
Aalto University

Florian Alt
Ludwig Maximilians Universität
University of the Bundeswehr Munich

Abstract

In this study, we empirically evaluate the effectiveness of soft moderation interventions on *X*—warning labels, warning bundles (labels combined with community notes), and warning covers—in reducing perceived accuracy and sharing intentions of inauthentic political content across two contexts: (1) the 2024 U.S. presidential election, and (2) a non-election setting. Using a sample of $n_1 = 925$ *X* users during the election campaign, we find that both the warning bundle and the warning cover significantly reduce the perceived accuracy of manipulated content related to each presidential candidate. A follow-up evaluation with $n_2 = 649$ *X* users after the election confirms these findings, reinforcing the role of such interventions as interaction frictions that effectively lower perceived accuracy of inauthentic content concerning both politically affiliated individuals and global conflict topics. Thematic analysis of participants' explanations suggests that warning labels and covers – especially those incorporating third-party fact-checks – are viewed as less trustworthy than the community notes included in the warning bundles. Across both contexts, no intervention significantly impacted participants' willingness to share the content, mainly due to concerns that sharing inauthentic material could harm self-presentation on *X*.

1 Introduction

The openness and anonymity of social media have led to several unintended consequences, including the proliferation of inauthentic accounts, behaviors, and, more recently, content. Inauthentic accounts (trolls, sockpuppets, and bots) regularly disrupt platforms by posting offensive material, spreading misinformation, or manufacturing the illusion of widespread discontent on polarizing issues. Often orchestrated by centralized entities, these accounts operate as coordinated botnets that engage in inauthentic behavior aimed at advancing divisive narratives, undermining civic and democratic institutions, and conducting information operations. This has been notably documented in foreign interference campaigns targeting the 2016 [52] and 2020 [19] U.S. presidential elections.

Platforms, in response, are forced to "remove" (hard moderate) inauthentic accounts, "reduce" visibility of problematic behavior, or "inform" (soft moderate) users about why content is problematic. Launching a disinformation campaign became a non-trivial task, and efforts, evidence shows, shifted towards proliferation of *inauthentic* or synthetically created, altered, or otherwise digitally manipulated photos, videos, or audio [62]. Inauthentic political content was seen as the main threat to the 2024 US elections [8] and was closely monitored for causing disruption, as memes and fake news did in the 2014 cycle [51]. Inauthentic content is a continuous threat that may result in widespread confusion on public or contentious issues – so platforms keep a close eye on it [45, 64].

X, in particular, implemented a structured response to detect, correct, and contain inauthentic content appearing on the platform [64]. As illustrated in Figure 3, a range of soft moderation interventions was deployed to inform users about problematic posts, including warning labels (Figure 3a), community notes (Figure 3b), and warning covers (Figures 3c, 3d, and 3e). From the perspective of user interaction effort [25], these interventions introduce increasing levels of friction: from low (warning labels), to medium (warning labels combined with community notes), to high (covers incorporating independently fact-checked information). At the lowest level, users encounter a simple warning message with no need for action; at the medium level, an extended community-contributed context about the post's authenticity is added; and at the highest level, they must actively click to reveal the content or consult fact-checks from two independent sources [11]. This graduated approach enables platforms like *X* to moderate content based on the severity of inauthenticity, while balancing for not too much obtrusiveness and user interruption.

As users objected to X's moderation both in the context of US elections [47] and in general [46], it is worth studying the aforementioned interventions applied to inauthentic content in both contexts. In the past, warning labels (low friction) did not achieve the intended correction and containment [34], but the community notes (medium friction) [13] and covers (high friction) did [27]. To the best of our knowledge, there is no

evidence yet of how effective the soft moderation approach is in decreasing the perceived accuracy and curbing the sharing intentions of inauthentic content, and with that influence their civic engagement or voting intentions.

To address this gap, we evaluated the effectiveness of soft moderation interventions both in the lead-up to the 2024 U.S. presidential election and in a non-election context. For the election setting, we selected two posts containing inauthentic content: (1) a photo of Kamala Harris [40], and (2) a video of Donald Trump [39]. Given that content characteristics influence perceived trustworthiness [60], and recognizing that inauthentic content is also moderated on *X* outside of electoral periods, we sought to include a diverse range of content types to mitigate potential bias from a single medium or event. Accordingly, for the non-election context, we selected six posts featuring inauthentic content: (1) an audio clip of JD Vance [48]; (2) a photo of Donald Trump [42]; (3) a photo of Tim Walz [41]; (4) a video of a staffer speaking on behalf of Kamala Harris [61]; (5) a video related to the Russo-Ukrainian conflict [22]; and (6) a photo from the Israeli-Palestinian conflict [23]. In both contexts, we investigated the following research questions:

- **RQ1:** How do *soft moderation interventions* in increasing order of friction – (i) a warning label; (ii) a warning label followed by a community note (bundle); and (iii) a warning cover – affect the *perceived accuracy* of an X post containing inauthentic political content, compared to a non-moderated version of the same X post?
- **RQ2:** How do *soft moderation interventions* in increasing order of friction – (i) a warning label; (ii) a warning label followed by a community note (bundle); and (iii) a warning cover – affect the *sharing intentions* of an X post containing inauthentic political content, compared to a non-moderated version of the same X post?
- **RQ3:** Controlling for the users' *intention to vote*, how do *soft moderation interventions* affect the (a) *perceived accuracy*; and (b) *sharing intentions* of an X post containing inauthentic political content, compared to a non-moderated version of the same X post?
- **RQ4:** Controlling for the users' *demographics*, how do *soft moderation interventions* affect the (a) *perceived accuracy*; and (b) *sharing intentions* of an X post containing inauthentic political content, compared to a non-moderated version of the same X post?
- **RQ5**: What is the nature of *meaning-making* around the (a)*perceived accuracy* and (b)*sharing intentions* on X posts containing inauthentic political content (some of which contain soft moderation interventions)?

Using a sample of $n_1 = 925$ participants, we found statistically significant evidence that the medium (warning bundle) or high level (warning covers including third-party fact-checked information) of frictions decreased the perceived accuracy of both the inauthentic photo and video compared to a condition where no soft moderation was applied on an X post *during* the US Election 2024. The follow-up evaluation *outside* of any election context ($n_2 = 649$) *confirmed* these results for all types of inauthentic content we tested (**RQ1**). Regardless of the election context, in both evaluations, none of the soft moderation interventions, however, had an effect on our participants' sharing intentions relative to any of the eight (two plus six) inauthentic content types we tested (**RQ2**).

As inauthentic content on social media, in the context of elections, attempts to influence voting outcomes [52], we controlled for the participants' voting intentions in both the US Elections 2024 and in general. We found that the friction-countering effect on the perception of accuracy does not depend on the intention to vote in both contexts. The soft moderation frictions had no effect on the sharing intentions, regardless of the voting intentions in both contexts (**RQ3**). We found that those participants who frequently share content on X were more likely to perceive the inauthentic content as accurate and share it, in both the US Election 2024 context and outside of it. In both contexts, participants living in the rural areas or suburbs were less inclined to share the inauthentic content compared to the urban living participants (**RQ4**).

Bringing the qualitative component of the participants' justifications of their accuracy/sharing decisions in conversation with the results above, in both contexts, we found that participants trusted the low and high friction (warning labels and warning covers that contain third-party fact checked information) *less* compared to the medium frictions (warning labels bundled with the community notes, which include debunking information provided by other X users). Sharing, in both the US Election 2024 and outside of it, was seen as not worthy of an endeavor as it might potentially hurt the participants' reputable persona they maintain on X (**RQ5**).
The contributions of this work are threefold:

1. We introduce a mixed-methods methodology for quantitatively and qualitatively evaluating the effects of soft moderation interventions on users' perceptions of accuracy and intentions to share real-world *inauthentic political multimedia content* on social media.
2. We provide an operational framework for assessing how inauthentic political content influences users' *voting intentions*—a core aim of information operations and disinformation campaigns—and offer empirical evidence that medium- and high-friction soft moderation interventions can effectively *counter* this influence.
3. We present evidence that *community notes* are the most trusted mechanism for debunking inauthentic political content on social media, both in election contexts and beyond, aligning with moderation strategies based on *crowdsourcing* adopted by, e.g., X and Meta [44, 59].

2 Background and Related Work

2.1 Social Media Content Moderation

Content moderation on social media is a process encompassing (1) *definition* of problematic content (terms of use, community guidelines) and moderation policies; (2) *detection* of content candidates for moderation (automatic, third-party/user input); and (3) *enforcement* of moderation policies (interventions per post, per account, per topic). While it is expected that content moderation might differ from platform to platform (assuming the platforms are willing to engage with it[1]), research shows that the content moderation yields platform-independent effects relative to the reception of such an intervention among social media users. Social media users do object to the opaqueness of the moderation policies, voice skepticism about unfair policy enforcement, and resist moderation interventions on the platforms [16].

Problematic content, with a high probability to cause individual and/or societal harm, pertains to hate speech, offensive language, bullying and harassment, m/disinformation, spam/phishing, violence, graphic content, sexual abuse, self-harm, or intellectual property violation. Platforms, thus, maintain policies to monitor for such candidate content usually through automatic means, though for topics such as m/disinformation, platforms receive input from third parties (e.g., fact-checkers) [33] or user communities [12]. Upon determination that content is problematic and in violation of the community guidelines/terms of use, platforms enforce the moderation policies – following the "remove, reduce, inform" approach – through either *hard moderation* (e.g., content removal, account suspension, reduced visibility, and down-ranking) [29] or *soft moderation* (e.g., provision of information cues such as contextual or interstitial content covers, warnings, labels, or tags) [27]. Platform moderation efforts are also influenced by legal requirements, for instance, the European Union's Digital Services Act, which requires large social media companies to put systems in place to control the spread of misinformation [21].

The user reception of algorithmic moderation decisions has been far from positive. The main objections to hard moderation come from experiences of unfair and unjust silencing, inability for self-recovery and obtaining community support, shadowbanning, disproportionate targeting of users from minority groups, and cumbersome appeal process. The main objections about soft moderation come from experiences of alternative narrative silencing, inability for self-expression and exercising the right for free speech, selective application, and the disproportionate targeting of political out-groups. The application of soft moderation interventions on social media also saw a dismissal of corrections (sustained or increased belief in falsehoods) and problematic content amplification (engagement with and increased sharing, retweeting, liking, or commenting of falsehoods) [16].

