Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

User-Side Interventions Reduce Harmful Content Exposure in Algorithmic Feeds

Lucen Li[1], **Anshuman Chhabra**[2], **Magdalena Wojcieszak**[3, 4]
[1]Department of Computer Science, University of California, Davis, USA
[2]Bellini College of AI, Cybersecurity, and Computing, University of South Florida, USA
[3]Department of Communication, University of California, Davis, USA
[4]Center for Excellence in Social Sciences, University of Warsaw, Poland
lcnli@ucdavis.edu, anshumanc@usf.edu, mwojcieszak@ucdavis.edu

Abstract

Recommendation algorithms on social media platforms optimize for user engagement, which can inadvertently amplify exposure to harmful content such as violence, sexual material, and hate speech. Platform-level moderation is often delayed, opaque, and uniform, motivating the need for complementary user-side interventions that allow individuals to reduce unwanted content in their feeds without relying on platform cooperation. Prior work largely relies on single-session or human-subject studies, limiting the ability to capture recursive recommendation feedback loops, control for users' baseline preferences for harmful content, or systematically compare intervention strategies across harm types. To address these gaps, we propose a sock puppet simulation framework that models 30 rounds of iterative recommendation and interaction. We evaluate two user-side interventions: Downranking and Replacement, on YouTube's Homepage and Up-Next interfaces, controlling for users' baseline harm exposure levels (0%, 25%, 50%), yielding 18 experimental conditions with 1,000 puppets each. Our results show that user-side interventions are effective relative to the baseline, with effects concentrated on the Homepage interface. In particular, Downranking emerges as the most robust and consistent strategy, producing statistically significant improvements in both final-state outcomes (net change) and cumulative harmful exposure across baseline preference levels. For example, Downranking reverses baseline increases in harm into significant decreases (e.g., from a $+0.6$ percentage-point increase to a -0.9 percentage-point reduction), and yields durable reductions in cumulative exposure over time. Replacement shows weaker and less consistent effects on final-state outcomes. We further find no strong evidence of heterogeneous intervention effects across harm types, and observe that the overall reduction in harmful recommendations is largely driven by declines in Physical harm. Our work establishes the long-term efficacy of user-side interventions and provides guidance for their design.

Code — https://github.com/lucenl/yt-intervention

1 Introduction

Social media platforms are reshaping how people consume information and form perceptions of the world (Van Dijck and Poell 2013). The majority of platforms use recommendation systems to maximize user engagement by matching the recommended content to users' inferred preferences (Covington, Adams, and Sargin 2016). However, users tend to engage more with negativity (Robertson et al. 2023), out-group hate (Yu, Wojcieszak, and Casas 2024), and other potentially problematic content (Beknazar-Yuzbashev, Jiménez-Durán, and Stalinski 2024; Roggenkamp 2025). Engagement-driven optimization can be socially problematic. Indeed, evidence suggests that recommender systems inadvertently expose users to harmful content, such as violence, self-harm material, sexual content, misinformation, or hate speech (Wu et al. 2019; Nigatu and Raji 2024; Döring 2020; Hussein, Juneja, and Mitra 2020). This can lead to severe real-world consequences, ranging from undermining public health (Denniss and Lindberg 2025), rising risk of self-harm and suicidal ideation in youth (Hamilton et al. 2025; Susi et al. 2023), offline violence and hate crimes targeting marginalized communities (Müller and Schwarz 2025). Consequently, the mitigation of algorithmically-driven harm on social media platforms becomes an important societal problem.

Among social media platforms, YouTube in particular emerges as the core platform for examining recommendations to harmful content, and approaches to mitigate such recommendations in particular. It is the most popular platform, used by 85% of adults and 90% of teens in the U.S. (Center 2025a,b). It has over 2.7 billion monthly active users globally who collectively watch more than 1 billion hours of videos every day (Goodrow 2017). YouTube's recommendation algorithms drive 70% of time users spend on the platform (Rodriguez 2022), primarily via its *Homepage* (a broad, personalized content feed generated by a user's preference) and the *Up-Next* queue (a video sequence related to the video currently being played). In addition, YouTube is the "Great Radicalizer" (Tufekci 2018) that leads users to extremism and conspiracy theories (Ribeiro et al. 2020; Haroon et al. 2022), and its recommendation system also directs vulnerable users to environments rich in violent, sexualized, and harassing content, which can exacerbate anxiety (BBC 2018; Gkolemi et al. 2022). Given YouTube's popularity and its algorithm's role in leading users to harmful content, it is essential to design and test interventions that mitigate algorithmic-driven harms.

Indeed, the platform itself already employs *platform-level*

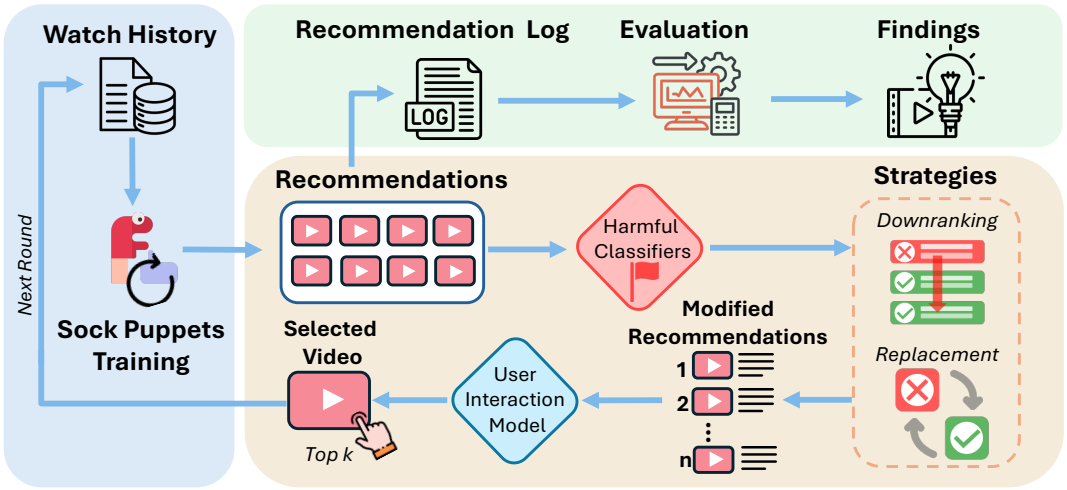

Figure 1: An overview of our experimental workflow for sock puppet-based intervention. The diagram illustrates the three stages of the simulation: (1) Pre-intervention (Blue) training, where puppets are initialized with different baseline content preferences using an initial watch history; (2) Intervention (Orange), where puppets receive YouTube recommendations that are classified in real time for harmful content and modified using Downranking or Replacement before a video is selected and added to the watch history; and (3) Post-intervention evaluation (Green), where original recommendations are logged and analyzed to measure changes in harmful content over time.

interventions: actions, tools, or strategies designed to minimize recommendations and consequent exposure to harmful content on the platform. The core mechanism is *content moderation* (YouTube 2025a; Roberts 2022; Vaccaro et al. 2021), an organized effort to screen and manage user-generated content based on community policies, typically relying on automated detection (Gorwa, Binns, and Katzenbach 2020; Gongane, Munot, and Anuse 2022), user reporting (Crawford and Gillespie 2016), and human review (Petrakaki and Kornelakis 2025; Barrett 2020).

While these efforts have proven effective in reducing overall harmful content (Buntain et al. 2021; Report 2025; Reddit 2024; X 2024), this centralized approach faces critical limitations. First, moderation actions suffer from delays, with 70.7% of moderation decisions taking more than 30 days (Trujillo, Fagni, and Cresci 2025), allowing harmful content to circulate and cause negative effects long before it is assessed. In addition, platform moderation remains a "black box" to scholars and users. Moderation decisions are often opaque and the overall process is inconsistent, evidenced by transparency reports showing that only approximately 37% of user-flagged videos result in actual removal(Report 2025). Last but not least, this centralized model leaves users powerless. Specifically, users may wish to diverge from the platform's engagement-driven recommendations and moderation standards. For example, users report wanting to see less divisive, negative, or moral-outrage-based content and more accurate, nuanced, and educational content on platforms (Rathje et al. 2024), yet they do not have control over what is recommended and whether the content they see as harmful is indeed minimized. Similarly, content that is permissible under general platform guidelines may still pose significant risks to vulnerable individuals, such as children or users with specific mental health conditions (e.g., eating disorders and PTSD) who may be triggered by certain types of content (Yasaroglu 2020; Urman, Hannak, and Makhortykh 2024).

