

USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

Source-Level Disengagement: A Usable Security Defense Against Misinformation

Zaid Hakami, *Florida International University and Jazan University;*
Yuzhou Feng and Bogdan Carbunar, *Florida International University*

https://www.usenix.org/conference/soups2026/presentation/hakami

This paper is included in the Proceedings of the
Twenty-Second Symposium on Usable Privacy and Security.

August 24–26, 2026 • Hannover, Germany

ISBN 978-1-939133-59-5

Open access to the Proceedings of the
Twenty-Second Symposium on Usable Privacy and Security
is sponsored by USENIX.

Source-Level Disengagement: A Usable Security Defense Against Misinformation

Zaid Hakami[†‡] *Yuzhou Feng*[†] *Bogdan Carbunar*[†]
[†]*Florida International University,* {zhaka001, yfeng015, carbunar}@fiu.edu
[‡]*Jazan University,* zhhakami@jazanu.edu.sa

Abstract

Misinformation defenses on social media mostly focus on individual posts, and do not address repeated exposure from the same misinformation-sharing accounts. We frame source-level disengagement as a usable security defense and study how interface design shapes adoption and escalation of such controls. We built TimeLineR, an instrumented timeline environment, and used it to run a within-subject lab study with 29 participants. We design and compare three interface designs that vary the visibility and timing of source-level defensive actions, and introduce a Filter action that hides misinformation-labeled posts from an account while preserving other content. We find that actionability increases defensive uptake: in-context controls substantially increased adoption in our controlled setting. Users select relationship-appropriate defenses, with the introduced Filter acting as a stable middle-ground endpoint.

1 Introduction

Misinformation on social media is amplified through platform-mediated diffusion and repeated exposure, driven in part by engagement dynamics [5, 38]. Many deployed and studied countermeasures operate at the level of individual posts [21], e.g., warning labels, fact-checks, or contextual notices. Post-level interventions can reduce belief or sharing in some settings [8, 24]. However, they do not address a central driver of exposure: continued connections to sources that repeatedly produce or amplify misleading content.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
USENIX Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23–26, 2026, Hannover, Germany.

As a result, even after recognizing a source as misleading, users may continue to encounter its posts. This exposure risk is relevant to usable security: repeated encounters create repeated opportunities for downstream harmful actions (e.g., re-sharing or acting on false operational information). Platforms do provide source-level controls such as mute, unfollow, and block. While in principle they can reduce future exposure, prior work suggests these controls are often used sparingly or avoided, especially for socially meaningful sources (e.g., acquaintances or community accounts) [4]. More broadly, users may struggle to discover, navigate, or confidently use feed control settings in routine use [18]. Thus, the defense goal is to reduce future exposure to repeat misinformation sources through usable, user-mediated account-level controls.

This paper builds on prior work that treats misinformation interventions as security defenses shaped by user behavior [20] to frame *source-level disengagement* as a *usable security* problem. In this framing, repeated exposure is the risk, user-mediated source-level controls are the mitigation, and usability-driven non-adoption constitutes a defense failure mode. We therefore study how interface-level design shapes adoption, escalation, and abandonment in practice.

We introduce and study source-level disengagement interventions that operate at the account level, including familiar controls (mute, unfollow, block, report) and a novel *filtering* mechanism that hides posts identified as misinformation from a given source while preserving other content from that account. Filtering is designed as a middle-ground source-level defense: it reduces exposure to repeated misinformation without severing social ties or requiring irreversible action. It sits between taking no action and stronger relationship-severing controls such as unfollowing or blocking. This design targets barriers that can limit adoption of stronger source-level controls. Guided by this, we ask the following research questions:

- **RQ1 (Defensive uptake)**. Under what interface conditions do users take source-level defensive actions against misinformation-sharing accounts?

- **RQ2 (Intervention design)**. How do interface-level design choices for source-level disengagement influence defensive uptake, escalation, and breakdowns in use?

To answer these questions, we developed *TimeLineR*, a controlled, instrumented timeline environment that reproduces core social-media interactions and enables measurement of defensive behavior. Using TimeLineR, participants browsed realistic timelines containing a mix of genuine posts and ethically scoped misinformation. We operationalize interface conditions as three intervention designs that vary the visibility, timing, and friction of source-level actions.

Behavioral interaction logs provide the primary evidence of defensive effectiveness, measured as uptake, follow-through, and control-choice trajectories under repeated exposure. Semi-structured interviews provide supporting context to explain observed behaviors and reveal constraints shaping adoption.

The study shows that source-level disengagement depends primarily on actionability at the moment of encounter. Warning-only designs reduce engagement and increase explanation-seeking, but rarely produce follow-through. In contrast, surfacing source-level controls inline substantially increases defensive uptake. In addition, users do not escalate along a severity ladder: filtering functions as a stable middle-ground endpoint and is not a precursor to unfollowing or blocking. We further find that justification use changes across repeated exposure, and that a subset of users remains non-adopting even when controls are prominent and actionable. These findings motivate a usable-security framing of source-level disengagement and inform the design of encounter-time controls. Accordingly, we make four contributions:

- Conceptual: We model source-level disengagement as a usable security defense against repeated exposure to misinformation sources.
- Design and empirical finding: We introduce and evaluate *Filter*, a middle-ground source-level defense that reduces exposure to misinformation-labeled posts without severing ties. We show that filtered sources were not subsequently unfollowed or blocked within the study.
- Empirical: We characterize how encounter-time actionability drives uptake and follow-through.
- Design constraints: We identify practical limits for source-level defenses, including relationship cost, uneven control selection, declining explanation use, and persistent non-adoption.

2 Related Work

Post-level Misinformation Interventions. A substantial body of prior work focuses on identifying and flagging misinformation at the level of individual posts. These defenses include automated detection [16, 17, 28, 39], third-party fact-checking [1, 2, 34], crowdsourced annotations such as Community Notes, warning labels, fact-check notices, and contextual prompts. Prior work shows that such interventions can reduce belief in or sharing of misinformation under certain conditions [6, 10, 25]. However, post-level interventions are tied to specific flagged items and may not reduce repeated exposure from the same source. Work on account-level credibility cues shows that users attend to source signals and platform policy, but do not provide actionable mechanisms for reducing repeated exposure from specific sources [30]. Our work assumes the availability of post-level identification signals and examines a complementary question: how users respond when misinformation is produced by identifiable sources.

Source-level Controls and Disengagement. Most social-media platforms provide *source-level* controls, e.g., muting, unfollowing, blocking, and reporting, that can reduce repeated exposure to problematic accounts. These controls can function as *user-mediated defenses*: once a source is identified as sharing misinformation, disengaging from that source can mitigate future risk without requiring post-level judgment.

User coping responses to misinformation are shaped by platform context and social ties [13, 33]. Source-level controls are used sparingly, even when users recognize repeat misinformation sources [4]. Stronger actions are often perceived as socially costly or overly blunt, particularly for socially or institutionally salient accounts [4]. Engagement with misleading content is often driven by affective or motivational factors and not accuracy judgments alone [7]. These findings suggest limits to awareness-based interventions.

What remains underexplored is whether *intervention design* can make source-level defenses more adoptable in everyday use. Existing controls are typically surfaced as generic account actions and are rarely framed, sequenced, or explained as part of a defensive response. As a result, we lack behavioral evidence about when users adopt source-level controls, how they use and combine available actions over time, and where usability-related breakdowns occur as defensive cues compete with routine scrolling. Addressing this gap, we examine how interface-level choices for surfacing and sequencing source-level actions, including a middle-ground *Filter* option, shape defensive uptake and non-adoption under exposure.

Usable Security Design Constraints. A large design space of misinformation interventions has been explored in prior work, including warning labels, fact-checking labels, source-credibility cues, and friction-based designs [21]. Comparative studies show that placement and specificity of warnings influence whether users notice and interpret them correctly. For example, interstitial warnings are often more effective at reducing engagement than embedded labels, which can be overlooked during routine browsing [20]. Other work shows that poorly placed or overly generic warnings may be ignored or misunderstood, while precise explanatory cues can increase perceived credibility [14, 15, 22].

Thus, defenses must not only exist, but must be actionable and comprehensible at the moment of decision [35, 37]. However, most prior work evaluates post-level signaling effects

on belief or sharing, and not whether users adopt source-level controls for accounts posting misleading content. As a result, we know comparatively little about how intervention design can support sustained source-level disengagement under misinformation exposure.