Misinformation not only tries to circumvent moderation, but might result in *non-conformity and resistance to correction/containment* among the user population. The disciplinary action of hard moderation and especially the corrective/containment action of soft moderation resulted in *continued influence effect* (using inaccurate information in reasoning after a credible correction has been presented), *ideological backfiring effect* (increasing belief based on pre-existing views due to information repetition within a correction). While the non-conforming response to hard moderation merely shifts the misinformation on platforms without any moderation (allowing *illusory truth effect* or increasing belief due to information repetition in the absence of correction), resistance to misinformation correction and containment remains an open soft moderation issue [16].

2.2 Moderating Misinformation

The resistance to misinformation correction and containment is caused by the platforms balancing between moderation and engagement as not to push users away. Correction-wise, platforms experimented with *content covers* that obscured the problematic misinformation content and required users to click through to see it. Evidence showed that these *content covers* worked [53], though they fell out of favor as too interruptive. Platforms moved to more of an unobtrusive approach with *interstitial warning tags and labels* [34] that do not interrupt the user interacting with the content or compel any action but offer information from third-party sources for topics often subject to misinformation [43, 57, 58, 66].

Platforms blended these warnings with the interface and used evidence from correcting misinformation offline, that is, linking to expert sources and correcting quickly and early. Platforms also avoided generalized interventions (e.g., "Disputed" tags) and instead organized them by per topic (e.g., "Get the facts about COVID-19" labels) or source associated with a topic (e.g., "Russia state-affiliated media" tags for content related to the Russo-Ukrainian conflict [1]). The lack of context in the warning text was a deliberate design choice to avoid ostracizing as the public correction might be experienced as embarrassing or confrontational. As platforms had to carry and attract as many users as possible, the minimal design of the soft moderation interventions also balanced for users' prior values, preferences, and beliefs, as well as their media literacy skills.

Despite all of this effort, platforms did not anticipate how the users would behave when the soft moderation interventions were applied to highly politicized or politically contentious topics. Though interstitial warnings were found to work for politically concordant content [38], they failed to reduce the perceived accuracy and intention to share content among the politically discordant user base [53]. An analysis

[1] Alt-platforms dismiss content moderation altogether under the "marketplace of ideas" model of free expression on social media.

of the engagement with the labeled Trump's election tweets from 2020, for example, shows that the soft moderation intervention did not decrease the sharing of the labeled false content [47] among politically concordant users. In addition, the application of the misinformation warning labels on mainstream platforms was seen as a deliberate attempt to silence right-leaning or other unpopular opinions in favor of a perceived left-leaning platform's ideology.

To avoid the impression of a punitive and politically-biased moderator, platforms such as X shifted the moderation responsibility to users instead, creating the so-called *community notes* [59]. The community notes retained the interstitial format of the platform-provided misinformation warnings and tags, though the context and correction were left to be determined by the (non-expert) users themselves. The moderation then became a crowdsourced fact-checking as it involved users with both left-leaning and right-leaning political worldviews who selected falsehoods, crafted the debunking information, and assigned the community notes themselves. The whole concept is consensus-dependent, so the community must first reach an agreement for a given content – which they choose based on a broader policy targeting falsehoods to be addressed and contextualized to the other users [59].

Initially, the shift to user-provisioned content moderation appeared successful, with many users demonstrating a reduced perception of the accuracy of the labeled content. [12]. But, later analyses found that this approach might be too slow to decrease the perceived accuracy and sharing intentions of misinformation content in the early (and most viral) stage of diffusion [7]. Another problem that beset the community notes was that the non-expert users often struggle to separate their political or ideological biases when evaluating claims, leading to disputes and biased fact-checks [2]. The dependence on consensus was shown to be selectively applied, usually to surface-level misinformation, leaving nuanced topics unaddressed for correction and containment. Though limited as such, evidence shows that the community notes are promising way of addressing the resistance to correction and contain misinformation on social media, as they allow for a more verbose context and explanation to be included in the warning labels substantiating misinformation content [13].

2.3 Moderating Inauthentic Content

Most misinformation content that was the subject of soft moderation resistance to correction and containment, was assumed to be created by deliberate, authentic methods. Usually, the misinformation content in question is textual or memetic in nature, and the soft moderation interventions have a limited effect towards decreasing the perceived accuracy and the sharing intention of the content [46]. But social media saw a surge in synthetically created, altered, or otherwise digitally manipulated content that also needed to be corrected and contained as it constitutes (1) misinformation by design (e.g., a deepfake); and (2) might be misleading nonetheless by including elements of authentic content (e.g., out-of-context speech or video, superimposed imagery, PhotoShop alternations) [63].

Figure 1: A "Manipulated Media" warning on Twitter [30]

Platforms started to add warnings about "Manipulated media," as shown in Figure 1 for X (then Twitter), though so far, the evidence about their effect to decrease the perception of accuracy and dissuade users from sharing is limited (platforms also recognize and label "AI-generated" content so one could distinguish between a partial or full fabrication). Early experiments show that warnings do help, though only marginally, as the majority of users (80%) were unable to spot an inauthentic video [37]. Sort of a *truth-tainting* effect was noticed when interstitial warnings were applied to political deepfake videos in that social media users believed that any substantiated video was fake, even if the video was real [55]. Unlike the ambiguous wording of the standard misinformation interventions, the interstitial warnings about synthetically generated and manipulated content were found to be comprehensible and well-received among social media users [18].

While these warnings have received at least some attention towards helping users discriminate between authentic and synthetic, altered, or manipulated content, there is no work done on the effectiveness of these warnings relative to their main goal: decrease the perceived accuracy of the content and dissuade users from sharing it. Absent from the current knowledge of whether soft moderation would comprehensively work for inauthentic content is evaluation with one found on platforms, as the one used in the aforementioned experiments was created by the researchers instead. Though the correction and containment were applied by platforms in the basic form as early as 2020 [30], there is no evidence of how they fare in the community note variant nor when covers obscure the inauthentic content, a soft moderation variant that recently found its way back on social media [6].

While the initial approach of soft moderating synthetic or manipulated media opted for a general text and unobtrusive warnings (e.g., "AI-generated," "Synthetic Media" [6] or the one shown in Figure 1), platforms felt that they need all hands

on deck, given that it becomes increasingly impossible for people to spot any synthetic content, be that audio, image, text, or video [10]. Therefore, platforms expanded the interstitial warning labels with reference to fact-check information (Figure 3a), bundled and assigned the basic warning labels with the community notes underneath altered content (Figure 3b), and added back the verbose covers for both images and videos (Figure 3c and 3d). Seeing the soft moderation from the perspective of a *counter-misinformation* friction[2] it appears that platforms departed from the conventional approach of minimal obtrusivness in the otherwise uninterrupted platform engagement. It remains to be seen whether the shift from a contextualized exposure for correction/containment (back) to a preventive avoidance might work. Therefore, we set out to explore the effectiveness of counter-misinformation frictions applied to inauthentic political content, in both election and non-election context.

3 Methodology

3.1 Recruitment and Study Protocol

Before initiating our recruitment and sampling, we obtained approval from the ethical review board to conduct surveys (Appendix A). We sought X (Twitter) users, aged 18 and above, from the US. The participants were recruited through Prolific to take a 5-minute (on average; maximum of 30 minutes) survey through Qualtrics for compensation of $\$12/hour$ (average \$1, maximum \$6; total of \$1888 overall).

For the *US Election 2024* setting (two evaluated posts), we conducted a power analysis aimed at detecting even a small effect size of $d = 0.25$, which indicated a requirement of approximately 116 participants per group (8 groups total). Data collection continued up to the final day of campaigning before Election Day (November 4$^{\text{th}}$, 2024), yielding 1168 survey responses. After excluding responses that failed attention checks or were deemed low quality (e.g., response times under 5 seconds), our final sample comprised $n_1 = 925$ participants. For the *non-election* setting (six evaluated posts), we revised the power analysis based on the observed effect sizes from the US Election 2024 study and aimed to detect at least a medium effect size of $d = 0.40$. We recruited 28 participants per group across 24 groups. Following the same cleaning procedure, the resulting sample was $n_2 = 649$ participants.

In both evaluations, after the consent, participants were randomly assigned to one of the 8 or 24 groups. Each participant was exposed to an interactive X post – either a photo, video, or audio – selected from a third-party fact-checked database of posts containing inauthentic media [22, 23, 39–42, 48, 61]. Participants had the option to click the links for more information (e.g., "Find Out More" in Figure 3a, see the link to the authentic content in Figure 3b, or "See Why" and "See Post" in Figure 3c, 3d, and 3e) as our goal was to emulate as realistic interaction as possible with the each of the study stimuli. Participants were then asked what is their perceived accuracy of the post (**RQ1**) and what are their sharing intentions if this post was seen in their natural X feed (**RQ2**). For each question, participants were also asked to provide an open-ended answer about how they determined the content accuracy and the reasoning behind the indicated sharing intentions (**RQ5**).

For **RQ3**, we specifically controlled for participants' *intention to vote*, as prior work on soft moderation (cf.Section 2) indicates users' political ideology influences how they engage with misinformation on social media. However, we chose to use voting intention as our variable rather than political ideology for two key reasons. First, the deliberate spread of misinformation on social media is often aimed at influencing voter behavior, as demonstrated in the previous two U.S. presidential election cycles [15]. Second, we sought to avoid potential misinterpretation of our findings under the politically charged claim that misinformation research disproportionately targets conservative viewpoints. Unlike political ideology, which tends to be stable over time, voting intention is subject to change between election cycles [28].

For **RQ4**, we additionally controlled for the participants' *demographics* or the following items: *age*, as it is a factor in sharing intentions of misinformation [26]. Equally, there is no concrete information on how the perception of accuracy and the sharing intentions of inauthentic content differ based on the *geographic area* where the participants reside, except that states or counties during the 2016 US elections where people engaged more with misinformation tended to vote for Donald Trump [20]. As engagement with misinformation on social media factors in the susceptibility and sharing propensity, we also collected the participants' *engagement on X*.

Participants were allowed to skip any question, request support during the survey, or abandon the survey at any point without penalty. Given that the posts presented to participants contained misleading claims and that the eight control groups (two election-related and six non-election-related) were shown content without any form of debunking or soft moderation, we provided an extensive debriefing at the end of each survey (Appendix D). This was important, especially considering that engagement with the soft moderation interventions in the other conditions could not be guaranteed. To support transparency, we provided participants with access to the authentic versions of the manipulated or altered content used in the study [22, 23, 39, 40, 42, 48, 61]. After the debriefing, participants had the option to request removal of their data before exiting the survey. Since the survey was conducted anonymously, data removal could not be processed after participants had left. We ensured participants had sufficient time to make this decision and issued completion codes regardless of whether they reviewed the debriefing content.