To overcome the latency and intransparency of centralized moderation as well as the lack of user agency, we propose a bottom-up, *user-side intervention* approach. This framework serves as a normative safety intervention: User agency is operationalized not through preference alignment, but by empowering users to enforce established platform norms more transparently and rapidly than centralized moderation. Unlike platform-level moderation which removes videos from servers, our interventions operate on the client side directly. They intervene at the point of recommendation, adjusting the visibility or ordering of content in the user's feed in real-time, without requiring cooperation from the platform. Prior work in this space ranges from "soft" approaches like nudging and warning labels (Donabauer et al. 2024; Zubairu, Abdou, and Matrawy 2025), designed to enhance user awareness of potential risks, to "hard" algorithmic interventions that directly modify the feed. This includes user-configurable filtering (Bhargava et al. 2019), automatic content hiding (Beknazar-Yuzbashev et al. 2025), and feed re-ranking based on alternative objectives, such as increasing diversity (Kolluri et al. 2026). Recent work focuses on "reductive" strategies, notably *Downranking* (systematically lowering the rank of target content) and *Replacement* (substituting target content with an alternative one). However, while these strategies have promise in reducing animosity (Jia et al. 2024), existing research largely focuses on short-term experiments and static contexts.

In particular, current research, often based on single sessions or over short durations (Oak et al. 2025; Jia et al.

2024), provides limited insights into how interventions might influence the algorithm's feedback loop over time. Without observing this feedback loop, we cannot know if an intervention actually retrains the platform to be safer, or just hides a single bad video in the moment. Second, existing studies struggle to isolate the influence of users' varied baseline preferences[1] for harmful content (Kolluri et al. 2026; Piccardi et al. 2024). More specifically, a user who rarely watches harmful content has a different starting point than a user with a deep history of seeking it out. Human-subject studies often face challenges in accurately measuring these preferences through surveys (Kolluri et al. 2026). Even when studies account for baseline heterogeneity by observing pre-treatment exposure (Piccardi et al. 2024), they lack direct experimental control over a user's digital history. Without the ability to exogenously manipulate these backgrounds, it remains difficult to answer whether the intervention worked because of the strategy itself, or simply because the user had no interest in harmful content to begin with. Lastly, although researchers target specific harms like misinformation or hate speech in isolation (Beknazar-Yuzbashev et al. 2025; Zubairu, Abdou, and Matrawy 2025), there is a lack of a more comprehensive investigation into how various intervention strategies perform across distinct categories of harm within a unified framework. For instance, is downranking similarly effective in reducing sexual content or physical harms as it is in reducing hate and harassment? Without comparing these categories side-by-side, it remains unclear if a strategy is universally effective, or if it fails against specific types of harm.

To address these gaps, we propose an intervention framework that operates on platform recommendations before users are exposed to or engage with harmful content. We use sock puppets, automated agents that mimic user interactions, to experimentally control user background and observe longitudinal feedback effects (Bandy 2021; Hussein, Juneja, and Mitra 2020). We train puppets with different baseline preferences by exposing them to initial video histories containing 0%, 25%, or 50% harmful content, and then run a 30-round simulation in which each round represents an iterative cycle of recommendation and user interaction. In each round, we test two different interventions side-by-side: (1) Downranking (moving harmful content lowest-ranked positions within our fixed-length feed) and (2) Replacement (substituting harmful content with benign alternatives), resulting in a modified recommendation feed for each sock puppet.

We evaluate intervention effectiveness using two complementary metrics: net change in harmful recommendations between Round 0 and Round 30, and cumulative harmful exposure aggregated across all 30 rounds. Analyses focus on three harm categories: Physical, Sexual, and Hate and Harassment, and are conducted across YouTube's two primary recommendation surfaces, the Homepage and Up-Next (see Appendix A.1 for category definitions). In summary, we make the following contributions:

- We contribute a sock puppet simulation framework that integrates multi-class real-time harm classification with recursive feedback loops, in which simulated users select programmatically from within the offered recommendations, offering a new standard for auditing algorithmic interventions, one that accounts for dynamic feedback loops and active user selection.
- We test whether "algorithmic steering" is possible: can specific user-side interventions reduce recommendations to harmful content generated by the platform's *original* recommendations compared to a baseline?
- We evaluate intervention effectiveness across different user baseline preferences (0%, 25%, 50% harmful training histories), assessing how Downranking and Replacement perform over time for users with varying prior exposure to harmful content.
- We examine whether intervention effects vary across harm categories: Physical, Sexual, and Hate and Harassment, testing whether reductions in harmful recommendations are category-specific or aggregate-driven.
- We provide empirical evidence that intervention design critically shapes long-term effectiveness: Downranking consistently outperforms Replacement, suggesting that reinforcing users' inferred benign preferences is more effective than substituting harmful content with context-agnostic alternatives.

2 Related Work

2.1 Social Media Harms

Prior studies define and measure online harm in different frameworks, ranging from platform-specific guidelines to academic taxonomies and text-based constructs such as toxicity and incivility. Jo and Wojcieszak (2025) propose a unified taxonomy of online harm for YouTube videos that synthesizes prior works and platform guidelines into six categories, including Physical harm, Sexual harm, Hate and harassment, etc. These categories map onto the major enforcement domains in YouTube's safety policies, including violent extremism. Platform transparency reports show that these domains are associated with large volumes of enforcement actions. For example, between July and September 2025, YouTube removed 392,847 videos for violent extremism, 7,073,677 videos for child safety violations, and 260,335 videos for hate speech (Report 2025). Consistent with this, user reports and audits reveal widespread exposure to such content online. For instance, 66% adults report that they witnessed harmful content online (Enock et al. 2023).

Exposure to such harmful content can have detrimental effects on individuals, groups, and society at large. These impacts are often disproportionately severe for vulnerable populations. For children and adolescents, exposure to harmful content is linked to increased risks of anxiety, depression, body dissatisfaction, and potentially mimicking dangerous behaviors (Office of the Surgeon General 2023; Wei-

[1]In this study, we use the term "Baseline Preference" to denote the initial exposure level derived from historical interactions, rather than internal content preferences. This only characterizes the starting state for subsequent recommendations, which are influenced by prior watch history.

gle and Shafi 2024). Similarly, for individuals with pre-existing mental health conditions such as eating disorders or post-traumatic stress disorder (PTSD), encountering triggering content can exacerbate symptoms (Abdalla et al. 2021; Secker and Braithwaite 2021).

In addition, algorithmic audits show that recommender systems can create detrimental feedback loops: initial user interactions with harmful content can lead the algorithm to recommend more such content in the future, potentially increasing users' exposure to harmful content over time. For instance, initial engagement with borderline, sensational, or extremist content is associated with systematic shifts in subsequent recommendations toward more harmful material (Ribeiro et al. 2020; Hussein, Juneja, and Mitra 2020). Recent work quantifies this process, showing that engagement with harmful content can increase the probability of receiving harmful recommendations (Griffiths et al. 2024).

For these reasons, many users, such as parents, caregivers, and members of marginalized communities, express a desire to reduce their exposure to harmful or triggering content and to have greater control over what they see on social media platforms (Rathje et al. 2024).

2.2 Platform-Level Interventions

Given the negative consequences of recommendations and subsequent consumption of harmful content, platforms themselves address multiple types of harms through content moderation systems that combine automated detection with human review (Report 2025). In addition, extant research has proposed varied classifiers to identify content such as hate speech (Mukherjee and Das 2023), violent content (Khan, Tahir, and Ahmed 2018; Constantin et al. 2020), erotic content (Tabone et al. 2021; Nguyen, Wilson, and Dalins 2024), and misinformation (Wu et al. 2019). Recent studies have explored the use of LLMs for detecting harmful content (Liu et al. 2025; Bonagiri et al. 2025).