Attention Capture and Habituation. Feed-based interfaces promote rapid, habitual consumption, which can reduce follow-through even when users notice risk cues. Prior work shows that salience and interruption can shift attention and behavior, but effects can attenuate under repeated exposure due to habituation and fatigue [11, 23, 26, 27, 31]. For misinformation, post-level warnings must compete with routine scrolling and can become background over time [11, 20, 24, 31]. Usable-security research shows that protective behavior depends on decision-time costs and fit with routine workflows. This motivates security-enhancing friction [9].

We build on these insights by comparing designs that vary cue timing (inline vs. gated) and action accessibility (menu vs. in-context) under repeated exposure.

3 Scope and Assumptions

Threat Model. We distinguish between *misinformation*, defined as false or misleading information shared without verified intent to deceive, and *disinformation*, defined as false information deliberately produced or disseminated with intent to mislead. Intent is generally unobservable in platform-scale settings. Thus, in our experimental design we use the term *misinformation* to refer to content treated as false by the system, without making claims about source intent.

We consider misinformation exposure on social media platforms as a repeated-exposure risk arising from continued connections to sources that share misinformation over time. In our model, the threat actor is a source account that repeatedly produces posts labeled as misinformation. We do not assume the source has a specific intent or strategic objective, nor that it controls the platform or client interface. The source can publish content that may be algorithmically amplified or encountered through existing social connections, resulting in recurring exposure for connected users. We also do not model adversarial adaptation to specific interface designs (e.g., content mutation or account switching). Instead, the source behavior is considered fixed during interaction.

System and Trust Assumptions. Our work focuses on the last mile of defense, where platform users must decide whether and how to disengage from misinformation sources using available account-level controls. We assume that some upstream mechanism, such as automated detection, third-party fact-checking, or platform moderation, identifies specific posts as misinformation and surfaces that signal to the user. The accuracy, fairness, and robustness of detection are outside the study scope.

We assume a fixed set of misinformation-labeled posts and stable source identities during interaction. We do not model coordinated campaigns or cross-platform propagation.

Defensive Goal and Scope. Our goal is not to prevent the creation or platform-wide spread of misinformation, but to reduce an individual user's repeated exposure to misinformation from known sources via source-level disengagement. We frame this as a *usable-security* problem: source-level controls are exposure-reducing defenses whose protection is realized only if users can reliably enact them during routine feed use. Accordingly, we evaluate defensive effectiveness in behavioral terms: whether users adopt, follow through on, or abandon source-level controls when they are surfaced in context. We do not measure belief change, accuracy judgments, or downstream platform-wide diffusion.

4 Intervention Designs and TimeLineR

This section describes the three intervention designs and the *TimeLineR* platform used to render them and log interactions.

4.1 Intervention Designs

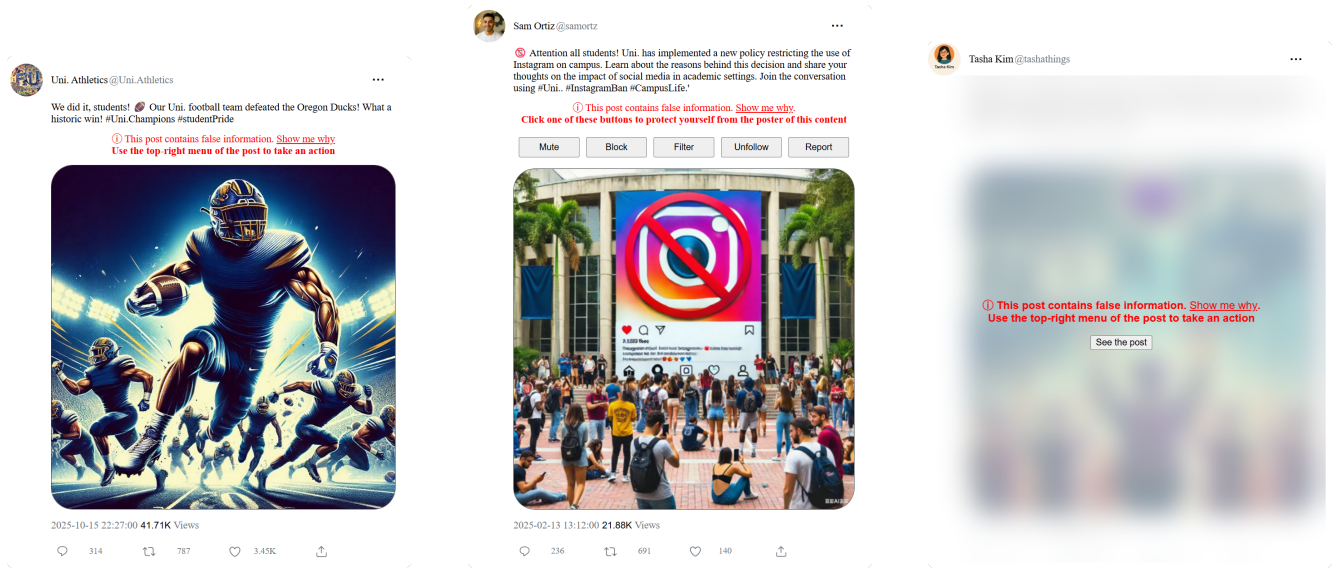
We evaluated three intervention designs that support source-level disengagement but differ in how defensive cues and actions are presented, see Figure 1. In all conditions, participants were exposed to the same source-level actions: Mute, Unfollow, Block, Report, and a middle-ground Filter option. What varied across designs was the visibility, timing, and interruptiveness of source-level controls.

Inline warning (Intervention 1). Posts identified as misinformation were accompanied by a warning label embedded directly above the post content, see Figure 1(a). The warning initially appeared in low-contrast gray text. If a post remained visible on screen for three seconds, the warning increased in visual salience (shifting to red text) to draw attention without interrupting scrolling. Source-level actions were accessible through the post's native menu, not explicitly foregrounded.

Action-oriented source-level controls (Intervention 2). This condition paired the inline warning with prominently displayed, inline action buttons beneath the post, see Figure 1(b). These buttons provided immediate access to source-level actions, including the Filter option. As in Intervention 1, the warning label employed delayed visual emphasis, to increase salience only after several seconds of dwell time.

Interstitial blur overlay (Intervention 3). Misinformation posts were obscured by a blur overlay before the content was visible (Figure 1(c)). The overlay informed participants that the source had shared false information and required an explicit action (e.g., "See post") to reveal the content. Source-level defensive actions were available only after the post was revealed. Unlike Interventions 1 and 2, this design gated access to the post instead of presenting a delayed warning.

Filter action. In addition to traditional source-level controls, we introduced a *Filter* option as a middle-ground source-level defense, see Figure 1(b). When a user filters a source,



(a) Inline Warning (b) Action Nudges (c) Blurred Overlay

Figure 1: Three misinformation intervention designs used in the study. (a) Inline warning above flagged posts, with source-level actions hidden in the post menu. (b) Action nudges with immediately accessible source-level controls (Mute, Block, Filter, Unfollow, Report). (c) Blurred overlay concealing the post until the user opts to reveal, with source-level actions available only after reveal. University logos were removed and the university name was replaced with "Uni" in the post texts. Poster names and profile images were AI-generated.

TimeLineR hides subsequent posts from that source only when they are labeled as misinformation, while leaving that source's unlabeled posts visible. Thus, Filter reduces exposure to misinformation-labeled content from a source without muting, unfollowing, or blocking the account. Filter is *source-conditioned* and *label-conditioned*: it suppresses posts from a specific account only when they are labeled as misinformation, and does not affect unlabeled posts from the source. Unlike keyword mutes or ranking-based "see less" controls, Filter is gated by an explicit misinformation label and scoped to a specific source.

Filter was available in all intervention conditions but was visually emphasized only in Intervention 2 and 3. During the study, participants were not pre-trained on Filter. Instead, they encountered it as an interface control, with hover text available to describe its function.

4.2 TimeLineR System

Platform APIs provide limited visibility into user interaction sequences. Experimental control over repeated exposure is also infeasible, and deploying misinformation interventions raises ethical concerns. To address these constraints, we implemented *TimeLineR*, a controlled timeline environment for evaluating how users interact with source-level defenses under repeated exposure to misinformation. Similar motivations have led prior work to evaluate security interactions in virtualized study environments [19].