[2] Counter-misinformation intervention is friction that is designed to mitigate the risk of incorrect belief or further misinformation spread by lowering the perceived accuracy and lowering the sharing intention without affecting the overall episodic user experience on a social media platform [11].

3.2 Study Stimuli

For all of our study stimuli, we used the AFP and Reuters databases of fact-checking 910 reports to select real-world content that has been tagged as synthetic, doctored, altered, or manipulated content and pertains to: (1) each of the candidates in the run-up of the US Elections 2024; and (2) right-leaning individuals, left-leaning individuals, or ongoing conflicts – topics with an inherent political connotation. In each of the eight (two plus six) cases, we first checked to see how each of the platforms applied soft moderation to the inauthentic media. Instagram, TikTok, and Facebook have not applied any warning labels, notes, or covers about the inauthentic photo or video in question. The only platform that did so was X.

The content in question is initially assigned the warning label with the "Manipulated Media" text and the warning favicon (Figures F.1, F.2, F.3, F.4, F.5, F.6, F.7, F.8), offering a link to find out more about the inauthenticity of the media included in the post[3]). Later, the non-expert users added community notes explaining the alteration in the media as "added context by the readers" that X urges users to consider while accessing the post. Usually, the context is a link to the unaltered or the original media.

A cover variant of soft moderation interventions was also added for the posts. The button "See why" opens a dialog window linking two independent fact-checked reports from AFP and Reuters with a brief description of the alteration for each of the media In each of the interventions that had links, participants were able to click on them. The participants in the warning label group were able to find out more about the inauthentic media policy at X; the ones in the group with both the warning label and the community note had the ability to also check the authentic content, and the ones in the cover were able to click the "See why" button and access both the fact-checked links.

3.2.1 US Election 2024 Settings

The inauthentic photo and video used in the *US Election 2024* context are shown in Figure F.1 [40] and Figure F.2 [39] (Appendix F). For Kamala Harris, we selected a photo, as a disinformation campaign during that period attempted to falsely associate her with Sean 'Diddy' Combs, who was facing criminal allegations. For Donald Trump, we used a video that falsely claimed he had staged an assassination attempt in order to generate political support.

3.2.2 Non-Election Settings

The inauthentic content selected for the *non-election* context included two posts targeting right-leaning figures: Vice President JD Vance (Figure F.3) and President Donald Trump (Figure F.4). The first was chosen to introduce media diversity, featuring an audio deepfake instead of a photo or video. The second served as an alternative modality to the video of Trump used in the US Election 2024 setting. We also included two posts targeting left-leaning individuals: one featuring Vice Presidential candidate Tim Walz (Figure F.5), and another depicting a staffer speaking on behalf of Kamala Harris (Figure F.6). The Walz post mirrors the role targeted in the JD Vance case, while the Harris staffer video complements the photo tested in the election context. We also included two posts tied to politically charged global conflicts: one manipulating a video about casualties in the Russo-Ukrainian conflict (Figure F.7), and another using a doctored photo of an alleged Israeli attack in the Israel-Palestinian conflict (Figure F.8).

3.2.3 Order of Friction

From a user interaction perspective [25], the design of the warning labels enabled a *low level* of friction in that it (1) allows the user to get exposed to the inauthentic content; and (2) it only warns the user that the content is manipulated, offering only general description on what that means on the "Find out more" link. The community notes, paired with warning labels, offer a *medium level* friction as they also allow for users to see the inauthentic content, but now users have the option to see the authentic part too as part of the moderation design that aims to mitigate the risk of poor accuracy perception or further misinformation spread. The covers, on the other hand, offer a *high level* of friction as they (1) prevent the user exposure to the inauthentic content itself, (2) the user is informed that independent fact-checkers agreed that the content is inauthentic (a credibility marker); (3) the user needs to explicitly choose to press "See Post" to see the content; and (4) the user has the option to press the "See Why" button to see the fact-checked information and access the authentic content.

The order of low, medium, and high further aligns with the concept of counter-misinformation (or general usable security-enhancing) frictions [11]. The momentary negative effect in the case of the warning label is the shortest one, while the user needs a bit more time to see and perhaps even read/access the community note to ensure that the content is inauthentic. In the case of the cover, the user might not even see the inauthentic content at all and proceed with scrolling down their feed, and even if they do, they have to actively press a button (in case they want to see why, this will take the longest as they have the option to verify both independent fact-checkers).

3.3 Risks and Ethical Considerations

Studying misinformation with political connotations requires careful ethical judgment, risks assessment, and appropriate protections. In addition to the IRB approval, we developed our ethical decisions and protection protocols through extensive deliberation on our study's goals, guided by known computer security research dilemmas [9, 56] (see our Ethical Deliber-

[3] https://help.x.com/en/rules-and-policies/authenticity

ation). We proceeded with the study involving the exposure to the manipulated social media content (that we sourced directly from X as it appeared in its natural form, but we took the precaution to blur the posters' avatars and X handles), though only under strictly controlled conditions.

We implemented protocols to balance between obtaining results that would benefit the social media user base, on one side, and respecting the participants' self-agency and dignity, on the other. As part of the consent process, we informed participants that the study aimed to evaluate the use of soft moderation on content posted to *X*. To avoid biasing responses, we did not explicitly state our focus on inauthentic content. However, this was clarified during the debriefing, where we established that the soft moderation examined pertained to digitally altered, manipulated, or synthetically generated media—commonly referred to as deepfakes, cheapfakes, or fake media [36]. We emphasized that participation was both anonymous and voluntary, and assistance is available any point during or after the study. Participants were also encouraged to consult the research team regarding any inauthentic content they encountered on social media, regardless of whether it was accompanied by a warning.

As part of the content selection process, we followed the fair user policy and focused on content that has already reached large audiences in the context of X (and possibly other platforms). We deliberately avoided exposure to content that might create excessive distress, offend, or potentially cause triggering effects as a result of the claims involved. As part of the debriefing, we communicated the potential difficulties and/or risks that might arise from the exposure to inauthentic content, even though the media we selected – at the time of the study – were independently demonstrated to be a misinformative content, in addition to warning labels, community notes, and sources participants had the opportunity to see during the survey. We stated that our study is not undertaken from any (geo)political or electoral perspective, even though it involves politically related content.

We debriefed participants that we do not act on behalf of or receive funding from X, other platforms, any political campaign, or party. We pointed out that we are impartial to any variant of moderation (soft and hard) and that we do not have a preference relative to how inauthentic content is handled from a moderation perspective in general. We also indicated that there is a possibility that, after the publication of the results of the study, one could misuse or misinterpret them in the broader context of social media participation. As such, an occurrence is out of our control; we notified the participants that we could not prevent it, but we are open to discussing how to mitigate potential adverse effects in coordination with X, interested platforms, or involved stakeholders. Here, we stressed that our role as misinformation researchers is to substantiate evidence in a proactive public conversation about moderating inauthentic content, a benefit that spans beyond the narrow, US-centric use for political or campaigning purposes.

3.4 Data Collection and Analysis

As our dependent variables were measured on a 10-point scale with only the extremes being labeled (see Appendix A), it is reasonable to assume that they represent equally spaced judgments. A meta-study shows that all response modes (scale items) robustly measure the accuracy and sharing construct when it comes to misinformation [49]. We avoided a smaller scale (e.g., categorical or 4/5-points) as we wanted to capture the granularity in decisions, factoring for the effect of the level of friction on the accuracy perceptions and sharing intentions (or the absence of it). Also, smaller scales might be misconstrued as an attention checks. The 10-point scale was also preferred as it didn't have a neutral value.

We thus treated them as interval measurements and used linear models. We ran linear regression models with displayed friction as an independent variable, and perceived accuracy as the dependent variable (**RQ1**), and friction as independent variable, and sharing intention as the dependent variable (**RQ2**). We used the R package "lmtest" to check the model assumptions through visual inspection. For RQ1 and RQ2, the assumptions of linear regression (linearity, independence, homoscedasticity, and lack of multicollinearity) were met. The residuals were not normally distributed, but violations of the assumption of normally distributed residuals have not been found to bias results with larger sample sizes (defined as, for instance, $N > 40$ total [3] at least 10 observations per variable [50]). For **RQ3** and **RQ4**, as we controlled for the participants' intention to vote and demographics, we conducted a robust linear regression to account for heteroscedasticity. Throughout the analysis, we used the customary p-value threshold of 0.05.

As collected open answers about how participants determined the content accuracy and reasoned about sharing it (**RQ5**), we followed the steps of the practical guide for doing thematic analysis outlined in [4]. The decision to do thematic analysis was part of the research design process, where we, collectively as a research team, recognized that the evaluation of soft moderation would be incomplete if we did not inquire and bring the evidence of the nature of contemporary meaning-making of these interventions in the context of inauthentic content. We felt that the degree of arbitrariness and the anonymous survey, in addition, allowed us to collect quality, insightful, and relevant information from participants that, in turn, offered sufficient *information power* for our analysis (as a sample sizing approach, instead of achieving saturation [5]).

Each of the research members situated themselves both as *insiders* (in the sense that we encounter inauthentic content that is subject to soft moderation on X) and *outsiders* (in the sense that we have examined misinformation, inauthentic content, and moderation of social media). We all felt quite comfortable disengaging in our responses to inauthentic content and soft moderation to push the responses to our questions about *perceived accuracy* and *sharing intentions*

into the analytical foreground, and interrogating them. In the familiarization phase of our analysis [4, p. 42], we immersed and critically engaged with the data. This helped us to approach the coding and develop both semantic and latent codes in relation to the data (Appendix C). Two researchers from the team engaged in a mostly inductive coding process (though drawing on evidence-driven knowledge for deduction where needed) through multiple rounds for the purpose of collaboratively gaining richer or more nuanced insights (instead of reaching an agreement about every code) [4, p. 55].

We then shifted our focus from codes to themes in order to develop the patterns or meaning across our dataset and cluster the codes around central organizing concepts about inauthentic content. We first generated initial themes [4, p.78], then, working through a thematic mapping, we developed and reviewed the resultant themes [4, p.97]. The continuous analytical refinement enables us to define and name the themes [4, p.108]. Lastly, we wrote our results around the themes, selecting data extracts to evidence our claims and allow one to independently judge the fit between the data and our understanding and interpretation of them [4, p.133].