Although platforms emphasize the scale of their content moderation efforts and report substantial removals of harmful content (Report 2025; Reddit 2024; X 2024), independent audits identify persistent limitations that undermine their effectiveness and allow harmful content to remain accessible. First, moderation actions often suffer from substantial operational delays. For example, more than 70% of moderation decisions on YouTube and Pinterest occur over 30 days after content creation, limiting their ability to prevent exposure (Trujillo, Fagni, and Cresci 2025). Second, platform self-reporting lacks meaningful transparency: comparative analyses show that transparency reports are frequently incomplete or inconsistent, constraining external evaluation of moderation practices (Mündges and Park 2024; Chang et al. 2025). Third, beyond opacity, evidence suggests that platform interface design can actively discourage user reporting, for instance through hard-to-find reporting channels or dissuasive warnings (Wagner et al. 2020). Together, these limitations indicate that platform-level moderation alone is insufficient to address online harms, motivating the need for complementary approaches.

2.3 User-Side Interventions as a Complementary Safeguard

As a supplement to platform-level moderation, researchers explore user-side interventions, which operate directly on the user's device to modify the information flow to a specific user without altering the source content on the server or the way it can be displayed to others.

On the "softer" end are approaches aimed at enhancing users' awareness or critical evaluation without directly altering the feed's content. This involves interface-level interventions, such as nudging and warning labels. For example, Donabauer et al. (2024) used nudging principles to design interface cues that improve users' recognition of harmful content. Zubairu, Abdou, and Matrawy (2025) designed and evaluated warning labels for misinformation, demonstrating their potential and limitations.

In addition, studies use browser extensions to implement interventions, enabling user-side content modification, such as customized content filtering from social media feeds for better user control (Bhargava et al. 2019) and automatically hiding toxic content from the interface (Beknazar-Yuzbashev et al. 2025).

Building on browser-based approaches, other work proposes algorithmic intervention strategies that directly modify social media feeds. Jia et al. (2024) developed democratic attitude feeds using downranking and content removal-and-replacement strategies to reduce anti-democratic content exposure. Kolluri et al. (2026) enabled user-customizable re-ranking through 78 pluralistic values with real-time content scoring. Piccardi et al. (2024) provide a comprehensive framework supporting re-ranking and content editing interventions. Recent work also applies LLMs to algorithmic intervention, with Oak et al. (2025) demonstrating pairwise re-ranking effectiveness across six harm categories using zero-shot and few-shot approaches.

Extending earlier efforts, we present a systematic framework for evaluating intervention effectiveness in controlled yet realistic platform environments over extended periods, with analysis of differential impact across harm categories.

3 Intervention on YouTube to Mitigate Exposure to Harmful Content

3.1 Problem Formulation

Simulated Users and Background Representation. We use u to represent a simulated user (i.e., a sock puppet agent). The simulation progresses in discrete time steps, denoted by $t \in \{1, 2, 3 \dots\}$. Simulated users vary in their baseline preference for harmful content. To characterize this baseline preference prior to intervention, we define their *Baseline preference (BP)*. This metric is calculated after an initial "training" period of K time steps, during which the simulated user interacts with the platform without intervention. It is defined as the proportion of harmful content within the history accumulated during this period, $\mathcal{H}(u, K)$:

$$BP(u) = \frac{\sum_{v \in \mathcal{H}(u,K)} h(v)}{|\mathcal{H}(u, K)|}.$$

The history $\mathcal{H}(u, K)$ serves as the initial state for the subsequent "intervention" phase ($t > K$).

Intervention Target: Recommendation Feed. While $\mathcal{H}(u, t-1)$ represents the user's cumulative *past* watching history, during the intervention phase, our intervention specifically targets the platform's *current* recommendations at time step t. We denote this recommendation feed as $\mathcal{R}^{(s)}(u, t)$, where $s \in \{\text{Homepage, Up-Next}\}$ represents the distinct YouTube interfaces being simulated. Note that these recommendations are generated based on the user's prior history, $\mathcal{H}(u, t-1)$.

Intervention Framework and Objective. We define the *Harmful Exposure (HE)* of a given recommendation feed $\mathcal{R}$ as the proportion of harmful videos it contains:

$$HE(\mathcal{R}) = \frac{\sum_{v \in \mathcal{R}} h(v)}{|\mathcal{R}|}.$$

Thus, the harmful exposure from the platform's original feed from recommendation interface s at time t is $HE(\mathcal{R}^{(s)}(u, t))$.

Let $\mathcal{F}$ denote the set of available intervention strategies, namely Downranking (re-ordering based on harm probability) and Replacement (substituting harmful items with safe alternatives), which we systematically test. An intervention function $f \in \mathcal{F}$ maps the original recommendation feed $\mathcal{R}$ generated by the platform to a modified feed $\mathcal{R}'$ presented to the user: $f : \mathcal{R}^{(s)}(u, t) \mapsto \mathcal{R}'^{(s)}(u, t)$

At time step t, the simulated user selects a video v_t from the modified feed $\mathcal{R}'^{(s)}(u, t)$. This selected video is then used to update the cumulative history: $\mathcal{H}^{(s)}(u, t) = \mathcal{H}^{(s)}(u, t-1) \cup \{v_t\}$. In this way, the intervention f indirectly shapes the user's watching history, which the platform uses to generate future recommendations.

Our goal is to examine whether intervention functions $f \in \mathcal{F}$ can effectively steer the platform's recommendations toward a less harmful state. Specifically, we aim to reduce the harmful exposure generated by the platform's recommendation algorithms, $HE(R^{(s)}(u, t))$, throughout the intervention period.

3.2 Sock Puppet Simulation Framework

Our intervention evaluation relies on sock puppet simulation to assess intervention effectiveness in YouTube's recommendation environment under controlled conditions. Sock puppets are automated agents widely used to audit platform behavior under controlled conditions (Haroon et al. 2023).

In our project, implemented as automated web browsers, each puppet maintains a persistent browser profile (cookies, watch history) to ensure that YouTube provides personalized and session-consistent recommendations. Our simulation follows a three-phase workflow: (1) a pre-intervention training phase to establish baseline user profiles; (2) an intervention phase in which a fixed strategy $f \in \mathcal{F}$ is applied at each time step; and (3) a post-intervention evaluation phase analyzing harmful exposure in collected recommendations.

3.3 Harm Classification Model

We employ a two-stage harm classification pipeline based on RoBERTa-large to identify harmful videos throughout training, intervention, and evaluation.

Binary Classifier. The primary classifier takes a video's textual metadata (title and description) as input and outputs a harm probability score $\sigma(v) \in [0, 1]$. A binary harm label $h(v)$ is assigned using a threshold τ, where $h(v) = 1$ indicates a harmful video and $h(v) = 0$ otherwise. This binary signal directly triggers the intervention strategies.

Category Classifier. For videos labeled as harmful, a secondary multi-class classifier assigns a harm category $c(v) \in \mathcal{C} = \{\text{Hate, Sexual, Physical}\}$ based on the highest predicted probability. These labels are used for category-specific post-hoc analysis. Training and validation details are provided in Appendix B.

The primary goal of our study is to evaluate the intervention framework itself, in which the classifier serves as a modular "plug-in". The threshold τ is a methodological choice to test the framework's effectiveness under a consistent safety standard. Consequently, the framework can accommodate any alternative classifier or threshold that aligns with a user's personal definition or specific context of harm.

3.4 Intervention Strategies

We implement two intervention strategies designed to minimize recommendations to and consequent exposure to harmful content in users' platform ecosystem lowering the likelihood of user interaction with such content.

Downranking. This strategy re-ranks the original recommendation list $\mathcal{R}^{(s)}(u, t)$ based on predictive harm scores, such that videos deemed more likely to be harmful based on our classification are placed further down the sequence. Formally, the new ranking is given by $\mathcal{R}'^{(s)}(u, t) = \text{sort}_{v \in \mathcal{R}^{(s)}(u,t)} (\sigma(v))_{\text{asc}}$, where $\sigma(v)$ denotes the classifier's predicted probability of video v being harmful.