RQ1 and RQ2 focus on whether users execute source-level defenses at the moment of exposure. Thus, the primary methodological requirement is precise control over content exposure, action availability, and interaction logging, and not a live social graph or real-world consequences. We use "timeline-based interface" to mean a feed-style, vertically scrolling environment in which posts from identifiable sources appear in sequence and users can interact with posts and source-level controls. Therefore, TimeLineR reproduces selected core interaction patterns of mainstream social-media timelines, including scrolling, post engagement, identifiable sources, and source-level controls. It is not intended to reproduce platform-specific ranking, mobile affordances, or the full interaction ecology of any single platform. TimeLineR presents participants with a vertically scrolling feed (see Figure 4 in Appendix C) that includes identifiable source accounts, standard engagement affordances (Reply, Like, Share, Repost), and commonly available source-level moderation actions (Mute, Block, Report), alongside the *Filter* action introduced in this study.

This design allows participants to encounter content from multiple sources across a timeline and to take account-level actions in realistic contexts, while maintaining experimental control over content, intervention timing, and action availability. To reduce distraction and isolate source-level defensive behavior, we intentionally omitted auxiliary features such as recommendation tuning controls and platform-specific secondary menus.

Figure 3 in Appendix C shows the high-level architecture of TimeLineR. Genuine and synthetic misinformation posts (Section 5.5), including LLM-generated content, are stored in a PostgreSQL posts dataset and served via a Flask backend to the TimeLineR front-end. This renders timelines with standard social-media interactions and the intervention designs described in Section 4.1. Intervention logic is injected at render time. This allows precise control over when and how warnings, action affordances, and interstitial overlays are presented, while keeping content and interface design decoupled across conditions.

All participant interactions are logged. This includes scrolling behavior, engagement actions, invocation of source-level controls, and the ordering of interactions (e.g., revealing blurred content before or after disengagement). Interaction events are captured via Clarity [3], streamed in real time to the moderator to support post-timeline interviews, and stored for offline analysis. This instrumentation enables reconstruction of participant-level defensive trajectories.

5 Method

5.1 Study Overview

We conducted a controlled, lab-based user study using TimeLineR. TimeLineR preserves repeated exposure, persistent sources, and executable account controls. We hold content and exposure fixed and vary only cue timing and visibility and action accessibility to isolate interface presentation and target comparative adoption mechanisms. Behavioral interaction logs serve as the primary evidence for defensive effectiveness. Post-task semi-structured interviews provide supporting context for interpreting observed behaviors and identifying usability constraints.

TimeLineR is therefore used to compare interface-mediated action mechanisms under controlled exposure, not to estimate adoption rates in participants' actual social-media feeds.

5.2 Participants

We recruited $N = 29$ participants from a university community through mailing lists and campus advertisements. A brief screening survey confirmed regular use of at least one timeline-based social-media platform and familiarity with common feed interactions (e.g., liking and sharing). Participants ranged in age from 18 to 46 ($M = 24$, $SD = 7$) and reported regular use of at least one timeline-based social-media platform (e.g., Instagram, Facebook, X). Participants were drawn from a mix of student, staff, and faculty roles within the university community. All participants provided informed consent and were compensated $30 for their time.

This population was selected to ensure familiarity with timeline-based feeds and with identifiable content sources. The study was intended to evaluate source-level defensive mechanisms in a controlled setting, not to estimate population-level behavior. We treat $N = 29$ as appropriate for a controlled within-subject usability study focused on mechanisms, not population prevalence. All participants completed all four conditions, enabling participant-level paired comparisons. Repeated post exposures support observation of action trajectories, but are not treated as independent participants.

This sample size is consistent with prior usable-security studies using similar participant counts for controlled or formative evaluations. Examples include a SOUPS study with a within-subjects 2FA setup usability component using $N = 30$ participants [32], a SOUPS evaluation of e-mail user security tools with $N = 27$ participants [40], and a security-usability within-subjects lab study with $N = 31$ participants [36].

5.3 Experimental Design

We employed a within-subjects design in which participants interacted with four timelines: a control timeline without intervention and three timelines that differed in how misinformation cues and source-level actions were presented. The specific intervention designs are described in Section 4.1. The three intervention timelines were counterbalanced using a Latin square to mitigate learning and carryover effects. The control timeline was presented first to establish a baseline prior to exposure to intervention cues.

Each timeline contained 12 posts: 9 genuine posts and 3 misinformation posts (Appendix B). Misinformation stimuli were fixed within each timeline and constructed to be similar in plausibility, and stakes. Misinformation was kept a minority of content to avoid overexposure to warnings.

5.4 Procedure

Participants completed the study individually in a controlled laboratory setting. After providing informed consent, participants received instructions that they would browse a simulated social media feed and could interact with posts as they normally would. Participants then browsed four timelines in sequence using the TimeLineR platform (Section 4.2). Participants interacted with TimeLineR through a desktop web browser using mouse-based scrolling and click-based controls. They were not given any specific task beyond using the feed naturally. During browsing, they were free to scroll, open posts, engage with content, and use any available account-level controls. Throughout the session, participants followed a concurrent think-aloud protocol, verbalizing their reasoning and reactions when encountering posts or interface cues.

After completing all timelines, participants took part in a semi-structured interview. The interview focused on how they interpreted the interventions, how they decided whether to take action against sources, and why they adopted, escalated, or avoided source-level defensive actions. Observed interaction events were referenced during the interview to

ground discussion in participant behavior. Table 3 lists the probe categories.

Finally, participants were debriefed about the presence of synthetic misinformation content and the purpose of the study.

Study sessions lasted between 57 and 95 minutes ($M = 75.6$, $SD = 9.8$), including consent, browsing the four timelines, the post-task interview, and debriefing.

5.5 Study Stimuli

Synthetic misinformation posts. To evaluate source-level disengagement under controlled and ethical conditions, we used *synthetic, low-stakes misinformation posts* tailored to a university context. These posts were designed to be realistic enough to prompt defensive decision-making, while avoiding sensitive domains such as politics, health, or real-world crises, see Table 2 in Appendix B for the list.

Misinformation posts were generated using a semi-automated pipeline. The generation prompts asked for plausible but false campus-related social media announcements in low-stakes domains, such as transportation access, campus amenities, or student events. Candidate posts were refined to resemble social-media posts and reviewed by the research team. We selected posts that were false, understandable without specialized knowledge, relevant to the participant population, and unlikely to cause foreseeable harm. We considered alternatives such as political, health, crisis, and identity-related misinformation, but excluded them to avoid stronger emotional reactions and ethical risks outside the scope of this controlled interface study. For each misinformation item, a debunking explanation linked to an authoritative university source was prepared and used during debriefing.

Genuine posts. Genuine posts were drawn from publicly available content published by official university-affiliated accounts (e.g., the university, departments, and student organizations) and focused on routine campus announcements, events, and achievements.

Source accounts. To support source-level disengagement decisions, all posts (both genuine and misinformation) were associated with persistent source accounts that could appear multiple times within a timeline. Synthetic misinformation posts were assigned to realistic but fictitious source accounts designed to resemble common campus-affiliated or peer accounts. This ensured that participants could reason about disengagement without involving real individuals.

The study population was intentionally drawn from a university community to ensure that participants were familiar with the types of sources and topics represented in the timelines. Participants regularly encounter content from institutional accounts, student organizations, and peer-like sources in their everyday social-media use. As a result, they were well positioned to interpret the plausibility of the synthetic posts and to reason about the perceived cost of disengaging from different source types. We therefore focus on interaction and control-use mechanisms instead of domain-specific belief contestation. Higher-stakes domains (politics, health) are a target for future replication, not a requirement for isolating interface-mediated defensive behavior. This alignment between stimuli and participant context supports ecological realism for evaluating source-level disengagement decisions without exposing participants to real or consequential misinformation.

5.6 Measures

Measures are computed directly from interaction logs. *Defensive uptake* indicates whether a participant invoked any source-level control (Filter, Mute, Unfollow, Block, or Report) in a condition. *Post-level follow-through* indicates whether exposure to a misinformation post resulted in a source-level action. *Action intensity* is the number of source-level actions performed per participant. *Control selection* records which control was chosen when an action occurred. *Escalation check* evaluates whether filtering a source was later followed by unfollowing or blocking the same source within the session. *Non-adoption* denotes participants who performed no source-level actions despite exposure. All measures rely solely on observable events and do not use self-reported judgments.