4 Results

4.1 Sample Composition

925 participants were included in the first evaluation dataset, as shown in Table 1. We were particularly interested in how the inauthentic content might affect the intention to vote from a disinformation perspective, so we balanced the sample according to the reported voting intentions. Each of the 8 conditions we tested was shown to 116 participants on average ($Min_1 = 109$, $Max_1 = 123$). 649 participants were included in the follow-up evaluation dataset as shown in Table 1. Using the results of the first evaluation and the adjusted effect size, we replicated the same methodology from above, with the assumption that the intention to vote is for future elections (as the Elections 2024 have passed), so we retained the sample balancing according to the reported voting intentions. Each of the 24 conditions we tested was shown to 27 participants on average ($Min_2 = 23$, $Ma_2 = 30$). Overall, both the perceived accuracy and shared intentions were right-skewed, thus tending toward more negative values for both.

Category	**US Election 2024**		**Non-Election**	
	Count	**(%)**	**Count**	**(%)**
Intention To Vote				
Democratic Party	329	35.57	197	30.35
Republican Party	277	29.95	191	29.43
I don't plan to vote	119	11.57	53	8.17
Other party / Independent	107	12.86	127	19.57
Undecided	93	10.05	127	19.57
Gender				
Female	430	46.49	331	51.00
Male	478	51.68	307	47.30
Non-Cis	17	1.84	11	1.69
Area of Residence				
Large City	227	24.54	209	32.2
Suburbs	360	38.92	268	41.29
Small city/town	195	21.08	92	14.18
Rural area	143	15.46	80	12.33
Voted Before				
Yes	679	73.41	59	9.09
No	227	24.54	583	89.83
Rather Not Say	19	2.05	7	1.08
Category	***M*** **(σ)**	*Median*	***M*** **(σ)**	*Median*
Age	38.31 (11.92)	37	40.31 (13.43)	38.00
X Engagement				
Spent Time	3.66 (2.75)	3	4.67 (3.03)	5.00
Sharing Posts	2.27 (2.06)	1	3.24 (2.58)	2.00

Table 1: Demographics

4.2 RQ1: Perceived Accuracy

The results of our linear regression analysis for the soft moderation interventions' effect, both in the context of the US Elections 2024 and in a non-election context, on the *perceived accuracy* are shown in Table 2. In both cases, we found that the warning covers as high friction (including the option to read fact-checked information) and the community notes bundled with warning labels as a medium level friction had a statistically significant negative effect on the perceived accuracy ($p < .001$). In the US Elections 2024 context, the overall effect, capturing the contribution of the soft moderation interventions into making the decision about the accuracy perception, was $d = .43$. The particular effect for the warning bundle was -0.8 and for the warning over was -1.2. For the non-elections, the overall effect was $d = .17$, with -1.06 effect for the warning bundle and -1.10 for the warning cover. We did not find a statistically significant effect of the warning label (low-level friction) on the perceived accuracy of any of the eight (two plus six) posts. These results show that the soft moderation interventions decrease the perceived accuracy when applied with medium or high levels of friction variants.

Key Finding 1: Warning labels bundled with community notes or warning covers effectively decrease the perceived accuracy of an inauthentic content on X, both ***during*** the US Election 2024 and ***outside*** of the election context. The medium and high frictions work regardless of the subject of the inauthentic content.

4.3 RQ2: Sharing Intentions

The results of our linear regression analysis for the effect of the soft moderation interventions, both in the US Elec-

US Election 2024 Settings	
Displayed Friction (Warning)	**Perceived Accuracy**
Label (F.1, F.2)	0.106(0.243)
Bundle (F.1, F.2)	-1.058^{***}(0.246)
Cover (F.1, F.2)	-1.104^{***}(0.244)
Constant	4.044^{***}(0.174)
Observations	925
R^2	0.045
Adjusted R^2	0.042
Non Election Settings	
Displayed Friction (Warning)	**Perceived Accuracy**
Label (F.3, F.4, F.5 F.6, F.7, F.8)	-0.441(0.275)
Bundle (F.3, F.4, F.5, F.6, F.7, F.8)	-0.843^{***} (0.274)
Cover (F.3, F.4, F.6, F.6, F.7, F.8)	-1.245^{***} (0.275)
Constant	4.560^{***} (0.192)
Observations	649
R^2	0.034
Adjusted R^2	0.029
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Soft Moderation Effect on Perceived Accuracy

US Election 2024 Settings	
Displayed friction	**Sharing Intentions**
Label (F.1, F.2)	0.144(0.213)
Bundle (F.1, F.2)	-0.161(0.217)
Cover (F.1, F.2)	-0.405^{*}(0.215)
Constant	2.302^{***}(0.153)
Observations	925
R^2	0.008
Adjusted R^2	0.005
Non-Election Settings	
Displayed friction	**Sharing Intentions**
Label (F.3, F.4, F.5 F.6, F.7, F.8)	-0.512^{*} (0.282)
Bundle (F.3, F.3, F.4, F.6, F.7, F.8)	-0.084(0.281)
Cover (F.3, F.4, F.6, F.6, F.7, F.8)	-0.421(0.282)
Constant	2.893^{***} (0.197)
Observations	649
R^2	0.007
Adjusted R^2	0.003
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Soft Moderation Effect on Sharing Intentions

tions 2024 and in the non-election context, on the *sharing intentions* are shown in Table 3. None of the soft moderation interventions had a statistically significant effect on the sharing intentions. That is, the level of friction has no containment effect, regardless of context or inauthentic content type.

Key Finding 2: Soft moderation interventions did not have a statistically significant effect on sharing intentions of inauthentic content on X, regardless of election context.

4.4 RQ3: Intention to Vote

The results of the robust linear regression investigating the soft moderation effect on the *perceived accuracy*, controlling for users' intention to vote in the US Elections 2024, are shown in Table E.1 and in general in Table E.2. For both contexts, we included an interaction term for the frictions and the intention to vote to check whether the soft moderation's effect varied depending on the voting intentions. The interaction term was not statistically significant, for any level of soft moderation intervention friction.

Key Finding 3: Warning labels bundled with community notes or warning covers effectively decrease the perceived accuracy of inauthentic content on X, regardless of voting intentions, both ***during*** the US Election 2024 and ***outside*** of the election context.

The inauthentic post depicting candidate Harris was perceived as more accurate than the post with candidate Trump, with a medium effect ($p < 0.01$, $d = .65$). The inauthentic posts depicting left-leaning individuals were perceived as more accurate than the posts depicting right-leaning individuals or conflict-related topics with a medium effect ($p < 0.01$, $d = .56$). In the US Elections 2024 context, participants with voting intentions Republican, Other/Independent, or did not plan to vote reported higher perceived accuracy compared to participants intending to vote Democrat or were undecided ($p < 0.05$, $d = .65$). In the non-election context, participants with voting intentions Other/Independent or were undecided reported higher perceived accuracy compared to those with intention to vote Democrat ($p < 0.05$, $d = .56$).

Key Finding 4: Inauthentic content targeting left-leaning individuals on X is perceived as more accurate, regardless of election context, and regardless of the intention to vote.

Key Finding 5: Those undecided or those intending to vote Democrat reported lower perceived accuracy of inauthentic content compared to the other voting intentions ***during*** the US Election 2024. In the non-election context, participants with voting intentions Other/Independent or were undecided reported higher perceived accuracy compared to those with intention to vote Democrat.

The results of the robust linear regression investigating the soft moderation effect on the *sharing intentions*, controlling for users' intention to vote in the forthcoming US Elections 2024, are shown in Table E.1. The corresponding results for the non-election context are given in Table E.2. We included an interaction term for the frictions as a variable and the intention to vote variable to check whether the soft moderation intervention varied across the voting intentions. None of the frictions had a statistically significant effect on sharing intention in either of the contexts or the content selected.

Key Finding 6: Soft moderation interventions had no effect on sharing intentions of inauthentic content on X, regardless of one's voting intentions, both ***during*** the US Election 2024 and ***outside*** of the election context.

4.5 RQ4: Demographics

We conducted a robust linear regression model with *perceived accuracy* as a dependent variable, controlling for demographic effects, in the context of the US Elections 2024, as shown in Table E.1. The corresponding results for the non-election context as shown in Table E.2. In both contexts, participants who shared content on X more frequently had a higher perceived accuracy score for all of the posts we evaluated with a medium effect ($p < 0.001$, $d \geq .63$).

Key Finding 7: Frequent content sharing on X was associated with a higher perceived accuracy of inauthentic content, both ***during*** the US Election 2024 and ***outside*** of the election context.

Participants who had voted in previous elections perceived both the US Elections 2024 content as less accurate than participants who had never voted before for a US president ($p < 0.01$, $d = 0.63$). Older participants showed a higher perceived accuracy score for all of the six posts we evaluated outside of the US Election 2024 context ($p < 0.05$, $d = 0.69$).

Key Finding 8: Participants who had voted in previous US election cycles showed lower perceived accuracy of inauthentic content compared to participants who had never voted before for a US president, ***during*** the US Election 2024; Older participants also showed higher perceived accuracy, ***outside*** of the election context.

We conducted a robust linear regression model with the *sharing intentions* as a dependent variable, controlling for demographic effects. The results are presented in Table E.1 and Table E.2. None of the soft moderation interventions had a statistically significant effect on the sharing intentions of either of the eight (two plus six) posts we tested.

Key Finding 9: We did not find a statistically significant effect of the soft moderation interventions on sharing intentions when controlling for demographic variables.

In the US Election 2024 context, participants living in suburbs near a large city reported lower sharing intention ($p < 0.01, d = 0.63$) compared to participants living in a large city. In the non-election context, participants living in rural areas reported lower sharing intention ($p < 0.01$, $d = 0.69$) compared to participants living in a large city. Participants who shared content on X more frequently reported higher sharing intention in both contexts ($p < 0.01$, $d \geq 0.63$).

Key Finding 10: Participants living in a large US city showed higher sharing intentions of inauthentic content, regardless of the election context.

Key Finding 11: Frequent content sharing on X was associated with a higher sharing intention of inauthentic content, both ***during*** the US Election 2024 and ***outside*** of the election context.

4.6 RQ5-a: Perception of Accuracy Formation

Research shows that users have no more than a guessing advantage in forming the accuracy perception of whether content is authentic or synthetically generated/manipulated [10]. If the content quality alone is insufficient, then one would reasonably attempt to look for external inputs – usually surrounding its presentation – in making a meaningful accuracy perception. This approach emerged as a broader pattern of meaning-making in both of our US Elections 2024 and non-election datasets, or what we named *reliance on situated trustworthiness*. Relative to this overarching theme, we developed two themes with their own central organizing concepts [4, p.113]:

(1) **X as an authenticity arbitrator** – focuses on not whether the *content* is authentic, but whether one could rely on X's authoritative credibility to moderate inauthentic political content in the first place, including the structuring and application of the community notes;

(2) **Content association cues of authenticity** – focuses on the intrinsic influence of the broader association with a given content, to accurately depict its accuracy when encountered on X.