Replacement. Unlike Downranking, this strategy operates on the strict binary label $h(v)$. It substitutes harmful videos with benign alternatives. We maintain a harmless video pool $\mathcal{P}$, initialized with harmless videos identified from the training set, and dynamically expanded with new harmless videos encountered during simulation. For each video v_t in $\mathcal{R}^{(s)}(u, t)$ labeled as harmful, the intervention replaces it with a randomly chosen video from the pool $\mathcal{P}$. Consequently, the modified feed $\mathcal{R}'^{(s)}(u, t)$ contains content that was not classified as containing hate, sexual, or physical harm.

3.5 User Interaction Model

In order to simulate how real users might interact with a ranked list of recommended videos (i.e., select a video to watch), we employ a position-based decay probabilistic model. This approach captures the position bias in recommendation systems, where users are more likely to interact with items ranked higher in the list (Zhou et al. 2016; Zhou,

Khemmarat, and Gao 2010). We provide the formal formulation of the simulation and evaluation metrics in the Appendix C.

3.6 Evaluating Intervention

To quantify the effectiveness of interventions, we derive metrics based on the harmful exposure definition $HE(\mathcal{R})$ established in Section 3.1. All evaluations are performed separately for each YouTube's interface $s \in \{\text{Homepage, Up-Next}\}$ comparing the intervention strategies ($f \in \mathcal{F}$) against the no-intervention baseline. We denote the start of the intervention as $t_{start} = K$ and the end as $t_{end} = T$.

Net Exposure Change. This metric assesses the intervention effect by computing the shift in the platform's recommendation state between the start to the end of the intervention. First, for each individual puppet u, we calculate the paired difference in harmful exposure generated by the platform's raw recommendations: $\Delta_u^{(s)} = HE(\mathcal{R}^{(s)}(u, t_{end})) - HE(\mathcal{R}^{(s)}(u, t_{start}))$.

We then aggregate these individual shifts to estimate the population-level Net Exposure Change:

$$\widehat{\Delta\mu}^{(s)} = \frac{1}{M}\sum_{u=1}^{M}\Delta_u^{(s)}$$

where M is the total number of puppets. A negative difference indicates that the intervention was effective compared to the control group.

Cumulative Harmful Exposure (CHE) This metric captures the cumulative exposure risk over the entire experimental window . It is defined as the total proportion of harmful videos encountered across all time steps and all puppets:

$$CHE^{(s)} = \frac{\sum_{u=1}^{M}\sum_{t=K+1}^{T}\sum_{v\in\mathcal{R}^{(s)}(u,t)} h(v)}{\sum_{u=1}^{M}\sum_{t=K+1}^{T}|\mathcal{R}^{(s)}(u,t)|}.$$

A lower CHE compared to the baseline indicates superior protective efficacy throughout the simulation.

4 Experimental Setup

4.1 Experimental Design

We design a multi-round intervention workflow to capture the long-term impact of intervention strategies. The experiment controlled for three key factors: (1) Intervention Strategy ($f \in \{\text{Downranking, Replacement, None}\}$); (2) User Background (Prior Harm Exposure) (0%, 25% and 50%); (3) Recommendation Interfaces ($s \in \{\text{Homepage, Up-Next}\}$). Crossing these factors resulted in 18 unique experimental conditions. For each condition, we simulated $N = 1,000$ independent sock puppets. Each simulation ran for $T = 30$ rounds to capture longitudinal effects.

4.2 Dataset and Classifier Validation

We constructed a composite dataset (1,268 harmful and 68,818 harmless) derived from two annotated YouTube collections. The harmful subset is sourced from the Meta-Harm dataset (Jo and Wojcieszak 2025), which contains six categories of online harm. From this collection, we focused on three categories: Hate, Sexual, and Physical harm, selecting only the samples where all three labeling actors reached a consensus. The harmless subset is derived from a large-scale audit conducted between November 2022 and January 2023, representing actual user watch histories (Yu et al. 2024). These videos span diverse non-problematic topics, including verified news and educational content. All videos were passed through a YouTube Data API v3 filter in May 2025 to exclude age-restricted, region-blocked, or removed content, ensuring each video remained accessible for automated "watching" actions. Consistent with our focus on metadata-based detection, we utilize video titles and descriptions as the primary textual input for our models. Detailed harm category definitions and granular dataset breakdowns are provided in Appendix A.

We fine-tuned the RoBERTa-large binary classifier on a balanced subset of this dataset (details on splits and hyperparameters are provided in Appendix B). During inference, a strict threshold of $\tau = 0.8$ was applied to the predicted probability score $\sigma(v)$ to determine the binary label $h(v)$. On a held-out test set, the binary classifier achieved both an F1-score and Accuracy of 93.9%. The category-specific classifier used for post-hoc analysis achieved both an F1-score and Accuracy of 92.5%. Given these high performance metrics, we integrate these models as the harm classifiers within our simulation framework.

4.3 Simulation Configuration and Pipeline Automation

Consistent with large-scale audits (Haroon et al. 2023; Hussein, Juneja, and Mitra 2020), our simulation environment was configured following established protocols. To isolate the algorithmic response from confounding variables (e.g., user background), each puppet operated in an isolated Docker container and operated in a separate Google Chrome browser environment with a persistent profile. Our puppets operated in a "signed-out" state but maintained persistent HTTP cookies and local storage throughout the simulation lifecycle.

During the pre-intervention training phase, we initialized user profiles with varying degrees of prior harm preferences. This phase consisted of $K \approx 100$ video interaction. For each interaction, the puppet navigated to the target video and "watched" it for a duration of 30 seconds. We select this threshold because technical assessments and empirical audits indicate that a 30-second duration is sufficient for a "view" to be registered by YouTube's backend (Funk 2020; Haroon et al. 2023), and extending watch time does not yield significant differences in influencing subsequent recommendations (Chandio, Dar, and Nithyanand 2024). This process aimed to create distinct user profiles with consistent recommendation patterns before interventions.

The framework's automated pipeline handled the real-time data processing, including extracting recommendation metadata (video IDs, titles and descriptions) from the rendered YouTube interface. This data was asynchronously passed to the inference engine for harm classification. Based

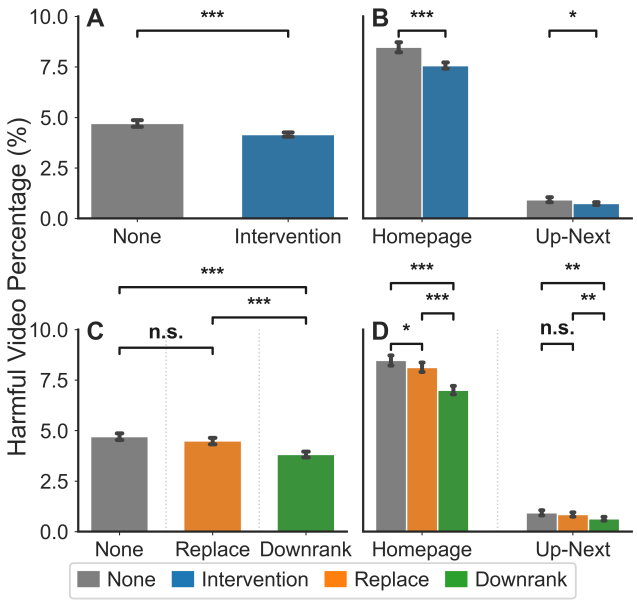

Figure 2: Overall effectiveness of user-side interventions after 30 rounds. Panel A compares final harmful video percentages between baseline and intervention conditions; Panel B shows results by surface (Homepage vs. Up-Next). Panels C–D compare intervention strategies across and within surfaces. Values reflect the final recommendation state at Round 30. Error bars show standard errors. Asterisks indicate significant pairwise differences (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

on the resulting labels, the pipeline executed the assigned intervention strategy and logged all interaction data (e.g., raw feed, modified feed, and user selection) for subsequent analysis.

5 Results

5.1 Overall Effectiveness

To establish the primary effects of user-side interventions, we first examine the final state of the recommendation ecosystem after 30 rounds of interaction across all users. Figure 2 presents a stepwise breakdown of the final-state effects of our interventions. We first compare intervention versus non-intervention conditions aggregated across all recommendation surfaces and user baseline preferences (Panel A). We then disaggregate results by surface type to assess whether intervention effects differ between the Homepage and Up-Next (Panel B). Finally, we compare specific intervention strategies both in aggregate and within each surface to evaluate their relative effectiveness (Panel C and D).