5.7 Interview Data

Semi-structured interviews were conducted to contextualize observed behavioral patterns and to identify usability constraints that influence defensive adoption. Interview data are used as behavior-linked explanatory evidence for interpreting logged actions and non-actions. Behavioral logs are the primary evidence of intervention effectiveness.

Two researchers reviewed recordings and notes while referencing the corresponding interaction logs. For each participant, the researchers reviewed interview comments alongside that participant's interaction logs. This allowed explanations to be interpreted in relation to observed actions or non-actions. Similar behavior-linked explanations were then grouped across participants into recurring themes. The resulting themes captured recurring reasons for decisions about warnings or source-level actions (e.g., perceived effort, social cost, curiosity, trust calibration). Researchers iteratively compared interpretations and resolved disagreements through discussion. The resulting themes were used to explain observed behavioral patterns and failure modes.

The qualitative analysis was therefore used to provide behavior-linked explanations for the logged interaction patterns, not to develop a standalone typology of participant attitudes or a general theory of misinformation response.

5.8 Ethics Statement

The study protocol was approved by the institutional review board. All misinformation content was low-stakes and de-

Interaction	Control	Inline	Action	Blur
Like	33	13	4	2
Share	7	1	1	2
Repost	8	4	1	0
Block	0	0	8	9
Unfollow	0	2	14	6
Mute	0	1	1	4
Report	3	11	25	19
Filter	0	0	9	5
Show me why	-	44	32	29
Learn more	-	12	3	5
See post (click)	-	-	-	79
Keep blurred (no click)	-	-	-	8

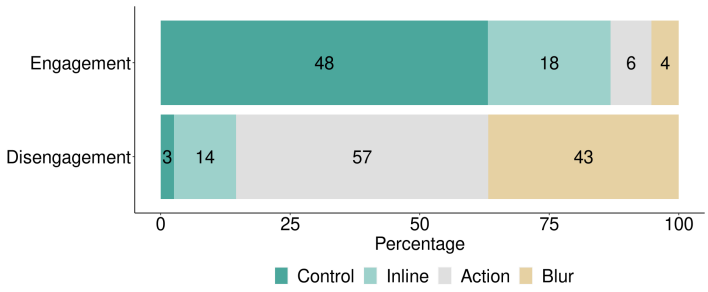


Figure 2: Behavioral log data from the TimeLineR study. The table (left) reports post-level interaction counts for misinformation posts and interventions. For the Blur condition, "See post" / "Keep blurred" indicate whether participants clicked to reveal blurred items or left them hidden. The bar chart (right) summarizes the relative balance between engagement (likes, shares, reposts) and source-level defensive actions (block, unfollow, report, filter). They illustrate how interventions shift behavior from passive engagement to active disengagement.

signed to avoid real-world harm. Participants were fully debriefed after the session. The debrief explicitly identified all synthetic misinformation items and provided the corresponding corrections. Participants were informed that no interactions affected real accounts or platforms.

6 Results

We report results by RQ, focusing on defensive uptake, post-level follow-through, control selection, escalation, and non-adoption across the three intervention designs.

6.1 Defensive Uptake Varies with Intervention Presentation

Figure 2 summarizes behavioral interactions across conditions, showing how interventions shift activity from engagement toward source-level action. Table 1 in Appendix A provides participant-level action distributions by condition and action type.

We first examine whether participants invoked any source-level defensive control under each intervention condition. Defensive uptake was rare in the absence of intervention: only 3.4% of participants (1/29) performed any source-level action in the Control timeline. Uptake increased to 27.6% (8/29) in the Inline warning condition, where controls were available but not foregrounded. When source-level controls were surfaced at the moment of encounter, uptake rose sharply. In the Action condition, 79.3% of participants (23/29) invoked at least one defensive action, and 55.2% (16/29) did so in the Blur condition. All participants completed all conditions. Thus, we used exact McNemar tests for paired binary uptake comparisons, with Holm correction across planned comparisons. For Inline vs. Action, 15 participants acted in Action but not Inline, while no participants acted in Inline but not Action ($p = 6.1 \times 10^{-5}$, Holm-corrected). For Inline vs. Blur, 10 participants acted in Blur but not Inline, while 2 participants acted in Inline but not Blur ($p = 0.039$, Holm-corrected). No participant-level condition data were missing.

Thus, the presence of misinformation cues alone did not reliably lead users to invoke source-level defensive actions . Defensive uptake increased substantially only when source-level actions were immediately visible and executable at the point of exposure.

6.2 Inline Warnings Reduce Engagement but Fail to Elicit Follow-through

Inline warnings altered how participants interacted with misinformation posts but did not translate increased attention into defensive action. As shown in Figure 2, engagement actions decreased relative to Control (e.g., likes dropped from 33 to 13), while participants frequently consulted explanations ("Show me why" was clicked 44 times).

However, despite this attention shift, defensive behavior remained limited (uptake 27.6%, Section 6.1). No participants blocked or filtered misinformation sources in this condition, and stronger account-level actions remained uncommon (unfollow occurred twice, report occurred 11 times). When actions occurred, they were sparse: participants averaged only 0.48 defensive actions each. These results indicate that inline warnings successfully prompted inspection but rarely led to source-level defensive action : users noticed and evaluated the content, but typically continued without acting on the source.

6.3 Action-oriented Source-level Controls Increase Follow-through

The Action-oriented source-level controls condition (hereafter Action) produced the highest volume and diversity of

source-level defensive actions, see Figure 2: Compared to Inline warnings, this condition sharply increased unfollow (14), introduced substantial use of block (8) and filter (9), and resulted in the highest report count (25).

The Action condition substantially increased *defensive follow-through* at the level of individual misinformation posts. Under the Action condition, 63.2% (55/87) of misinformation posts elicited a direct defensive action, compared to 16.1% (14/87) in the Inline condition and 3.4% (3/87) in Control. Since each participant contributed three misinformation-post encounters per condition, these aggregate counts are descriptive. We therefore compared participant-level paired follow-through rates using an exact Wilcoxon signed-rank test, appropriate for the small sample and discrete rate values. Action exceeded Inline by 47.1 percentage points on average (95% CI: [31.0, 63.2]; $p = 2.4 \times 10^{-6}$). Participant counts for 0/3, 1/3, 2/3, and 3/3 acted-on posts were 21, 2, 6, and 0 in Inline, versus 6, 5, 4, and 14 in Action. No participant-level condition data were missing.

Interview accounts help explain this increase in follow-through. Participants described the inline controls as reducing the effort required to act and making a source-level response salient at the moment of encounter. P9 explained, "*if the button is accessible then more people will take action right away*". P22 contrasted the Action condition with warning-only presentation: "*In the version without buttons, I didn't take any action. But with the buttons, I actually blocked an account. Having the option right there encouraged me to act*".

To characterize follow-through, we classify each participant-post encounter based on whether a source-level defensive action occurred without prior engagement with explanatory content (e.g., "Show me why" or "Learn more"). Defensive responses in the Action condition were exclusively direct within our observations: none of the misinformation posts in this condition exhibited an explanation-mediated pattern in which participants first engaged with explanatory content and then took a source-level action.

Action-oriented controls elicited defensive actions from seven participants who did not act under Inline warnings or Blur overlays. This suggests that immediate, low-friction access to source-level controls can increase uptake for participants who otherwise did not use defensive actions.

6.4 Blur Overlays Gate Exposure but do not Prevent Viewing

The Interstitial Blur Overlay condition introduced an explicit gate-before-viewing moment but did not function as a hard barrier to content exposure. Across 87 blurred misinformation posts, participants clicked "See post" in 79 cases (90.8%, Figure 2). Thus, while blur reliably interrupted scrolling, most users ultimately chose to reveal the content.

Despite this high reveal rate, the Blur condition supported substantial source-level defensive action. Participants invoked block (9), report (19), and filter (5) actions under Blur. This exceeds the uptake observed in the Inline warning condition and approaches that of the Action controls. This indicates that the interstitial design increased the salience and accessibility of defensive responses, even when users viewed the content.