The US Election 2024 participants did not forget the disturbance that the X's (then Twitter's) soft (and hard) moderation caused during the 2020 US elections [47]. The objections against the labeling then-president Donald Trump, banning the New York Post article about Hunter Biden's laptop (son of Joseph Biden, then-democratic presidential candidate for US Elections 2024), and deplatforming prominent far-right voices (e.g., Ben Garrison, Marjorie Taylor Greene) lingered in the memory of the users who did not intend to vote Democrat. Here, $\mathbf{P_1435}$ (Republican; warning cover video: F.2 and F.2) and $\mathbf{P_1262}$ (Undecided; warning label photo: F.1) recalled these incidents as a reason "*not to trust any of the X's labels since their reputation for moderation is really bad*".

Participants not intending to vote Democrat reasoned that fact-checkers are humans and could fall for inauthentic content. Participant $\mathbf{P_1402}$ (Republican; warning cover video: F.2 and F.2) expressed they "*don't necessarily trust X's 'fact checkers' because a lot of what they are fact-checking can be considered opinion or they themselves cannot tell if something is accurate or has been altered*". $\mathbf{P_1718}$ (Other Party/Independent; warning cover photo F.1 and F.1) noted:

They "*haven't done much research regarding 'independent fact checkers' so before trusting them [they] would need to determine where their intentions lie*."

The mistrust in fact-checkers was also present in the follow-up evaluation. Participant **$P_2$634** (Other Party/Independent; audio warning cover F.3 and F.3) expressed: "*I believe that [the content] is edited, but the organizations who are 'fact-checking' lean right typically and even use language which suggests that they have a certain bias (e.g., 'phony')"*. Participant **$P_2$316** (Republican; photo warning cover F.5 and F.5) said: "*Tim Walz is a freak so I would not put anything pass him and I do not rely on biased fact checkers; I will search for my own truth*." And participant **$P_2$232** (I don't plan to vote; video warning cover F.7 and F.7) pointed they "*don't trust most information off of X so even with the independent fact checker I am still not sure what to believe*"

There was consensus among the US Election 2024 participants, regardless of their voting intentions, that the community notes could be trusted to help determine the authenticity of content on X as a user, rather than a third-party-based, authenticity double-check. **$P_1$141** (Democrat; photo warning bundle F.1) noted that they "*tend to trust X's community notes as they go by community agreement and therefore it's harder for the system to be outright biased*." **$P_1$720** (Republican; video warning bundle F.2) and **$P_1$163** (Other Party/Independent; video warning bundle F.2) also surmised that "*if a post has had context added through community notes, which are generally somewhat reliable, it most probably contains manipulated content*." Participant **$P_1$704** (Republican, photo warning bundle F.1) praised the community notes as "*they helped [them] take the post with a grain of salt*."

The bundle of a warning label and the community note (medium friction) also emerged as a "*solid and reliable information*" in the follow-up evaluation too, as **$P_2$296** (Republican; warning bundle, no warning F.3) put it. Regardless of the follow-up participants' voting intentions or political identity, it is the application of the "*community notes [that helps with noting the content's inauthenticity]*" because **$P_2$465** (Other Party/Independent; audio warning bundle F.3) felt one "*could not be sure to trust much X these days*" to do a proper moderation. The "third-party" fact-checks worked because with the community notes, these are "*multiple-party" fact-checks*", according to **$P_2$107** (Democrat; video warning bundle F.6).

Interestingly, the candidates in our US Election 2024 stimuli emerged as an input of trustworthiness of the content itself, often in tension with the applied soft moderation. Participants like **$P_1$132** (Republican; warning label photo F.1) thought the photo was authentic, unaffected by the warning label, stating that "*this photo is just another proof that Kamala Harris lacks judgment and has no integrity; she knew exactly who she was standing next to in this photo and people need to know the truth*." For participant **$P_1$265** (Republican; warning bundle photo F.1) the content was authentic, again despite both the label and the community note, "*because people need to know that Kamala has ties to Diddy, who's a criminal*." Participant **$P_1$762** (Republican; warning cover photo F.1 and F.1) conceded to the high friction soft moderation, but nonetheless commented that "*It's inaccurate, yes, but Kamala is crooked enough, pics do not need to be altered to show that*."

Similarly for Donald Trump, **$P_1$223** (Democrat; warning bundle video F.2) ignored both the label and the community note, stating that they "*wouldn't believe it even with the fact check on the bottom, he might be crazy and he has committed fraud*". For **$P_1$221** (Other Party/Independent; warning label F.2), the perceived accuracy was determined by the manipulation of an already inauthentic event: "*leave it to Trump to do something stupid like fake an assassination attempt, and you'll get the truth*." Participant **$P_1$632** (Democrat; warning bundle video F.2 and F.2) agreed with the community notes that the video was mirrored, but nonetheless stated that "*the mirroring doesn't in and of itself makes the content inaccurate, Trump would do anything to rile up his voter base; faking the assassination is totally plausible*"

Participants in the follow-up evaluation responded in a similar manner. Participant **$P_2$308** (Other Party/Independent; no warning F.3) stated: "*JD Vance is his own worst enemy – Obviously Elon has affected him, but the content is accurate just by the nature of the subject and JD Vance's idiocy.*" Participant **$P_2$482** (Democrat; no warning F.4) dismissed the warning label, commenting: "*I think it's a little exaggerated, but basically accurate – Trump is very authoritarian and he is not for anyone but himself and the rich*." Participant **$P_2$382** (Undecided; no warning F.5) also saw no issue with the content about Tim Walz: "*based on what [they] already know about [him], seeing a Little Black Sambo cake (or whatever that is) in the background did not surprise [them] a bit – he is definitely a weirdo*." In a reversal of the warning bundle debunking, participant **$P_2$13** (Republican; warning bundle video F.6) remarked: "*I don't think it's true because she's a part of the Democratic Party, and they absolutely hate and despise guns and weaponry of any form – It would be shocking to me if Kamala actually owned a firearm*."

Key Finding 12: Perceptions of accuracy depend on the perceived trustworthiness of the surrounding context. Compared to community notes, labels, and third-party fact-checks were viewed as less credible. A politically incongruent personal disposition toward the subject of the content tended to shift perceived accuracy toward authenticity, even when the content was demonstrably inauthentic.

In both contexts we evaluated, the *prevalence of opinion* regarding the warning labels (low friction) was that casting content as a "manipulated media" might not be a sufficient cue alone against the perceived accuracy. Though the labeled helped the participants suspect all the content we tested, many expressed that "*it might be manipulated media [what they saw], but parts of it could be true*" (**$P_1$631** – I don't plan to

vote; warning label photo: F.1). The opposite opinion prevailed when the warning label was reinforced by an explanation offered in the community notes (medium friction). What resonated with most of the users is that the community notes allow not for "one link supplied by X" or "couple of links supplied by independent fact-checkers" but "*a lot of different links that show a content is fake supplied by other X users*" (**P$_2$263** – Republican; warning bundle audio: F.3). In simple words, the community notes are less probable to be *biased* compared to the labels and covers with independent fact-checked information (high friction) because "*no one fact-checks the fact-checkers*" (**P$_1$421** – Republican; warning cover photo: F.1 and F.1), while "*X users fact-check among each other and break down the inauthenticity for impartial judgment*" (**P$_2$207** – Republican; warning bundle audio: F.3).

4.7 RQ5-b: Determining Sharing Intentions

The decision to share (in)authentic content was primarily driven by participants' concerns about their *reputation or self-presentation* on social media, across both datasets. Prior research on misinformation sharing has largely focused on its relationship to perceived accuracy and the intention to inform others about salient developments [24]. However, accurately assessing inauthentic content is often more challenging than identifying conventional misinformation or fake news. This makes it more difficult for users to maintain a consistent self-presentation, as they risk damaging interpersonal relationships built through online interactions [14].

Importantly, a user's *reputation or self-presentation* on social media reveals little about the *nature* of that reputation. While some individuals may strive to project a serious or credible persona, others may opt to appear humorous or build social ties through entertainment. Yet others, might see a social media presence bound to responsibility in informing followers and other users about deceiving content. To capture this nuance, we identified two additional themes, each with its own central organizing concept [4, p.113]:

(1) **Maintain serious social media persona** – focuses on the *absence* of sharing intentions of (in)authentic (soft moderated) political content, in any context, due to the threat of reputation loss and damaging interpersonal relationships on social media;

(2) **Maintain responsible social media persona** – focuses on *presence* of sharing intentions of (in)authentic, (soft moderated) political content, in any context, due to the potential of reputation gain and strengthening interpersonal relationships on social media.

Sharing (in)authentic political content might be damaging to self-presentation, even if it was not moderated, because one could not regain their reputation by removing or subsequently sharing verified, authentic content later. Participant **P$_1$887** (Democrat; no warning F.1) statement testifies, here, that "*don't feel comfortable sharing political content even if I agree with it because it just seems like it would invite more attention and negative comments, even if I retract my action, and I don't want to deal with that.*" Participant **P$_2$15** (Undecided; no warning F.6) seconded: "*I don't share political news. I don't want to stir up drama with friends.*"

The soft moderation, even with a minimal level of friction, made **P$_1$556** (Undecided; warning label video F.2) abstain from sharing because "*the manipulation was done to rage bait others on social media, and [they] don't want to be part of it.*" Likewise, participant **P$_2$144** (I don't plan to vote; warning label audio F.3) added: "*I rarely share anything and this would make me look stupid to share because it is too outlandish.*" The damaging effect was well pronounced by **P$_1$190** (Don't plan to vote; warning label video F.2) who said that they "*wouldn't share it only to avoid all of the backlash from people thinking [they] support one side of the political spectrum more than the other; It's not about politics anymore it's about hate nowadays on social media.*"

As to why one could share inauthentic content, **P$_1$153** (Democrat; warning bundle photo F.1) said they "*would only share this post because the photo is funny; However, [they] feel like everyone should do more research before taking news articles and photos at face value.*" Participant **P$_2$154** (Democrat; warning bundle photo F.8) agreed, and added: *It's a witty and relatable way to describe instability, using humor to comment on real-life turbulence; I'd share it because it captures a common feeling in a memorable, lighthearted way.*" The participants in the follow-up dataset also felt that share inauthentic content, according to participant **P$_2$379** (Democrat; warning bundle video F.6) is needed "*to inform others of the incorrect information being spread on social media.*" Sharing, in the view of **P$_2$380** (Democrat; warning cover video F.8 and F.8), should be done "*to can see other people's perspective on the same if they agree that it is accurate or not.*"

Participants found that the soft moderation interventions could factor reputation- and relationship-wise, even if one shares the content for humor or entertainment purposes. Participant **P$_1$197** (Republican, warning label video F.2) hinted at the contested nature of labels in the past: "*I would share this Trump meme on because it's too funny not to; Plus, I know the labels would spark some conversation, and maybe even a few laughs.*" The humor, with an option for serious factual discernment using the soft moderation interventions, was seen as a good opportunity for **P$_1$89** (Democrat; warning bundle photo F.1) to strengthen interpersonal relationships: "*I am confident that the post contains an altered photo; I might share it just as an example to my own followers and mutuals as election misinformation and how the community has called it already.*" Here, **P$_2$638** (Other party/Independent; warning bundle audio F.3) felt sharing would "amplify" the effect of the community notes "*because other people need to see [the note] a lot as people don't look at these kind of things.*"

Key Finding 13: Sharing inauthentic (soft moderated) content on social media is restrained by the perceived risk of reputation and relationship damage: Sharing might happen as an act of responsibility, though, counting on the warning labels and community notes as a counterbalance against legitimizing misinformation on social media.