As shown in Figure 2A, intervention conditions result in a significantly lower final percentage of harmful recommendations compared to the non-intervention baseline (4.16% vs. 4.70%, $p < 0.001$). This confirms the overall effectiveness of user-side interventions in reducing harmful content in the final recommendation state. Figure 2B further breaks down these effects by recommendation surface. Our interventions significantly reduce harmful video percentages on both Homepage ($p < 0.001$) and Up-Next ($p < 0.05$). Notably, on the Homepage, harmful content constitutes a substantial share of baseline recommendations (8.48%), which is approximately 9.1 times higher than that of Up-Next (0.93%). The low baseline on Up-Next limits the absolute room for improvement, which helps explain the weaker statistical effects observed on this surface. Complete statistics are reported in Tables 4 and 5.

Consequently, while interventions are effective on both surfaces, their statistical impact is most pronounced on the Homepage. Figure 2C and D compare the effectiveness of specific strategies. Across both aggregated results and surface-specific analysis, Downranking consistently achieves significant reduction in harmful recommendations. While Replacement can significantly reduce harmful video percentage on Homepage, its effects are less consistent, especially showing non-significance on Up-Next. Overall, Downranking consistently outperforms Replacement in achieving a lower final harmful video percentage.

5.2 Intervention Effectiveness Across User Baseline Preferences

To account for heterogeneity in users' baseline preferences for harmful content, we evaluate intervention effectiveness across three training conditions (0%, 25%, 50%) using two complementary metrics: (1) Net Change ($\Delta\mu$), which compares the harmful video percentage between Round 0 and Round 30; and (2) Cumulative Exposure, which aggregates harmful exposure over all 30 rounds. Importantly, baseline preference levels reflect the proportion of harmful content in users' historical interactions used to train the puppets, representing controlled levels of prior exposure rather than internal preferences. These training levels do not directly correspond to the harmful percentage observed at Round 0. As a result, even users trained with 0% harmful content may receive non-trivial harmful recommendations at the start of the simulation.

We first examine net change in harmful recommendations ($\Delta\mu$). Figure 3 shows that on the Homepage, Downranking consistently outperforms the baseline across all training conditions, both mitigating algorithmic worsening for clean preferences and amplifying reductions under higher baseline exposure. For example, in the 0% condition, the baseline harm rate increases by $+2.5$ percentage points (pp), whereas Downranking limits this increase to $+0.8$ pp ($p < 0.001$), reducing the magnitude of worsening by approximately 67%. In the 50% condition, Downranking nearly doubles the net reduction in harm compared to the baseline (-3.1 pp vs. -1.7 pp, $p < 0.01$), representing an additional 81% reduction beyond the baseline trend. In the 25% condition, Downranking reverses a baseline increase of $+0.6$ pp to a reduction of -0.9 pp. Furthermore, direct pairwise comparisons between Downranking and Replacement show that Downranking achieves significantly larger net reductions than Replacement across all three training conditions, yielding an additional decrease of 1.4 pp, 1.2 pp, and 0.9 pp for the 0%, 25%, and 50% groups, respectively (all $p < 0.05$, see Table 7 for the full results). In contrast, in-

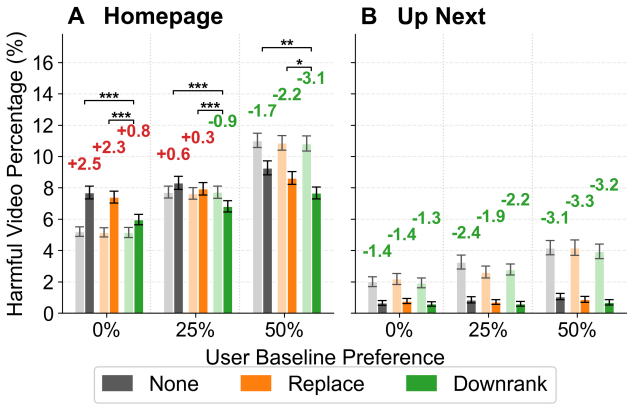

Figure 3: Net change in harmful video percentage ($\Delta\mu$) between Round 0 and Round 30 across Homepage and Up-Next recommendations. Bars show harmful video percentages across user baseline preference conditions (0%, 25%, 50%) and intervention strategies, with bootstrapped 95% confidence intervals. Values above the bars are the net change ($\Delta\mu$) between Round 0 and Round 30. Positive values (red) indicate worsening, while negative values (green) indicate improvement. Asterisks indicate statistically significant pairwise differences within each condition (two-sided permutation tests: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). The absence of brackets indicates non-significant differences (n.s.). On the Homepage, Downranking consistently achieves significant reductions relative to the baseline and outperforms Replacement across all conditions.

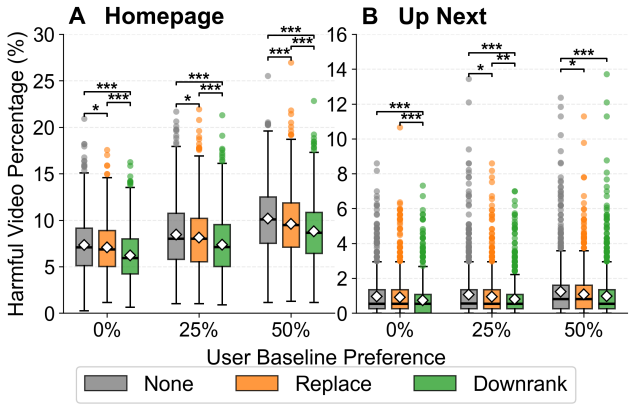

Figure 4: Distribution of cumulative harmful exposure (CHE) over the 30-round intervention. The figure shows results for (A) Homepage and (B) Up-Next recommendations, grouped by training condition (0%, 25%, 50%) and intervention strategy: No Intervention (gray), Replacement (orange), and Downranking (green). Box plots show the median and interquartile range (IQR); white diamonds indicate the mean. Asterisks denote significant differences for all pairwise comparisons within each training group (two-sided permutation tests: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).The absence of brackets indicates non-significant differences (n.s.).

tervention effects in the Up-Next environment are substantially weaker, none of the conditions reaches statistical significance in net change on Up-Next.

While net change captures differences between the final and initial states, cumulative exposure reveals a different pattern. Figure 4 shows the distribution of per-user cumulative harmful exposure, computed by summing harmful video percentages across all rounds. This metric captures sustained exposure over time and is particularly informative for Up-Next, where net changes are small.

Across both recommendation surfaces and all training conditions, Downranking yields consistent and statistically significant reductions in cumulative harmful exposure relative to the baseline. Replacement also produces significant cumulative reductions in most conditions, though its effects are weaker and less consistent. For instance, under the 25% training condition in Up-Next, cumulative harmful exposure decreases from 1.06% in the baseline to 0.94% with Replacement and 0.80% with Downranking, corresponding to relative reductions of 11.32% and 24.53%, respectively (detailed in Table 8 and 9). Direct comparisons confirm that Downranking yields significantly lower cumulative exposure than Replacement across all Homepage conditions (additional reductions of 0.75–0.96 pp; all $p < 0.001$) and most Up-Next conditions (additional 0.14–0.17 pp; $p < 0.01$), except for the 50% training condition on Up-Next where the difference is not statistically significant. Complete pairwise tests are reported in Table 10.

Importantly, these cumulative reductions occur even in the Up-Next interface, where baseline harmfulness is low and final-state net changes are correspondingly small. Taken together, these results highlight three key insights: (1) net change and cumulative exposure capture distinct but complementary aspects of intervention effectiveness; (2) Downranking is the most robust strategy, consistently improving both final recommendation states and total exposure; and (3) even interventions with limited net change effects, such as Replacement, can meaningfully reduce cumulative harm.

5.3 Category-Specific Analysis

To determine whether intervention effects vary by harm type, we decompose the aggregated results into three distinct categories: Physical Harm, Sexual Harm, as well as Hate and Harassment Harm. We analyze the net change in the harmful video percentages between Round 0 and Round 30, as shown in Figure 5. We further provide a decomposition and explanation in the Appendix E.5.