To assess whether blur encouraged action *prior to exposure*, we examine the ordering of interactions within each participant–post sequence. Across blurred misinformation posts, we observe no instances where participants took a source-level defensive action *before* clicking "See post" and then later revealed the content. Instead, defensive actions either occurred *after* reveal or without revealing the post at all. Specifically, in 39.1% (34/87) of blurred misinformation posts, participants clicked "See post" and subsequently took a source-level defensive action. In an additional 4.6% (4/87) of cases, participants took a defensive action without revealing the post.

Thus, while the blur overlay reliably inserts a gate-before-viewing interaction, when defensive actions occur, they typically follow *post inspection*. This indicates that blur alters interaction ordering by requiring an explicit reveal step, but it does not primarily shift defensive decision-making to a pre-exposure checkpoint.

Blur overlays also enabled defensive action for a small subset of participants (2/29) who did not take any source-level actions under either Inline warnings or Action-oriented controls. This suggests that gate-before-viewing interruption can occasionally prompt disengagement for a minority of otherwise inactive users.

6.5 Skewed Selection of Source-level Controls

Behavioral logs reveal a strong skew in how participants selected source-level defensive controls when responding to misinformation posts. In the Inline condition, defensive actions were dominated by reporting: report accounts for 78.6% (11/14) of all defensive actions, while unfollow appeared only twice and block or filter were not used. Even when controls were made visible and actionable (Action and Blur), reporting remained the most frequently selected defensive option (Action: 25/57, 43.9%; Blur: 19/43, 44.2%). Other source-level actions remained comparatively rare (e.g., mute occurred only once in Action and four times in Blur).

Participant-level rates show the same pattern. In Inline, 24.1% (7/29) of participants used report at least once, compared to 3.4% (1/29) who used unfollow and 3.4% (1/29) who used mute. No participants used block or filter. In contrast, the Action condition increased uptake across multiple controls (report: 48.3%, unfollow: 24.1%, block: 20.7%, filter: 20.7%). The Blur condition showed a similar distribution (report: 31.0%, block: 17.2%, unfollow: 10.3%, mute: 10.3%, filter: 10.3%). Reporting remained the modal choice across all intervention designs.

Thus, making source-level defenses visible and actionable increases uptake, but does not lead to uniform or balanced use

of available controls. Instead, defensive behavior concentrates on a small subset of options, even when multiple alternatives are equally accessible.

6.6 Justification Use Declines Unevenly

Participants frequently consulted justifications for why posts were flagged. However, this consultation changed across repeated exposures. In all conditions, the justification affordance was used heavily in aggregate ("Show me why" clicks: 44 in Inline, 32 in Action, and 29 in Blur). Deeper explanatory links were used sparingly ("Learn more": 12, 3, and 5 respectively).

To examine repeated-exposure patterns, we analyze "Show me why" click probability as a function of the exposure index to misinformation posts *within each intervention timeline*. In Inline, click rates were 55.2% (16/29) on the first exposure, 44.8% (13/29) on the second, and 51.7% (15/29) on the third. In Action, click rates changed from 41.4% (12/29) on the first exposure to 34.5% (10/29) on the second, then to 34.5% (10/29) on the third. In Blur, click rates changed from 37.9% (11/29) to 34.5% (10/29) to 27.6% (8/29) across the first, second, and third exposures.

The Blur overlay makes the warning unavoidable prior to viewing. Thus, repeated clicks in this condition indicate whether participants continued to inspect the explanation after noticing the warning. In Blur, among participants who clicked "Show me why" on the first flagged post, 54.5% (6/11) clicked again on the second, and 53.3% (8/15) clicked on the third given that they had clicked at least once earlier in the same timeline. This pattern is consistent with a repetition or habituation effect under interstitial designs. Other explanations, however, include reduced informational value of repeated explanations and ordinary scrolling behavior.

6.7 Filtering is Adopted as a Middle-Ground Defense when Made Visible

Filtering was used only when the interface explicitly presented it as a first-class option. Log data show zero filter actions in Inline, but non-trivial use in Action (9) and Blur (5). This suggests that visibility and placement were important conditions for adoption in our study.

To assess whether filtering acts as an intermediate step or endpoint, we analyze action sequences at the participant $\times$ source account level (grouping multiple misinformation posts from the same source). For both Action and Blur, filtering never preceded unfollowing or blocking for the *same account*: 0/13 participant-source account pairs that were filtered were later unfollowed or blocked within the session; filtering was final with respect to these stronger actions for the same source.

Participants frequently took stronger actions without first filtering the same account. In Action and Blur, we observe 31 participant-source pairs with unfollow/block and no prior filter (22 in Action; 9 in Blur). Among source-exposure reduction decisions (filter, unfollow, or block), filtering accounted for 29.0% (9/31) in Action and 30.8% (4/13) in Blur.

These results show that filtering is adopted when made visible. Further, filtering functioned primarily as a lower-cost final alternative: filtered sources were not later blocked or unfollowed by the same participant.

Interview accounts help explain why Filter functioned as a final alternative in these sequences. Participants described it as preserving access to ordinary posts from a source while suppressing misinformation-labeled content from that source, e.g., "*if I already have good experience with the account, I prefer to filter it instead of block it*" (P9), "*Filtering is better than blocking because it maintains the relationship but removes false content*" (P1). This relationship-preserving interpretation helps explain why filtered sources were not later unfollowed or blocked within the study session.

6.8 Non-adoption Persists for a Subset of Users Despite Salient Interventions

Behavioral logs show that a subset of participants did not adopt source-level defensive controls even when warnings were salient and actions were immediately accessible. We define non-adopters as participants who took zero source-level actions (Mute, Unfollow, Block, Report, or Filter). Under this definition, 13.8% (4/29) of participants were non-adopters.

All non-adopters encountered the full set of 12 misinformation posts presented during the study. Thus, non-adoption was not due to limited exposure. Importantly, the logs also show that non-adopters sometimes actively engaged with misinformation content, performing an average of 1.5 engagement actions (likes, shares, or reposts) per participant, compared to 0.76 among adopters. This suggests a behavioral trajectory in which some users continue to interact with misinformation posts while never invoking source-level defensive controls, despite repeated exposure and visible defensive affordances.

Interview accounts suggest several reasons for this non-adoption pattern. Some participants described low perceived payoff or indifference toward acting, e.g., "*Honestly, I just scrolled past. I didn't really care*" (P23). Others described action as too effortful under repeated exposure: "*There's too many posts. I would have to take action for every post*" (P11). In some cases, participants described continued engagement with false content because it was entertaining or socially appealing, "*If it's false but funny, I'll still share it*" (P27). This suggests that non-adoption can reflect low motivation, perceived action burden, or perceived value in the content, while not ruling out confusion or misunderstanding in some cases.

6.9 Summary by Research Question

RQ1 (Defensive uptake). Source-level defensive actions occurred rarely without intervention (3.4% Control), increased

with warning-only presentation (27.6% Inline), and rose further when actions were immediately accessible (55.2% Blur; 79.3% Action), with the highest uptake in the Action condition. Thus, intervention presentation affected whether participants invoked any source-level defense.

RQ2 (Intervention design). Interface design shaped both *follow-through* and *control selection*. Inline warnings reduced engagement and increased explanation-seeking, but rarely produced follow-through to source-level action. Action-oriented controls produced the highest follow-through and broadened the set of defenses participants used (including Filter). Blur reliably interrupted scrolling but did not prevent viewing (90.8% reveal). Filter appeared only in conditions where it was directly surfaced. No filtered source was later unfollowed or blocked within the session. Across designs, reporting remained the modal choice. A subset of participants remained non-adopters across intervention designs.

7 Discussion

The results reveal a mismatch between how misinformation interventions are typically designed and how users respond in routine use. In all conditions, recognizing misinformation rarely translated into protective action. Instead, responses were shaped by the availability, cost, and perceived meaning of source-level controls. We interpret these findings through actionability, justification, selective defenses (including filtering), non-adoption, and interruption.

7.1 Actionability as a Mechanism for Defensive Uptake and Follow-Through

A central finding of this study is that *actionability*, and not awareness alone, drives defensive uptake and follow-through. The Action-oriented intervention produced both the highest volume of source-level actions and the highest rate of direct follow-through at the post level. Unlike Inline warnings, which frequently prompted explanation-seeking without subsequent action, Action designs enabled users to disengage immediately at the moment of encounter.