In both contexts we evaluated, the *prevalence of opinion* regarding all the soft moderation interventions was that they "reinforce" the feeling one should refrain from sharing inauthentic content on social media, regardless of who is the target of manipulation or the context. Participants, in both datasets, were reluctant to contribute to an inauthentic behavior as "*sharing [the content] would risk amplifying a misleading narrative and could contribute to public confusion or manipulation*" (**$P_2$74** – Democrat; warning cover video: F.7 and F.7). In simple words, "*sharing inauthentic content, even with warnings, notes, or covers*" (**$P_2$78** – Republican; warning bundle: F.8) is, for the time being, "*too controversial*" (**$P_1$402** – Republican; warning cover video: F.2 and F.2).

5 Discussion

5.1 Implications

The results presented per research question (for brevity) are inherently hard to decouple and treat in isolation. As such, they deserve to be brought into the conversation as the analysis in Sections 4.2, 4.3, 4.4, and 4.5 uncovered *whether* the soft moderation interventions work in any context, while Section 4.6 and 4.7 offer the dimensions of *why* (some of) the soft moderation interventions work among ordinary X users, regardless of elections and regardless of the content's topic.

5.1.1 Perceived Accuracy

Our results across both evaluation contexts indicate that perceptions of media authenticity are supported by medium-to-high-friction soft moderation warnings. In their absence, users on *X*—and on social media more broadly—face significant difficulty in discerning deceptive content, particularly when motivated reasoning is involved [17]. Elements of such reasoning were especially evident in participants' assessments of authenticity in the *US Election 2024* context, as revealed by our qualitative data. Although susceptibility to misinformation is linked more to a lack of analytical thinking than to one's political alignment [31], our findings suggest a more complex picture. Specifically, when considering users' voting intentions and perceptions of accuracy, we find that warning labels paired with community notes, as well as warning covers, contribute to—but do not solely drive—analytical engagement in evaluating whether content is accurate.

The perception of *X* as an unreliable arbiter of authenticity –particularly among participants who did not intend to vote Democrat in the (*US Election 2024*) or who intended to vote Republican (*non-election*) suggests that trust in the platform's moderation practices has not been fully restored, despite the elaborate options for soft moderation. In this light, Meta's recent decision to discontinue third-party fact-checking and adopt a community notes model instead [30] is perhaps unsurprising. As our findings reinforce the prior evidence [12]), community notes' consensus-driven corrections contributed by users across the political spectrum [59] represent a promising approach for mitigating the effects of motivated reasoning, both *during* election periods and *outside* them.

Consistent with past evidence, we also show that people use political individuals as a heuristic to guide the evaluation of what is true or false, yet do not necessarily insist on verifiable, platform-provided soft moderation as a prerequisite for supporting (or opposing) political candidates or officials [54]. The problem of "political animus" is perhaps harder to address, at least in the context of election campaigning [32] but also outside of the election context (as we showed), and remains an open question on how to deal with the increasingly convincing inauthentic content as a potential catalyst of affective polarization [16]. An interesting avenue of potential future exploration is to retain the soft moderation interventions, though to include not just a link – according to the misinformation inoculation theory – but also information about how inauthentic content could drive affective polarization [35].

5.1.2 Sharing Intentions

The containment of inauthentic content, as suggested by our results across both contexts, appears to be a promising outcome. Past research indicates that to effectively discourage users from sharing misinformation is to prompt reflection, typically by asking them to confirm the content's accuracy or commit to pro-social behavior at the moment of sharing [27]. While our study did not evaluate such interactive interventions, our mixed-method findings suggest that these mechanisms may not be universally effective or even applicable. Regardless of intervention type, our participants often expressed reluctance to share inauthentic content simply because they could not confidently assess its authenticity. This inability to objectively verify accuracy may, in itself, be enough to deter sharing, even without additional prompts or mechanisms.

What may warrant further attention is the inclusion of explicit information about how sharing a post could potentially mislead other users on the platform. Based on such information, users could then be asked to actively endorse their intention to share. This approach offers an alternative to conventional soft moderation, shifting from reliance on fact-checker-provided information – which, as our qualitative findings show, is often met with skepticism – to a more user-centered reflection mechanism. Importantly, this type of intervention would still depend on the presence of at least a warning bundle combining a label and a community note.

Such a bundle would serve as a necessary "fallback," offering contextual correction in cases where users object to the act of endorsement itself or perceive it as implicating them in the spread of misinformation, regardless of their intent.

5.2 Reflections on Soft Moderation Design

Given the clear departure from standalone warning labels—which, according to our findings and prior research, have limited effectiveness [34]—and the declining reliance on third-party fact-checkers, which remain effective but increasingly distrusted [27], community notes appear to be the most viable soft moderation intervention currently available to users. While community notes may lag in addressing viral content [7], the nature of inauthenticity as verifiably false information, due to the existence of authentic counterparts, may help prevent disputes, biased fact-checks, and issues of selective application [2]. Their strength lies in their flexibility, contextual depth, and foundation in user consensus [13].

To maximize the potential of community notes, we see value in designing them to explicitly "inform" users about how inauthentic content may contribute to political polarization, as illustrated in Figure 2. The format and wording of such warnings aim to preserve the perceived neutrality of the community note system while encouraging user reflection on the broader implications of manipulated content. This proposal builds on empirical evidence showing that reflection can be effective when integrated into soft moderation strategies. Of course, such a design might "backfire" or yield unintended sharing effects, but given the substantial threat posed by highly deceptive political content—not only in elections but across civic discourse—we argue that evolving user-driven soft moderation is both necessary and timely.

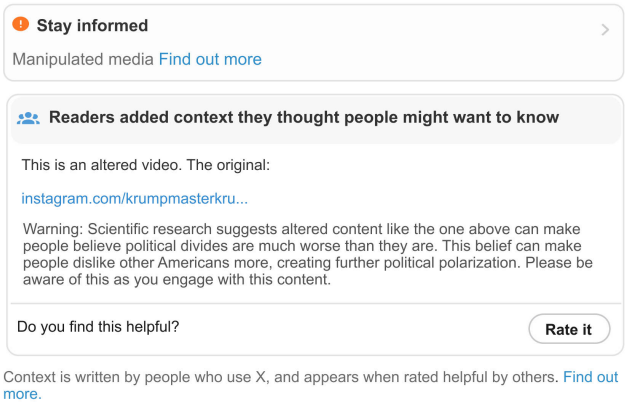


Figure 2: Community notes with political polarization context

5.3 Limitations

Several limitations pertain to our study. Although both our samples were balanced and representative, they nonetheless limit the generalization of the results as other samples of X users might yield varying effects. Another limitation of our study concerns the selection of topics and content across the eight evaluated posts (two in the election context and six in the non-election context), the specific type of community notes associated with each post, and the fact-checking responses provided. Although we balance between the initial and follow-up evaluations, we acknowledge that our findings may not generalize to non-political inauthentic content or to future elections contexts on *X* or other platforms. While we controlled for confounding factors such as removing engagement metrics and user comments, these elements are known to influence both perceived accuracy and sharing intentions. As such, participants' behavior in our study may diverge from how they might act in naturalistic settings on *X*. Additionally, the use of a 10-point scale to measure perceived accuracy and sharing intentions may carry the risk of attenuating the results by introducing unnecessary granularity or ambiguity.

We did not directly measure "friction," but instead inferred friction levels based on the required user interaction effort, following the framework proposed by [25]. A further limitation is participants' prior exposure to moderated content, which may have introduced a confounding effect, leading some to more readily judge content as inaccurate or to avoid sharing it altogether. Additionally, we did not evaluate soft moderation frictions in the context of repeated exposure, leaving open the possibility of habituation effects that could diminish intervention efficacy over time. Our findings remain specific to the controlled experimental conditions we employed, both within and outside the election context. For example, participants may have had prior familiarity with some of the inauthentic content we presented. Nevertheless, by employing a randomized control design, we were able to draw causal inferences regarding the impact of soft moderation interventions relative to the degree of interaction friction. All reported effects should, however, be interpreted as correlational rather than predictive. The same caveat applies to the thematic insights derived from our qualitative analysis.

6 Conclusion

Our study examined the impact of soft moderation interventions on the *perceived accuracy* and *sharing intentions* of inauthentic political content on *X*, both *during* and *outside* of election periods. From a user interaction perspective, we found that medium- and high-friction interventions led participants to view such content as less accurate. However, these interventions had no observable effect on sharing intentions, as many participants refrained from sharing inauthentic content altogether as a matter of principle. Our findings indicate that misleading content with political connotations can be effectively countered through community notes' consensus-based annotations developed by non-expert users from across the U.S. political spectrum.

7 Ethical Deliberation

Relative to the ***Respect for Persons*** principle of the Belmont Report [56], our protocol was non-full disclosure in that we *withheld* the information whether (1) each of the eight (six plus two) posts we evaluated was inauthentic or not; (2) and that these are real-world posts we sourced from X. This information, including the authentic counterpart for each content was tested, was offered to participants as part of the debriefing to mitigate a potential violation of the participants' autonomy. Here, we offered justification about the need to withhold the full nature and type of content they would see, as it was essential for us to obtain results under *realistic conditions*. Participants were able to remove their data while still being compensated if they felt the full disclosure and the mitigation were insufficient to address their loss of autonomy (none of them did). They also had the option to contact us and X with concerns about any inauthentic content encounters.