The results show a natural algorithmic drift in the baseline condition. Specifically, the platform demonstrates a substantial reduction in Physical Harm across both recommendation surfaces and user baseline preferences. In contrast, Sexual Harm and Hate and Harassment Harm exhibit different trends: On the Homepage, their percentages in the recommendation feed increase, while on Up-Next, their percentages decrease. Given that Physical Harm constitutes the majority of the initial recommendation feed in Round 0 across all conditions (e.g., 49.2% with 0% baseline

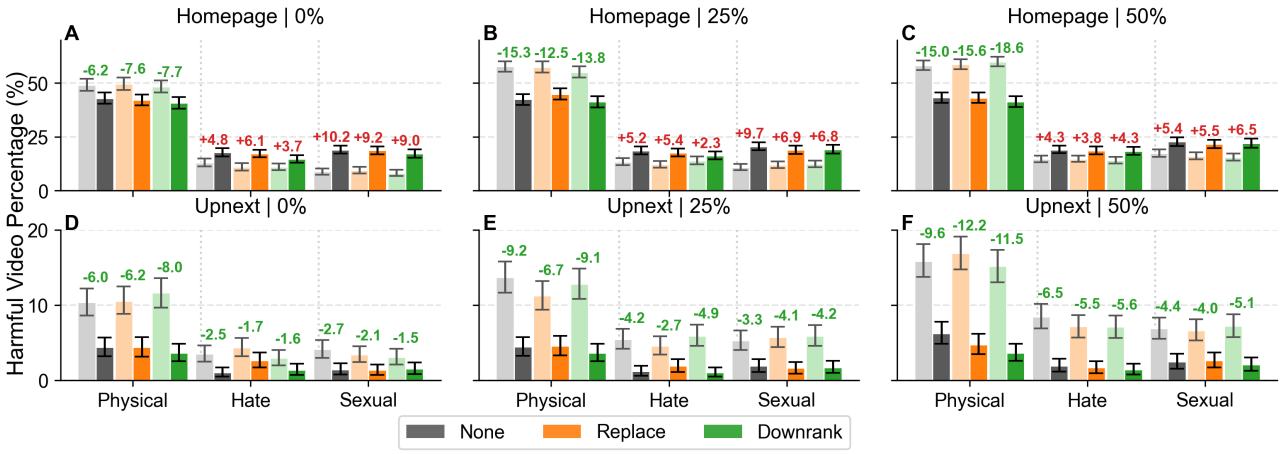

Figure 5: Category-specific net change in harmful content (Round 0 - Round 30). Physical harm shows strong baseline reductions, while Hate and Sexual harms exhibit weaker, surface-dependent trends. No intervention produces statistically significant additional reductions relative to the baseline at the category level ($p > 0.05$).

preference without intervention), its steep reduction mathematically contributes to the aggregate reduction in harmful videos. Thus, the effectiveness of our interventions is driven primarily by the reduction of Physical Harm, rather than a uniform mitigation across all categories.

More importantly, despite the superiority of Downranking observed in the aggregate results, we find that neither Downranking nor Replacement is statistically distinguishable from the baseline within individual categories. We report cumulative harmful video exposure in the Appendix as a complementary analysis. Consistent with the net change results, category-level cumulative patterns do not show a clear and uniform intervention advantage across harm categories (see Figure 6). These results suggest that while user-side interventions provide a general signal in reducing harmful content in recommendations over feedback loops, they struggle to override the algorithm's potential content-level biases. The platform appears to be highly responsive to mitigating Physical Harm, but remains less sensitive to the suppression of Sexual and Hate content.

6 Discussion

We present a simulation framework for evaluating user-side interventions aimed at mitigating exposure to harmful content in algorithmic feeds. By simulating longitudinal user interactions without relying on human subjects, the framework enables scalable and controlled evaluation of how interventions shape both short-term recommendation outcomes and feedback loops over repeated interactions.

Comparing Intervention Strategies. Our results highlight that intervention effectiveness depends on the evaluation perspective. Endpoint comparisons capture how interventions alter the final recommendation state, whereas cumulative exposure reflects sustained protection over repeated interactions. Across both metrics, Downranking consistently achieves larger reductions in harmful exposure than Replacement. These findings suggest that the observed reductions are not a purely "mechanical" result of position bias. While our framework incorporates a probabilistic interaction model, a mechanical interpretation would imply that Replacement, which completely removes harmful items from the feed, should be more effective at reducing exposure. However, our finding shows that Replacement underperforms Downranking. This suggests that replacing harmful videos with context-agnostic alternatives introduces noise that disrupts the recommendation logic and potentially trigger unpredictable harmful recommendations. Despite this noise, Replacement remains a viable strategy for reducing cumulative exposure in some settings. These findings underscore the importance of multi-dimensional evaluation: strategies that rapidly suppress harmful content may differ from those that primarily reduce aggregate exposure over repeated interactions.

Limited Category-Specific Effects. When decomposing results by harm type, we find no statistically significant differences in intervention effectiveness across categories. Reductions in aggregate harmful content are driven primarily by large baseline decreases in Physical harm, while Hate and Sexual harms exhibit weaker or mixed responses. This suggests that user-side interventions in our setting operate at an aggregate level rather than selectively mitigating specific harm categories, pointing to limits in their ability to override category-level dynamics of the underlying recommender.

Recommendation Surface Differences. Finally, we observe systematic differences between Homepage and Up-Next recommendations. While baseline harmfulness is substantially lower in Up-Next, interventions still reduce cumulative exposure over time. These differences likely reflect surface-specific recommendation mechanisms, with Up-Next being more directly shaped by the immediately preceding video. Our framework thus enables comparative evaluation of intervention effects across multiple recommendation contexts.

Protection Without Harmful Histories and Governance Implications. Notably, interventions remain effective even for users trained with no prior exposure to harmful content. In the 0% training condition, we observe that "clean" puppets still encounter a non-trivial baseline increase in harmful recommendations over time. This finding highlights a critical gap: because algorithmic feeds optimize primarily for engagement, platform-level moderation may fail to fully protect even users who actively avoid harmful content. Our results demonstrate that user-side approach effectively mitigate this risk: both Downranking and Replacement significantly reduce harmful recommendations, providing an autonomous layer of protection that empowers individuals. This property is particularly relevant for safeguarding vulnerable or restricted user populations. For policymakers, these findings establish a rationale for supporting user-side interventions as a flexible alternative to the inherent limitations of top-down, "one-size-fits-all" platform moderation.

7 Conclusion

We present a simulation-based framework for evaluating user-side interventions aimed at reducing exposure to harmful content in recommendation systems. By enabling controlled, large-scale experiments without involving real users, the framework supports systematic comparison of intervention strategies across different user training conditions and recommendation surfaces. Our results show that user-side interventions such as downranking and replacement can meaningfully reduce harmful exposure, including for users with limited prior interaction with harmful content. The findings highlight the importance of multi-metric evaluation, reveal heterogeneous effects across harm categories, and demonstrate effectiveness on both the Homepage and Up-Next interfaces. Future work will extend this framework to multimodal harm detection, explore category-specific intervention designs, and validate the findings through controlled studies with real users on live platforms.

8 Limitations

While the user-side interventions show significant effects in mitigating harmful content exposure, this study has certain limitations. several limitations. First, our harm detection relies on text-based metadata, which may not generalize to multimodal signals that dominate video platforms. Second, although simulation enables controlled and scalable experimentation, simulated user behaviors cannot fully capture the complexity of real user interactions with recommendation systems. Third, our analysis focuses on three harm categories that are reliably detectable through textual features, and does not cover other harm types. Finally, the dynamic nature of recommendation algorithms and continuously changing content availability limits reproducibility, as recommendations depend on evolving trends and platform updates.

References

Abdalla, S. M.; Cohen, G. H.; Tamrakar, S.; Koya, S. F.; and Galea, S. 2021. Media exposure and the risk of post-traumatic stress disorder following a mass traumatic event: an in-Silico experiment. *Frontiers in Psychiatry*, 12: 674263.

Bandy, J. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5(CSCW1): 1–34.

Barrett, P. M. 2020. Who moderates the social media giants. *Center for Business*, 102.

BBC. 2018. The disturbing YouTube videos that are tricking children. BBC NEWS.

Beknazar-Yuzbashev, G.; Jiménez-Durán, R.; McCrosky, J.; and Stalinski, M. 2025. Toxic content and user engagement on social media: Evidence from a field experiment. Technical report, CESifo Working Paper.