Behavioral logs show that under Action-oriented controls, defensive responses were exclusively direct: users acted without first consulting justifications or navigating away from the post. As reported in Section 6.3, interview accounts indicate that visible, proximate controls reduced effort and made source-level action salient at the moment of encounter. This supports the interpretation that actionability reduced the gap between recognizing risk and responding to it.

From a usable security perspective, these results show that interaction cost governs not only whether users comply with single-instance warnings but also whether they disengage from misinformation sources. Prior work demonstrates that protective actions and control use depend on how easily controls can be discovered, understood, and executed [9, 18, 37, 40]. Our findings extend this to ongoing source relationships: making source-level actions immediately available converts recognition of misinformation into account-level disengagement, while warning-only designs leave users aware but inactive.

7.2 Justification as Early Support with Diminishing Returns

Behavioral logs show that justifications ("Show me why" links) are often consulted early but are consulted less consistently across repeated exposures. In all conditions, justification click rates are highest on the first exposure to flagged content and are lower on later encounters. The decrease is most consistent in the Blur condition, where the warning is presented before the post can be viewed.

Interview accounts help explain this pattern. Participants described using justifications early to establish trust in the system and calibrate their judgment, e.g., "*At first, the 'Show Me Why' feature gave me some trust and made me believe the warning*" (P22). However, repeated explanations could also become less useful once participants had seen similar rationale multiple times, e.g., "*After reading the same explanation a few times, I stopped clicking and just skipped the posts*" (P22). Another participant similarly explained, "*If it's like warning symbol then I immediately understand okay it's false information. I just skip that*" (P14).

These accounts suggest that justification can serve both as an early verification aid and as trust-building support. However, our data do not establish a single mechanism for declining consultation. Lower repeated use may reflect several overlapping factors, including habituation, reduced informational value of repeated explanations, ordinary scrolling behavior, or cases in which warnings were noticed but disregarded.

Prior usable-security work shows that repeated alerts and explanations can become background noise over time [11, 20, 31]. Our findings are consistent with that concern: explanations were useful to some participants early in the interaction sequence, but repeated generic explanations appeared to lose value. A design implication is that explanation mechanisms should remain informative across repeated encounters. Real-world warning systems also use content-specific or context-specific explanations, such as X Community Notes [1], which attach explanatory context to particular claims.

For system designers, this suggests that justification should not be treated as a persistent prerequisite for action. Lightweight explanation early in an interaction sequence may help establish trust. However, repeated exposure requires explanations that remain informative, not merely repetitive. Subsequently, defensive effectiveness depends primarily on maintaining actionability and minimizing friction, while preserving access to content-specific explanation when users want to inspect the basis of the warning.

7.3 Filtering as a Middle-Ground Defense

Filtering emerged as a viable disengagement option only when it was made visible and legible. This pattern is consistent with its role as a middle-ground source-level defense. Behavioral logs show that filtering was never used in conditions where it was hidden, but was adopted in both Action and Blur conditions once surfaced as a first-class option. However, Filter was novel and participants were not pre-trained on it. Thus, unfamiliarity or limited discoverability may have constrained its adoption.

Filtering functioned as a source-level action that participants did not follow with unfollowing or blocking within the study session. Once a source was filtered, participants did not subsequently invoke stronger disengagement actions for that account. Conversely, unfollowing and blocking were typically selected directly, without prior filtering. This indicates that filtering was not used as an intermediate step toward stronger controls within the observed interaction window.

Instead of forming a severity ladder, source-level defenses partitioned into relationship-preserving and relationship-severing strategies. Filtering was consistently used as a terminal action for sources participants wished to keep, while blocking and unfollowing were chosen directly for sources they rejected. We did not observe within-session escalation from filtering to stronger actions. This suggests that users selected a disengagement strategy at the moment of encounter instead of progressively strengthening their response.

As reported in Section 6.7, interview accounts indicate that participants interpreted Filter as a relationship-preserving control that suppresses misinformation-labeled content while preserving ordinary posts from the same source. In contrast, unfollowing or blocking were described as higher-cost, final actions, reserved for sources perceived as persistently misleading: "*If an account posts one or two false posts occasionally, filtering is fine. But if they post 20 or 30 misleading posts daily, I'd prefer to unfollow or block*" (P1).

From a usable security perspective, these findings position filtering as a distinct defensive endpoint within the study session. Filtering lowers the cost of source-level defense for socially embedded or ambiguous sources by offering a selective and reversible way to reduce exposure without requiring users to sever ties or incur social penalties. Prior work shows that users may avoid stronger actions against misinformation-sharing sources when those actions carry social or relational costs [4]. This may be especially important in real feeds, where misinformation may come from friends, family members, coworkers, or community figures, and where unfollowing or blocking can carry stronger social meaning.

Our findings support a view of source-level defenses as qualitatively different protective strategies. Users may select among these strategies based on context, relationships, and perceived risk. To preserve this agency and avoid silent suppression, filtering mechanisms should be reversible and user-inspectable, with visible indicators when filtering is active for a source. This aligns with prior work that low-cost, flexible controls increase adoption [9].

7.4 Non-Adoption Despite Actionable Designs

Even when source-level controls were made prominent and action-forward, a subset of participants never invoked source-level defenses. Importantly, behavioral logs show that these non-adopters were not passive: they engaged with misinformation posts more than adopters, performing nearly twice as many likes, shares, or reposts.

As reported in Section 6.8, interview accounts suggest that some non-adoption reflected low perceived payoff, action burden, or perceived value in the content. These explanations are consistent with the higher engagement observed among non-adopters, but they do not rule out confusion, misunderstanding, or uncertainty in other cases. This aligns with prior usable-security work showing that awareness and usability improvements do not reliably translate into protective action, particularly when users are affectively engaged, fatigued, or perceive low payoff [7, 29, 35].

From a security perspective, these results identify a boundary condition for user-mediated defenses. Prominent source-level controls increased adoption for many participants, but did not eliminate non-adoption. This suggests that defensive adoption depends not only on interface usability, but also on whether the action aligns with user incentives at the moment of encounter. When continued interaction provides perceived value, clearer or more actionable controls may have limited effect. In such cases, source-level controls may need to be complemented by platform-level mechanisms such as moderation, ranking changes, or exposure reduction.

7.5 Interaction Ordering under Interruption

The Blur intervention altered interaction ordering by introducing a gate-before-viewing step. Interview participants consistently described the blur overlay as interrupting fast, habitual scrolling and creating a momentary pause, e.g., "*If you add an extra step before someone can interact with a post, it makes them stop and think. When you're scrolling, it's kind of mindless. If you have to click to unblur the image, it interrupts that flow*" (P23). Another noted that blur forced attention in a way inline warnings did not: "*The blur came so I can stop there and actually think about what's happening*" (P14). This indicates that blur reliably disrupts automatic engagement and captures attention at the moment of encounter.

However, the Blur did not function as a hard barrier to exposure. Behavioral logs show that most participants chose to reveal blurred posts, and defensive actions typically occurred after inspection and not before. Participants described the blur as an obstacle that had to be cleared prior to evaluation, e.g., "*If it wasn't blurred, it would be easier for me to judge [..]*

It just adds more work" (P22). Others explicitly preferred to inspect content before deciding whether to act. This suggests that interruption shifted *when* engagement occurred, but not necessarily *whether* it occurred.

Prior work shows that interruption and visual salience can pause interaction and disrupt habitual scrolling, but do not by themselves guarantee protective follow-through [9, 12, 20]. Our findings extend this pattern to source-level misinformation defenses by showing that interruption advances attention but does not move the decision boundary: users still evaluate the content before invoking a source-level defense.

From a usable security perspective, interruption therefore changes timing but not decision criteria. Interstitial designs must be paired with immediately accessible and clearly legible action pathways if they are to translate attention capture into defensive behavior. Without such pathways, interruption risks becoming a transient delay instead of a mechanism for sustained source-level disengagement.

7.6 Summary

Source-level defensive actions follow a conditional defensive process shaped by three factors: trust in the warning, cost of acting at the moment of encounter, and relationship context with the source. These factors jointly determine whether users disengage and which control they choose.