Relative to the ***Beneficence*** principle, we acknowledge that the *burden* of exposure to inauthentic content was needed to test the *unbiased* utility of the warnings towards assessing the posts' accuracy and the decision to share them or not. This burden might manifest as a feel shame, guilt, stress, embarrassment, or a sense of being manipulated when participants realize they have been not been fully informed about the nature of the content – inherently false, inauthentic – they saw, potentially impacting their self-esteem. If we had showed the participants only the warnings or used a laboratory generated inauthentic content, then their attitudes or behaviors would have been *biased* as they won't represent the real exposure to inauthentic content around the general discourse on political topics on X during both evaluation studies, and in turn, would not represent the real effect on perceptions and intentions.

Here, it is worth mentioning that we didn't ask participants to actually access any inauthentic content from their *own* X accounts in order to avoid the knock-on effect (i.e., getting recommendations for subsequent, related content they have no interest in) to minimize the *burden*. The unbiased utility of the warnings, as evidenced in both of our evaluations, is a *benefit* that is available for X (and its entire user base), and from March 2025, Meta too [44]. In times when fact-checkers are seen as unfavorable approach of handling misinformation [30], the *benefit* of the study is the evidence of what users – independently from the platforms' governance – prefer as a warning about content inauthenticity (i.e., community notes)

Relative to the ***Justice*** principle, beside the IRB approval, we offer evidence that the *benefits* of the soft moderation interventions – avoid deception by an inauthentic content – are equally enjoyed by the participants that bore the *burdens* as they regularly encounter such content on X and on the Meta during a usual platforms access. Our sampling criteria strictly stipulated anonymous participation, and we did not recruit by any other demographic criterion that might have resulted in social, racial, sexual, and cultural biases. A demonstration of this is our balanced sample relative to the intention to vote, gender, area of residence, and voting experience (Table 1).

Relative to ***Respect for Law and Public Interest*** principles from the Menlo Report [9], we borrowed a real-world content under the fair user policy of X content [65] (***compliance***). We transparently communicated the research purpose and situated our study in the broader context of harms associated with non-moderated inauthentic content on X (as well as other platforms) – widespread deception, civic unrest, and public health disturbances. We made it clear that we need exposures to real-world inauthentic content to learn the users' preferences about moderation when facing such recommendations in their X's feeds (***transparency***). We will use our results to advocate for the moderation of inauthentic content, as the detection of such content, so far, has been difficult for users. We also offer our artifacts, datasets, methodology, and all research instruments for the purpose of replication, but also future research, policy-making, or public knowledge (***accountability***).

Our *ethical position*, as authors, is that we see a greater benefit in studying soft moderation interventions to real-world content than the burden imposed in the study. If platforms and third-party fact-checking don't work, then our study shows that users on X and social media have at least the community notes at their disposal in navigating feeds that inevitably involve inauthentic, manipulated, or synthetically created content. The potential, brief psychological discomfort resulting from the exposure to the inauthentic content (and lasting 5 minutes on average till the participants reached the debriefing, where a full disclosure took place) is no greater than what a participant might experience in content encounters on X in their everyday life. We also see the exposure to inauthentic content (and soft moderation) as an opportunity for the participants to learn about what exists on X from their experience participating in a study using our methodology.

8 Open Science Policy

We are making both of our datasets as a research artifact *available* by default. We will offer the survey results and also provide the scripts we used to analyze the data to allow for both functionality checks and reproducibility. The study stimuli are already publicly available.

9 Acknowledgments

This work received funding from the Technology Research Center of the Bundeswehr (Voice of Wisdom). dtec.bw is funded by the European Union – NextGenerationEU.

References

[1] Aguerri, J. C., Santisteban, M., and Miró-Llinares, F. The fight against disinformation and

its consequences: measuring the impact of "Russia state-affiliated media"on Twitter. *Crime Science 13*, 1 (2024), 17.

[2] ALLEN, J., ARECHAR, A. A., PENNYCOOK, G., AND RAND, D. G. Scaling up fact-checking using the wisdom of crowds. *Science Advances 7*, 36 (2024), eabf4393.

[3] BACKHAUS, K., ERICHSON, B., GENSLER, S., WEIBER, R., WEIBER, T., ET AL. Multivariate analysis. *Springer Books 10* (2021), 978–3.

[4] BRAUN, V., AND CLARKE, V. *Thematic Analysis: A Practical Guide*. SAGE Publications, 2021.

[5] BRAUN, V., AND CLARKE, V. To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative research in sport, exercise and health 13*, 2 (2021), 201–216.

[6] BURRUS, O., CURTIS, A., AND HERMAN, L. Unmasking AI: Informing Authenticity Decisions by Labeling AI-Generated Content. *Interactions 31*, 4 (June 2024), 38–42.

[7] CHUAI, Y., TIAN, H., PRÖLLOCHS, N., AND LENZINI, G. Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction 8*, CSCW2 (Nov. 2024), 1–52.

[8] CYBERSECURITY & INFRASTUCTURE SECURITY AGENCY. Risk in Focus: Generative A.I. and the 2024 Election Cycle, 2024. https://www.cisa.gov/resources-tools/resources/risk-focus-generative-ai-and-2024-election-cycle.

[9] DEPARTMENT OF HOMELAND SECURITY. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research, 2011.

[10] DIEL, A., LALGI, T., SCHRÖTER, I. C., MACDORMAN, K. F., TEUFEL, M., AND BÄUERLE, A. Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports 16* (2024), 100538.

[11] DISTLER, V., LENZINI, G., LALLEMAND, C., AND KOENIG, V. The Framework of Security-Enhancing Friction: How UX Can Help Users Behave More Securely. In *Proceedings of the New Security Paradigms Workshop 2020* (New York, NY, USA, 2021), NSPW '20, Association for Computing Machinery, p. 45–58.

[12] DROLSBACH, C. P., AND PRÖLLOCHS, N. Diffusion of Community Fact-Checked Misinformation on Twitter. *Proc. ACM Hum.-Comput. Interact. 7*, CSCW2 (oct 2023).

[13] DROLSBACH, C. P., SOLOVEV, K., AND PRÖLLOCHS, N. Community notes increase trust in fact-checking on social media. *PNAS Nexus 3*, 7 (05 2024), pgae217.

[14] DUFFY, A., TANDOC, E., AND LING, R. Too good to be true, too good not to share: the social utility of fake news. *Information, Communication & Society 23*, 13 (11 2020), 1965–1979.

[15] EADY, G., PASKHALIS, T., ZILINSKY, J., BONNEAU, R., NAGLER, J., AND TUCKER, J. A. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications 14*, 1 (2023), 62.

[16] ECKER, U. K. H., LEWANDOWSKY, S., COOK, J., SCHMID, P., FAZIO, L. K., BRASHIER, N., KENDEOU, P., VRAGA, E. K., AND AMAZEEN, M. A. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology 1*, 1 (2022), 13–29.

[17] EFSTRATIOU, A., AND DE CRISTOFARO, E. Adherence to misinformation on social media through socio-cognitive and group-based processes. *Proc. ACM Hum.-Comput. Interact. 6*, CSCW2 (Nov. 2022).

[18] EPSTEIN, Z., ARECHAR, A. A., AND RAND, D. What label should be applied to content produced by generative AI? https://osf.io/preprints/psyarxiv/v4mfz_v1.

[19] FERRARA, E., CHANG, H., CHEN, E., MURIC, G., AND PATEL, J. Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday 25*, 11 (2024/12/17 2020).

[20] FOURNEY, A., RACZ, M. Z., RANADE, G., MOBIUS, M., AND HORVITZ, E. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (New York, NY, USA, 2017), CIKM '17, Association for Computing Machinery, p. 2071–2074.

[21] FREEDMAN, R. Sweeping EU digital misinformation law takes effect, Aug. 2023. https://www.legaldive.com/news/digital-services-act-dsa-eu-misinformation-law-propaganda-compliance-facebook-gdpr/691657/.

[22] GALGANO, P. ABC News segment altered to inflate Ukrainian soldier deaths, 2025. https://factcheck.afp.com/doc.afp.com.36QG8VU.

[23] GALGANO, P. AI-generated images falsely claimed to show bombing near Beirut airport, 2025. https://factcheck.afp.com/doc.afp.com.36KL83M.

[24] GEENG, C., YEE, S., AND ROESNER, F. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2020), CHI '20, Association for Computing Machinery, p. 1–14.

[25] GOULD, S. J. J., CHUANG, L. L., IACOVIDES, I., GARAIALDE, D., CECCHINATO, M. E., COWAN, B. R., AND COX, A. L. A Special Interest Group on Designed and Engineered Friction in Interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2021), CHI EA '21, Association for Computing Machinery.

[26] GUESS, A., NAGLER, J., AND TUCKER, J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances 5*, 1 (2019), 4586.

[27] HARTWIG, K., DOELL, F., AND REUTER, C. The Landscape of User-centered Misinformation Interventions - A Systematic Literature Review. *ACM Comput. Surv. 56*, 11 (jul 2024).

[28] HENNESSY, M., DELLI CARPINI, M. X., BLANK, M. B., WINNEG, K., AND JAMIESON, K. H. Using psychological theory to predict voting intentions. *Journal of Community Psychology 43*, 4 (2024/12/18 2015), 466–483.

[29] HORTA RIBEIRO, M., CHENG, J., AND WEST, R. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM Web Conference 2023* (New York, NY, USA, 2023), WWW '23, Association for Computing Machinery, p. 2666–2676.

[30] ISAAC, M., AND SCHLEIFER, T. Meta says it will end its fact-checking program on social media posts, 2025. https://www.nytimes.com/live/2025/01/07/business/meta-fact-checking.

[31] JAHANBAKHSH, F., ZHANG, A. X., AND KARGER, D. R. Leveraging structured trusted-peer assessments to combat misinformation. *Proc. ACM Hum.-Comput. Interact. 6*, CSCW2 (nov 2022).

[32] JENKE, L. Affective polarization and misinformation belief. *Political Behavior 46*, 2 (2024), 825–884.

[33] JIA, C., BOLTZ, A., ZHANG, A., CHEN, A., AND LEE, M. K. Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation. *Proc. ACM Hum.-Comput. Interact. 6*, CSCW2 (nov 2022).

[34] KAISER, B., WEI, J., LUCHERINI, E., LEE, K., MATIAS, J. N., AND MAYER, J. Adapting security warnings to counter online disinformation. In *30th USENIX Security Symposium (USENIX Security 21)* (Aug. 2021), USENIX Association, pp. 1163–1180.