Beknazar-Yuzbashev, G.; Jiménez-Durán, R.; and Stalinski, M. 2024. A model of harmful yet engaging content on social media. In *AEA Papers and Proceedings*, volume 114, 678–683. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Bhargava, R.; Chung, A.; Gaikwad, N. S.; Hope, A.; Jen, D.; Rubinovitz, J.; Saldías-Fuentes, B.; and Zuckerman, E. 2019. Gobo: A system for exploring user control of invisible algorithms in social media. In *Companion publication of the 2019 conference on computer supported cooperative work and social computing*, 151–155.

Bonagiri, A.; Li, L.; Oak, R.; Babar, Z.; Wojcieszak, M.; and Chhabra, A. 2025. Towards Safer Social Media Platforms: Scalable and Performant Few-Shot Harmful Content Moderation Using Large Language Models. *arXiv preprint arXiv:2501.13976*.

Buntain, C.; Bonneau, R.; Nagler, J.; and Tucker, J. A. 2021. YouTube recommendations and effects on sharing across online social platforms. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–26.

Center, P. R. 2025a. 5 facts about Americans and YouTube. Pew.

Center, P. R. 2025b. Teens and Social Media Fact Sheet. Pew.

Chandio, S.; Dar, M. D. P.; and Nithyanand, R. 2024. How audit methods impact our understanding of youtube's recommendation systems. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 241–253.

Chang, T.; Trybala III, J. J.; Bassan, S.; and Razi, A. 2025. Opaque Transparency: Gaps and Discrepancies in the Report of Social Media Harms. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–12.

Constantin, M. G.; Ştefan, L.-D.; Ionescu, B.; Demarty, C.-H.; Sjöberg, M.; Schedl, M.; and Gravier, G. 2020. Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing*, 13(1): 347–366.

Covington, P.; Adams, J.; and Sargin, E. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, 191–198.

Crawford, K.; and Gillespie, T. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *new media & society*, 18(3): 410–428.

Denniss, E.; and Lindberg, R. 2025. Social media and the spread of misinformation: infectious and a threat to public health. *Health promotion international*, 40(2): daaf023.

Donabauer, G.; Theophilou, E.; Lomonaco, F.; Bursic, S.; Taibi, D.; Hernández-Leo, D.; Kruschwitz, U.; and Ognibene, D. 2024. Empowering Users and Mitigating Harm: Leveraging Nudging Principles to Enhance Social Media Safety. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying@ LREC-COLING-2024*, 155–166.

Döring, N. 2020. Gendered hate speech in YouTube and YouNow comments: Results of two content analyses.

Enock, F.; Johansson, P.; Bright, J.; and Margetts, H. Z. 2023. Tracking experiences of online harms and attitudes towards online safety interventions: findings from a large-scale, nationally representative survey of the British public. *Nationally Representative Survey of the British Public (March 21, 2023)*.

Funk, M. 2020. How Does YouTube Count Views? Tubics.

Gkolemi, M.; Papadopoulos, P.; Markatos, E.; and Kourtellis, N. 2022. YouTubers Not MadeForKids: Detecting channels sharing inappropriate videos targeting children. In *Proceedings of the 14th ACM Web Science Conference 2022*, 370–381.

Gongane, V. U.; Munot, M. V.; and Anuse, A. D. 2022. Detection and moderation of detrimental content on social media platforms: current status and future directions. *Social Network Analysis and Mining*, 12(1): 129.

Goodrow, C. 2017. You know what's cool? A billion hours. YouTube.

Gorwa, R.; Binns, R.; and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945.

Griffiths, S.; Harris, E. A.; Whitehead, G.; Angelopoulos, F.; Stone, B.; Grey, W.; and Dennis, S. 2024. Does TikTok contribute to eating disorders? A comparison of the TikTok algorithms belonging to individuals with eating disorders versus healthy controls. *Body image*, 51: 101807.

Hamilton, J. L.; Untawale, S.; Dalack, M. N.; Thai, A. B.; Kleiman, E. M.; and Yao, A. 2025. Self-harm content on social media and proximal risk for self-injurious thoughts and behaviors among adolescents. *JAACAP Open*.

Haroon, M.; Chhabra, A.; Liu, X.; Mohapatra, P.; Shafiq, Z.; and Wojcieszak, M. 2022. Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations. *arXiv preprint arXiv:2203.10666*.

Haroon, M.; Wojcieszak, M.; Chhabra, A.; Liu, X.; Mohapatra, P.; and Shafiq, Z. 2023. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the national academy of sciences*, 120(50): e2213020120.

Hussein, E.; Juneja, P.; and Mitra, T. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on human-computer interaction*, 4(CSCW1): 1–27.

Jia, C.; Lam, M. S.; Mai, M. C.; Hancock, J. T.; and Bernstein, M. S. 2024. Embedding democratic values into social media AIs via societal objective functions. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–36.

Jo, W.; and Wojcieszak, M. 2025. MetaHarm: Harmful YouTube Video Dataset Annotated by Domain Experts, GPT-4-Turbo, and Crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2496–2509.

Khan, M.; Tahir, M. A.; and Ahmed, Z. 2018. Detection of violent content in cartoon videos using multimedia content detection techniques. In *2018 IEEE 21st International Multi-Topic Conference (INMIC)*, 1–5. IEEE.

Kolluri, A.; Su, R.; Jahanbakhsh, F.; Zhao, D.; Piccardi, T.; and Bernstein, M. S. 2026. Alexandria: A Library of Pluralistic Values for Realtime Re-Ranking of Social Media Feeds. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Liu, Y.; Zhu, J.; Liu, X.; Tang, H.; Zhang, Y.; Zhang, K.; Zhou, X.; and Chen, E. 2025. Detect, investigate, judge and determine: A knowledge-guided framework for few-shot fake news detection. In *2025 IEEE International Conference on Data Mining (ICDM)*, 477–486. IEEE.

Mukherjee, S.; and Das, S. 2023. Application of transformer-based language models to detect hate speech in social media. *Journal of Computational and Cognitive Engineering*, 2(4): 278–286.

Müller, K.; and Schwarz, C. 2025. Online Hate Speech, Offline Harm, and the Case for Content Moderation. In *EconPol Forum*, volume 26.

Mündges, S.; and Park, K. 2024. But did they really? Platforms' compliance with the code of practice on disinformation in review. *Internet Policy Review*, 13(3): 1–21.

Nguyen, T. T.; Wilson, C.; and Dalins, J. 2024. Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts. In *Proceedings of the 32nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2024)*.

Nigatu, H. H.; and Raji, I. D. 2024. "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 141–160.

Oak, R.; Haroon, M.; Jo, C. W.; Wojcieszak, M.; and Chhabra, A. 2025. Re-ranking using large language models for mitigating exposure to harmful content on social media platforms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 894–908.

Office of the Surgeon General. 2023. Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory.

Petrakaki, D.; and Kornelakis, A. 2025. What do content moderators do? Emotion work and control on a digital health platform. *Journal of Management Studies*.

Piccardi, T.; Saveski, M.; Jia, C.; Hancock, J. T.; Tsai, J.; and Bernstein, M. 2024. Reranking social media feeds: A practical guide for field experiments. *ACM Transactions on Social Computing*.

Rathje, S.; Robertson, C.; Brady, W. J.; and Van Bavel, J. J. 2024. People think that social media platforms do (but should not) amplify divisive content. *Perspectives on Psychological Science*, 19(5): 781–795.

Reddit. 2024. Transparency Report. Reddit.

Report, G. T. 2025. YouTube Community Guidelines enforcement. YouTube.

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A.; and Meira Jr, W. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 131–141.

Roberts, S. T. 2022. Content moderation. In *Encyclopedia of big data*, 211–214. Springer.

Robertson, C. E.; Pröllochs, N.; Schwarzenegger, K.; Pärnamets, P.; Van Bavel, J. J.; and Feuerriegel, S. 2023. Negativity drives online news consumption. *Nature human behaviour*, 7(5): 812–822.

Rodriguez, A. 2022. YouTube's recommendations drive 70% of what we watch. Quartz.