Actionability primarily affects adoption by lowering immediate response cost. Justification establishes trust early but is not repeatedly consulted. Filtering operates as a relationship-preserving endpoint. In our session logs, filtered sources were not later unfollowed or blocked by the same participant. Interruption changes attention timing without altering decision criteria. Interface improvements shaped how users disengaged, with limits when continued interaction was valuable.

7.7 Limitations

Our findings are subject to several limitations.

Simulated environment and synthetic content. Our study used a controlled, simulated timeline with synthetically generated misinformation. This design allowed us to fix which posts counted as misinformation and isolate how users respond to source-level disengagement tools, independent of detection accuracy. The use of synthetic, low-stakes content, designed to avoid emotionally charged domains, minimized potential harm to participants. However, this setup cannot fully capture the emotional intensity, social stakes, longitudinal dynamics, or mobile-specific affordances of real-world misinformation exposure. Thus, our findings characterize controlled interaction mechanisms. Deployed adoption rates require field or longitudinal validation.

Repeated exposure and explanation use. The controlled timeline may also have shaped explanation use: early "Show me why" clicks may partly reflect initial curiosity about the feature, while later decline may reflect habituation, reduced informational value over the session, or other study-specific dynamics that may differ from everyday feed use.

Misinformation density. Each timeline contained three misinformation posts out of twelve total posts. While this density supported lab observation, it likely over-represented misinformation relative to many organic feeds.

Fixed stimuli and content effects. Misinformation stimuli were fixed within each timeline. This controls cognitive-difficulty variation across conditions, but introduces possible content-specific effects. To reduce this risk, posts were designed to be comparable in plausibility, stakes, and campus relevance, and outcomes are analyzed at the action level.

Non-adopters. We treat non-adopters as a small explanatory subset, not a typology of resistance. Larger studies are needed to distinguish confusion, low motivation, action burden, warning disagreement, and intentional engagement.

Dependence on upstream detection. For all studied source-level interventions, we assume that misinformation has already been identified by a detection or fact-checking pipeline. In deployments, upstream errors would directly affect which sources are filtered or flagged. Our results show how users engage with source-level defenses once surfaced. Future work should examine these mechanisms under imperfect detection.

Sample and context. Participants were drawn from a single university population and interacted with campus-related content. This population is younger and relatively homogeneous in education and platform familiarity, and participation was voluntary. These characteristics may affect perceived social cost, trust in warnings, and willingness to intervene. Our claims therefore target behavioral mechanisms and not population prevalence. Replication in broader and more heterogeneous populations is necessary to assess generalizability.

Scope of defenses studied. We focus on source-level disengagement tools embedded in timeline workflows. Complementary platform-level mechanisms such as enforcement, reputation systems, or community moderation remain outside our scope.

8 Conclusion

This work frames *source-level disengagement* as a usable-security defense against repeated exposure to misinformation sources. Using *TimeLineR*, we studied three ways of surfacing source-level controls and introduced *Filter*, a middle-ground option that hides misinformation-labeled posts from a source while preserving other content. Behavioral results show that defensive uptake depends primarily on *actionability*: warning-only designs reduced engagement but rarely elicited follow-through, while immediate, in-context controls substantially increased adoption. Deployable defenses should treat source-level controls as first-class security mechanisms, prioritizing actionability and relationship-preserving options.

References

[1] About Community Notes on X. X, https://help.twitter.com/en/using-x/community-notes.

[2] Facebook's Approach to Misinformation: Partnering with Third-Party Fact-Checkers. https://www.facebook.com/journalismproject/programs/third-party-fact-checking/selecting-partners.

[3] Clarity. Microsoft, https://clarity.microsoft.com, Last accessed 2025.

[4] Joshua Ashkinaze, Eric Gilbert, and Ceren Budak. The dynamics of (not) unfollowing misinformation spreaders. In *Proceedings of the ACM on Web Conference 2024*, pages 1115–1125, 2024.

[5] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*, 2020.

[6] Nadia M Brashier, Gordon Pennycook, Adam J Berinsky, and David G Rand. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5), 2021.

[7] Andrew Chadwick, Cristian Vaccari, and Johannes Kaiser. The amplification of exaggerated and false news on social media: The roles of platform use, motivations, affect, and ideology. *American Behavioral Scientist*, 69(2):113–130, 2025.

[8] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political behavior*, 42:1073–1095, 2020.

[9] Verena Distler, Gabriele Lenzini, Carine Lallemand, and Vincent Koenig. The framework of security-enhancing friction: How ux can help users behave more securely. In *Proceedings of the New Security Paradigms Workshop 2020*, pages 45–58, 2020.

[10] Ullrich K. H. Ecker, Ziggy O'Reilly, Jesse S. Reid, and Ee Pin Chang. The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*, 111(1):36–54, 2020.

[11] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1065–1074, 2008.

[12] Wolfgang Einhäuser, Christiane R Neubert, Sabine Grimm, and Alexandra Bendixen. High visual salience of alert signals can lead to a counterintuitive increase of reaction times. *Scientific Reports*, 14(1):8858, 2024.

[13] KJ Kevin Feng, Kevin Song, Kejing Li, Oishee Chakrabarti, and Marshini Chetty. Investigating how university students in the united states encounter and deal with misinformation in private {WhatsApp} chats during {COVID-19}. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 427–446, 2022.

[14] Chen Guo, Zifei Guo, Nan Zheng, Chengqi Guo, et al. All warnings are not equal: A user-centered approach to comparing general and specific contextual warnings against misinformation. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, 2024.

[15] Chen Guo, Nan Zheng, and Chengqi Guo. Seeing is not believing: a nuanced view of misinformation warning efficacy on video-sharing social media platforms. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–35, 2023.

[16] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[17] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12), 2017.

[18] Silas Hsu, Kristen Vaccaro, Yin Yue, Aimee Rickman, and Karrie Karahalios. Awareness, navigation, and use of feed control settings online. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[19] Nicolas Huaman, Alexander Krause, Dominik Wermke, Jan H Klemmer, Christian Stransky, Yasemin Acar, and Sascha Fahl. If you {can't} get them to the lab: Evaluating a virtual study environment with security information workers. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 313–330, 2022.

[20] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J Nathan Matias, and Jonathan Mayer. Adapting security warnings to counter online disinformation. In

30th USENIX Security Symposium (USENIX Security 21), pages 1163–1180, 2021.

[21] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan M Herzog, Ullrich KH Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, et al. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, 8(6):1044–1052, 2024.

[22] Chen Ling, Krishna P Gummadi, and Savvas Zannettou. "learn the facts about covid-19": Analyzing the use of warning labels on tiktok videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 554–565, 2023.

[23] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J Vera Liao, James Choi, Kaiyue Fan, Sean A Munson, and Alexis Hiniker. How the design of youtube influences user sense of agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[24] Cameron Martel and David G Rand. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, page 101710, 2023.

[25] Cameron Martel and David G Rand. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, 8(10):1957–1967, 2024.

[26] Alberto Monge Roffarello, Luigi De Russis, and Kai Lukoff. The digital attention heuristics: Supporting the user's attention by design. *ACM Transactions on Computer-Human Interaction*, 32(4):1–41, 2025.

[27] Alberto Monge Roffarello, Kai Lukoff, and Luigi De Russis. Defining and identifying attention capture deceptive designs in digital interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[28] Pujan Paudel, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Lambretta: learning to rank for twitter soft moderation. In *2023 IEEE Symposium on Security and Privacy*, 2023.

[29] Gordon Pennycook and David G Rand. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402, 2021.

[30] Carson Powers, Nickolas Gravel, Christopher Pellegrini, Micah Sherr, Michelle L Mazurek, and Daniel Votipka. " i can say i'm john travolta... but i'm not john travolta": Investigating the impact of changes to social media verification policies on user perceptions of verified accounts. In *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, pages 353–372, 2024.

[31] Robert W Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018.

[32] Ken Reese, Trevor Smith, Jonathan Dutson, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. A usability study of five two-factor authentication methods. In *Proceedings of the Fifteenth USENIX Symposium on Usable Privacy and Security*, pages 357–370, 2019.

[33] Filipo Sharevski and Benjamin Kessell. Fight fire with fire: hacktivists' take on social media misinformation. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, pages 19–36, 2023.