[35] KUBIN, E., AND VON SIKORSKI, C. The polarizing content warning: how the media can reduce affective polarization. *Human Communication Research 50*, 3 (05 2024), 404–418.

[36] LAYTON, S., TUCKER, T., OLSZEWSKI, D., WARREN, K., BUTLER, K., AND TRAYNOR, P. SoK: The good, the bad, and the unbalanced: Measuring structural limitations of deepfake media datasets. In *33rd USENIX Security Symposium (USENIX Security 24)* (Philadelphia, PA, Aug. 2024), USENIX Association, pp. 1027–1044.

[37] LEWIS, A., VU, P., DUCH, R. M., AND CHOWDHURY, A. Do content warnings help people spot a deepfake? evidence from two experiments. *OSF Preprint* (2022).

[38] MARTEL, C., AND RAND, D. G. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology 54* (2023), 101710.

[39] MCCARTHY, B. Doctored RNC video fuels falsehoods about Trump assassination attempt, 2024. https://factcheck.afp.com/doc.afp.com.36786WG.

[40] MCCARTHY, B. Kamala Harris did not pose with Sean 'Diddy' Combs in 2001 photo, 2024. https://factcheck.afp.com/doc.afp.com.36GH38V.

[41] MCCARTHY, B. Photo of Tim Walz altered to add racist object, 2024. https://factcheck.afp.com/doc.afp.com.36RL2UB.

[42] MCCARTHY, B. Illustration of Trump banners outside Kennedy Center mistaken as real, 2025. https://factcheck.afp.com/doc.afp.com.372D28N.

[43] META. How we're tackling misinformation across our apps, 2022. https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/.

[44] META. Community Standards, 2024. https://transparency.meta.com/policies/community-standards/.

[45] META. Inauthentic Behavior, 2025. https://transparency.meta.com/policies/community-standards/inauthentic-behavior/.

[46] OELDORF-HIRSCH, A., SCHMIERBACH, M., APPELMAN, A., AND BOYLE, M. P. The ineffectiveness of

fact-checking labels on news memes and articles. *Mass Communication and Society 23*, 5 (09 2020), 682–704.

[47] PAPAKYRIAKOPOULOS, O., AND GOODMAN, E. The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets. In *Proceedings of the ACM Web Conference 2022* (2022), pp. 2541–2551.

[48] PATRICK, D. Analysis shows 'leaked' Vance rant about Musk is likely fake, 2025. https://factcheck.afp.com/doc.afp.com.38AT862.

[49] ROOZENBEEK, J., MAERTENS, R., HERZOG, S. M., GEERS, M., KURVERS, R., SULTAN, M., AND VAN DER LINDEN, S. Susceptibility to misinformation is consistent across questionframings and response modes and better explained by myside bias and partisanshipthan analytical thinking. *Judgment and Decision Making 17*, 3 (2022), 547–573.

[50] SCHMIDT, A. F., AND FINAN, C. Linear regression and the normality assumption. *Journal of Clinical Epidemiology 98* (June 2018), 146–151.

[51] SCHNEIER, B., AND SANDERS, N. The apocalypse that wasn't: AI was everywhere in 2024's elections, but deepfakes and misinformation were only part of the picture, 2024. https://theconversation.com/the-apocalypse-that-wasnt-ai-was-everywhere-in-2024s-elections-but-deepfakes-and-misinformation-were-only-part-of-the-picture-244225.

[52] SELECT COMMITEE ON INTELLEGENCE, US SENATE. Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, 2020. https://www.intelligence.senate.gov/publications/report-select-committee-intelligence-united-states-senate-russian-active-measures.

[53] SHAREVSKI, F., HUFF, A., JACHIM, P., AND PIERONI, E. (Mis)perceptions and engagement on Twitter: COVID-19 vaccine rumors on efficacy and mass immunization effort. *International Journal of Information Management Data Insights 2*, 1 (2022), 100059.

[54] SWIRE, B., BERINSKY, A. J., LEWANDOWSKY, S., AND ECKER, U. K. Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society open science 4*, 3 (2017), 160802.

[55] TERNOVSKI, J., KALLA, J., AND ARONOW, P. M. Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments. *OSF Preprints 10* (2021).

[56] THE BELMONT REPORT. Ethical Principles and Guidelines for the Protection of Human Subjects of Research, 1979.

[57] TIKTOK. TikTok Safety, 2023. https://www.tiktok.com/safety/en-us/topics/.

[58] TWITTER. How we address misinformation on Twitter, 2022. https://help.twitter.com/en/resources/addressing-misleading-info.

[59] TWITTER. Community Notes, 2024. https://communitynotes.x.com/guide/en/about/introduction.

[60] VAIDYA, T., VOTIPKA, D., MAZUREK, M. L., AND SHERR, M. Does Being Verified Make You More Credible? Account Verification's Effect on Tweet Credibility. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI '19, Association for Computing Machinery, p. 1–13.

[61] WADE, N. Posts altered to misquote Harris spokesperson on gun ownership, 2024. https://factcheck.afp.com/doc.afp.com.36GY48B.

[62] WIRED MAGAZINE. The WIRED AI Elections Project, 2024. https://www.wired.com/story/generative-ai-global-elections/.

[63] WITTENBERG, C., EPSTEIN, Z., BERINSKY, A. J., AND RAND, D. G. Labeling AI-Generated Content: Promises, Perils, and Future Directions. *An MIT Exploration of Generative AI* (mar 27 2024). https://mit-genai.pubpub.org/pub/hu71se89.

[64] X. Authenticity, 2025. https://help.x.com/en/rules-and-policies/authenticity.

[65] X. Fair Use Policy, 2025. https://help.x.com/en/rules-and-policies/fair-use-policy.

[66] YOUTUBE. Misinformation policies, 2023. https://support.google.com/youtube/answer/10834785?hl=en.

Appendix

A Survey

- US Elections 2024: **8 groups**
- Non Election: **24 groups**

B Replication Files

The Replication Files including our dataset and code used in our quantitative analysis.

C Codebook

The Codebook used in our thematic analysis.

D Debriefing

The Debriefing used in both studies.

E Analysis Tables

E.1 US Elections 2024

- Predictor: Intention to Vote [US Elections 2024]
- Predictor: Demographics [US Elections 2024]

E.2 Non Election

- Predictor: Intention to Vote [Non Election]
- Predictor: Demographics [Non Election]

F Study Stimuli

F.1 US Elections 2024 – Kamala Harris

Figure 4: Inauthentic Photo and Soft Moderation

- Control (No Warnings)
- Warning Label
- Warning Bundle (Label + Community Note)
- Warning Cover
- Warning Cover Fact Check

F.2 US Elections 2024 – Donald Trump

Figure 5 Inauthentic Video and Soft Moderation

- Control (No Warnings)
- Warning Label
- Warning Bundle (Label + Community Note)
- Warning Cover
- Warning Cover Fact Check

F.3 Non Election – JD Vance

Figure 6 Inauthentic Audio and Soft Moderation :

- Control (No Warnings)
- Warning Label
- Warning Bundle (Label + Community Note)
- Warning Cover
- Warning Cover Fact Check

F.4 Non Election – Donald Trump

Figure 7 Inauthentic Photo and Soft Moderation :

- Control (No Warnings)
- Warning Label
- Warning Bundle (Label + Community Note)
- Warning Cover
- Warning Cover Fact Check

F.5 Non Election – Tim Walz

Figure 8 Inauthentic Photo and Soft Moderation :

- Control (No Warnings)
- Warning Label
- Warning Bundle (Label + Community Note)
- Warning Cover
- Warning Cover Fact Check

F.6 Non Election – Kamala Harris

Figure 9 Inauthentic Video and Soft Moderation :

- Control (No Warnings)
- Warning Label
- Warning Bundle (Label + Community Note)
- Warning Cover
- Warning Cover Fact Check

F.7 Non Election – Russo/Ukrainian

Figure 10 Inauthentic Audio and Soft Moderation

- Control (No Warnings)
- Warning Label
- Warning Bundle (Label + Community Note)
- Warning Cover
- Warning Cover Fact Check

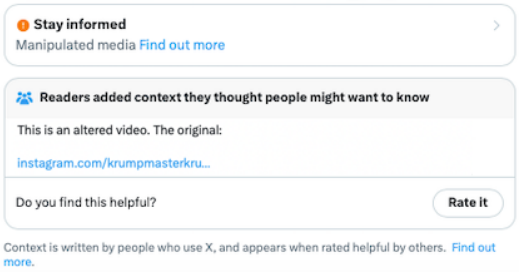
F.8 Non Election – Israel/Palestine

Figure 11 Inauthentic Audio and Soft Moderation

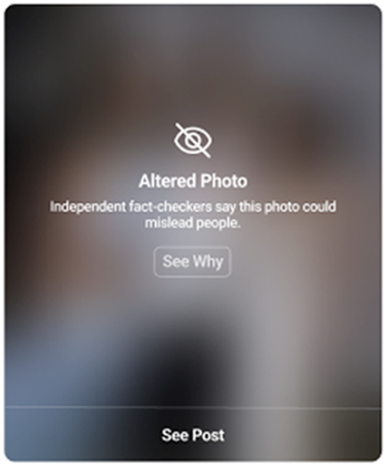
- Control (No Warnings)
- Warning Label
- Warning Bundle (Label + Community Note)
- Warning Cover
- Warning Cover Fact Check

G Soft Moderation Interventions

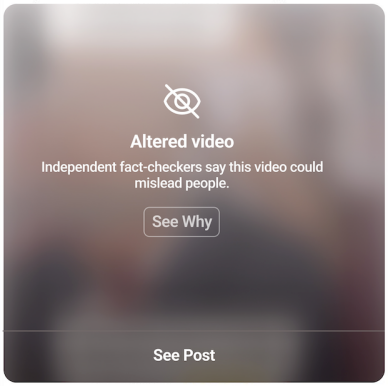
(a) Inauthentic Warning Label



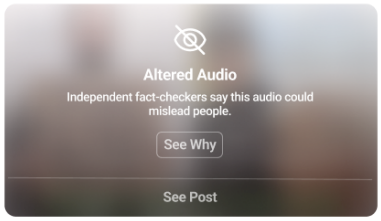
(b) Warning Bundle: Label & Community Note



(c) Warning Cover – Photo



(d) Warning Cover – Video



(e) Inauthentic Audio Cover

Figure 3: Soft Moderation Interventions for Inauthentic Media