[34] Aaron Sharockman. We started fact-checking in partnership with Facebook a year ago today. Here's what we've learned. https://www.facebook.com/journalismproject/programs/third-party-fact-checking/selecting-partners, 2017.

[35] Sarah Tabassum, Cori Faklaris, and Heather Richter Lipford. What drives SMiShing susceptibility? A U.S. interview study of how and why mobile phone users judge text messages to be real or fake. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security*, 2024.

[36] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A usability evaluation of Let's Encrypt and Certbot: Usable security done right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, page 1971–1988, 2019.

[37] Kami Vaniea, Lujo Bauer, Lorrie Faith Cranor, and Michael K Reiter. Out of sight, out of mind: Effects of displaying access-control information near the item it controls. In *2012 Tenth Annual International Conference on Privacy, Security and Trust*, pages 128–136. IEEE, 2012.

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

[39] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

Table 1: Participant-level distribution of source-level defensive actions by condition. Cells report the number of participants who used each action at least once. Total action counts are shown in parentheses.

Action	Control	Inline	Action	Blur
Any source-level action	1	8	23	16
Report	1 (3)	7 (11)	14 (25)	9 (19)
Filter	0 (0)	0 (0)	6 (9)	3 (5)
Block	0 (0)	0 (0)	6 (8)	5 (9)
Unfollow	0 (0)	1 (2)	7 (14)	3 (6)
Mute	0 (0)	1 (1)	1 (1)	3 (4)

[40] Sarah Y. Zheng and Ingolf Becker. Checking, nudging or scoring? Evaluating e-mail user security tools. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, pages 57–76, Anaheim, CA, 2023.

A Participant-Level Action Distribution

Table 1 reports participant-level use of source-level defensive actions by condition. This table complements the aggregate action counts reported in Figure 2 by distinguishing total action volume from the number of participants who used each action at least once. The distribution shows that source-level defensive actions were not driven solely by one or two highly active participants. In Action, 23/29 participants used at least one source-level action, and in Blur, 16/29 did so. Control selection is uneven: reporting involved the largest number of participants and total actions across intervention conditions, while Filter, Block, Unfollow, and Mute were used by smaller subsets of participants.

B Study Stimuli and Explanatory Probes

C TimeLineR Interface and Architecture

Figure 3 summarizes the TimeLineR architecture. Posts are stored in the backend and served to a controlled timeline, where intervention logic is injected at render time. This design allows the study to hold content fixed while varying warning presentation, action visibility, and blur gating across conditions.

Figure 4 shows the TimeLineR feed interface used in the study. The interface presents posts in a vertically scrolling timeline with source accounts, standard post engagement actions, and source-level controls. This example illustrates how a flagged misinformation post appeared alongside ordinary feed content.

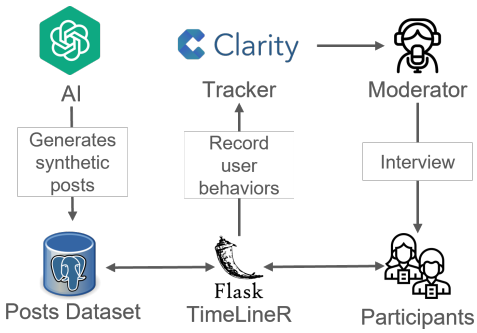


Figure 3: High-level architecture of TimeLineR. Posts are served to a controlled timeline where intervention logic is injected at render time. This enables precise control over exposure and action availability. All interactions (scrolling, engagement, source-level actions) are logged, to support reconstruction of defensive behavior and action ordering.



Figure 4: Example TimeLineR interface shown to participants. A flagged synthetic misinformation post with inline source-level defensive controls appears alongside a genuine post with standard account controls available via the menu.

Text of False Post	Debunk False Post	Intervention
1. Big steps for sustainability! From 2025, gas powered cars will be banned on campus. Let's embrace a greener future together! #EcoFriendly #SustainableCampus		Control
2. Great news, [UNI mascot name]! Starting August, all public transport in [redacted city name] is FREE for students. Get ready to explore the city! #FreeRides #StudentBenefits		Control
3. Exciting news! A [UNI] PhD student has been nominated for the Nobel Prize. #[UNI] #NobelPrizeNominee		Control
1. We did it, [UNI mascot name]! Our [University name] football team defeated the [Current College-Level Champions]! What a historic win! #[UNI]Champions #PantherPride	There is no record of this victory in the current or past seasons. Check out the NCAA Football Schedule.	Inline Warning
2. Double the fun, [UNI mascot name]! Next year's Spring Break is extended to two weeks. Time to plan some adventures! #[UNI]SpringBreak #ExtendedVacation	According to the official [UNI] academic calendar, Spring Break remains the standard one-week break. Check out [UNI] official academic calendar.	Inline Warning
3. Weekend treats just got better! Enjoy FREE food on campus every weekend. Cheers to endless delicious weekends! #[UNI]FreeFood #WeekendVibes	We found no evidence that [UNI] offers free meals to students every weekend. Dining services at [UNI] require payment or a meal plan. Check out the [UNI] Dining's official website.	Inline Warning
1. Attention all students! [UNI] has implemented a new policy restricting the use of Instagram on campus. Learn about the reasons behind this decision and share your thoughts on the impact of social media in academic settings. Join the conversation using #[UNI]. #InstagramBan #CampusLife.	We found no evidence that "UNI" restricted the use of Instagram. Check out the official [UNI] website for prohibited technologies.	Action Nudges
2. Dream come true! Messi will be playing for the [UNI] soccer team next season. Get ready for some magic on the field! #[UNI]Messi #SoccerLegends	Messi is a professional soccer player under contract with Inter Miami CF in Major League Soccer (MLS). He has no association with [UNI] or its soccer program. Check out official Messi professional career details.	Action Nudges
3. Challenge accepted! Get a 50% tuition reduction next year by sharing this post and gathering 500+ likes! Let's make education accessible. #[UNI]Challenge #TuitionReduction	[UNI] does not have any such program or initiative tied to social media activity. Check out the [UNI] tuition policies and financial aid options through the official financial aid page.	Action Nudges
1. Breaking news! [UNI] is making waves by changing its iconic [redacted name] mascot to a Crocodile! Get ready to show your new school spirit and stand tall with the mighty reptile. Share your excitement about this bold mascot change using #[UNI]Crocodile. Go team! #UniversityPride	The [redacted name] has been [UNI]'s official mascot since the university's founding, and no announcements or plans have been made to change it. Check out the [UNI] official logos information	Blurred Overlay
2. Impressive findings! A recent study reveals that graduates from [UNI] earn a staggering 10 times more than graduates from other institutions. Explore the factors behind this remarkable achievement and how it can impact your future career prospects. #Education #Success #Career	We found no study or data to support this claim. Check out the accurate salary and career outcomes for [UNI] graduates through U.S. Department of Education's College Scorecard.	Blurred Overlay
3. Important update! [UNI] system has made an announcement regarding a new dress code policy. Starting soon, t-shirts will no longer be allowed on campus. The goal is to foster a more professional and respectful environment. Share your thoughts on this change using #[UNI]. #DressCode #CampusLife	[UNI] has no policy, and no official announcement supports this claim. Check out the [UNI] official news website about campus and community.	Blurred Overlay

Table 2: The list of misinformation posts used in our study includes the debunking statements, and the type of intervention applied (Inline Warning, Action Nudges, or Blurred Overlay). The control timeline did not debunk the misinformation.

Probe Category	Purpose	Topics Covered
Non-engagement with flagged content	Understand why participants viewed misinformation but took no action	Noticing the warning, perceived credibility, attention, perceived relevance, visibility of controls
Engagement without defensive action	Explain interactions (like/share/reply) without disengagement	Motivation for engagement, perceived payoff, effort required to act, interpretation of warning
Defensive action selection	Explain why a specific control was chosen	Reason for mute/unfollow/block/report/filter, perceived severity, relationship considerations, reversibility
Interruption response	Assess reaction to blur gating	Attention capture, curiosity, annoyance, inspection before decision
Interface preference and usability	Compare intervention designs	Perceived usefulness, trust, clarity, effort, real-world likelihood of use
Real-world behavior calibration	Relate lab behavior to prior experience	Past encounters with misinformation, prior defensive actions, willingness to disengage

Table 3: Summary of post-task explanatory probes used to contextualize behavioral logs. The interview collected structured explanations for observed behaviors.