Roggenkamp, H. 2025. Toxic Language Captures Attention: Evidence from 200 Million News Headline Impressions.

Secker, R.; and Braithwaite, E. 2021. Social media-induced secondary traumatic stress: Can viewing news relating to knife crime via social media induce PTSD symptoms. *Psychreg Journal of Psychology*, 5(2).

Susi, K.; Glover-Ford, F.; Stewart, A.; Knowles Bevis, R.; and Hawton, K. 2023. Research review: Viewing self-harm images on the internet and social media platforms: Systematic review of the impact and associated psychological mechanisms. *Journal of child psychology and psychiatry*, 64(8): 1115–1139.

Tabone, A.; Camilleri, K.; Bonnici, A.; Cristina, S.; Farrugia, R.; and Borg, M. 2021. Pornographic content classification using deep-learning. In *Proceedings of the 21st ACM symposium on document engineering*, 1–10.

Trujillo, A.; Fagni, T.; and Cresci, S. 2025. The DSA Transparency Database: Auditing self-reported moderation actions by social media. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–28.

Tufekci, Z. 2018. YouTube, the great radicalizer. *The New York Times*, 10(3): 2018.

Urman, A.; Hannak, A.; and Makhortykh, M. 2024. User Attitudes to Content Moderation in Web Search. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–27.

Vaccaro, K.; Xiao, Z.; Hamilton, K.; and Karahalios, K. 2021. Contestability for content moderation. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2): 1–28.

Van Dijck, J.; and Poell, T. 2013. Understanding social media logic. *Media and communication*, 1(1): 2–14.

Wagner, B.; Rozgonyi, K.; Sekwenz, M.-T.; Cobbe, J.; and Singh, J. 2020. Regulating transparency? Facebook, twitter and the German network enforcement act. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 261–271.

Weigle, P. E.; and Shafi, R. M. 2024. Social media and youth mental health. *Current psychiatry reports*, 26(1): 1–8.

Wu, L.; Morstatter, F.; Carley, K. M.; and Liu, H. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2): 80–90.

X. 2024. Global Transparency Report H2 2024. X.

Yasaroglu, C. 2020. Youtubers' effect on children's values: Parents' views. *European Journal of Educational Sciences*, 7(4): 1–15.

YouTube. 2025a. How YouTube reviews content. YouTube.

YouTube. 2025b. YouTube Data API v3. YouTube.

Yu, X.; Haroon, M.; Menchen-Trevino, E.; and Wojcieszak, M. 2024. Nudging recommendation algorithms increases news consumption and diversity on YouTube. *PNAS nexus*, 3(12): pgae518.

Yu, X.; Wojcieszak, M.; and Casas, A. 2024. Partisanship on Social Media: In-Party Love Among American Politicians, Greater Engagement with Out-Party Hate Among Ordinary Users. *Political Behavior*, 46(2): 799–824.

Zhou, R.; Khemmarat, S.; and Gao, L. 2010. The impact of YouTube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 404–410.

Zhou, R.; Khemmarat, S.; Gao, L.; Wan, J.; and Zhang, J. 2016. How YouTube videos are discovered and its impact on video views. *Multimedia Tools and Applications*, 75(10): 6035–6058.

Zubairu, H.; Abdou, A.; and Matrawy, A. 2025. Evaluation Metrics for Misinformation Warning Interventions: Challenges and Prospects. *International Journal of Human–Computer Interaction*, 1–16.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) NA

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) NA

2.4. Proofs of all novel claims are included (yes/partial/no) NA

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) NA

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) NA

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) NA

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) NA

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) partial

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) partial

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) partial

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) Yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) partial

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) partial

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) partial

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) no

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes

Ethical Statement

This work studies user-side interventions for reducing harmful content exposure in algorithmic recommendations as a complement to platform-level moderation. The simulation framework does not involve real users and does not influence live recommendation systems. All data used were publicly available, and no personal or private information was collected or shared. Our goal is to develop scalable tools that enhance user agency and safety without replacing platform responsibility. We emphasize that user-side interventions are complementary safeguards rather than substitutes for platform moderation, particularly for vulnerable populations.

A Dataset Overview for Sock Puppet Training

To evaluate intervention effects of mitigating harmful content, we use sock puppets to simulate realistic users with different profiles. Specifically, we focus on harmful content users encountered during they watching trails. To supprot such simulation, we constructed a training dataset that combines harmful and harmless YouTube videos from two distinct sources.

For harmful content, we used a publicly available dataset containing YouTube videos labeled across various harm categories (Jo and Wojcieszak 2025). Labels were assigned independently by three sources: Amazon Mechanical Turk workers, fine-tuned large language models (LLMs), and social science related domain experts. We selected a subset of the dataset in which the binary harmful label was unanimously agreed upon by all three sources. Additionally, we filtered the harmful subset to include only videos that were categorized as containing Hate and Harassment harm (HH), Sexual Harm (SXL), or Physical Harm (PH) by majority agreement across annotators.

While the dataset contains videos labels as "harmless", they are typically collected from potential harmful sources. Thus, we used different dataset for harmless videos.

Since sock puppets must perform web-based "watching" actions during training, all videos in the training set must remain available on YouTube. We used the YouTube Data API v3 (YouTube 2025b) to check video availability. Videos were excluded if they were: (1) Age-restricted, (2) Region-restricted in the U.S. (3) Removed or resulted in retrieval errors.

As of the experimental period (May 2025), 1,268 harmful videos and 68,818 harmless videos passed the availability filter and were retained for training process.

A.1 Harmful Category Breakdown

Harmful Content Category Composition. Regarding the composition of harm categories in the sock puppet training dataset, we counted the number of relative proportions of each harm category content among all available harmful videos. Among the 1,268 videos, HH accounted for 37.5% (476), SXL for 34.8% (441), and PH for 27.7% (351).

Harm Category Definitions. **Sexual Harm (SXL):** Content containing erotic scenes or explicit imagery, depictions of sexual acts and nudity, sexual abuse scenarios, and sexually exploitative material. Examples include videos with suggestive thumbnails, explicit sexual content, and materials that sexualize individuals without consent. **Physical Harm (PH):** Content promoting or depicting self-injury, suicide, eating disorder promotion, dangerous challenges and pranks, and other activities that could result in physical harm to viewers or participants. Examples include self-harm tutorials, dangerous viral challenges, and content glorifying eating disorders. **Hate and Harassment (HH):** Content containing insults and obscenities, identity attacks or misrepresentation, and hate speech targeting individuals or groups based on gender, race, ethnicity, age, religion, political ideology, disability, or sexual orientation. Examples include discriminatory commentary, targeted harassment campaigns, and content promoting intolerance toward specific communities.

A.2 Sock Puppet Training

To simulate different user risk profiles, we defined three training conditions based on the percentage of harmful videos in each sock puppet's viewing history: 0%, 25%, and 50%. For example, in the 25% condition, each puppet watches 100 training videos, 25 of which are randomly sampled from the harmful dataset described above. These harmful videos are shuffled with 75 harmless videos to form the final training sequence. This graduated exposure design enables systematic exploration of how different levels of historical harmful content exposure influence subsequent recommendations and intervention effectiveness.

B Training Details for RoBERTa-based Classifier

B.1 Model Overview

We developed a two-stage classification pipeline using fine-tuned RoBERTa-large models[2] to identify and categorize harmful content in YouTube videos. The first stage employs a binary classifier to detect harmful content, while the second stage applies a multi-class classifier to further categorize detected harmful content into three specific harm types: Hate and Harassment (HH), Sexual Harm (SXL), and Physical Harm (PH). Both models operate exclusively on textual features extracted from YouTube video metadata, including titles, descriptions, and available transcript segments.

B.2 Training Dataset

We used the same data sources described in Appendix A, with slight modifications for model training. Unlike sock puppet training, video availability was not a constraint, as the classification task operates independently of video playback functionality. To ensure training quality, we implemented additional filtering to remove low-quality samples that are: (1) Visual-dependent videos with insufficient textual content; (2) Music, film, or art content lacking clear harm-related indicators; (3) Ambiguous samples requiring full transcripts beyond the RoBERTa token limit.

[2]https://huggingface.co/FacebookAI/roberta-large

Table 1: RoBERTa classifier training configuration and evaluation results.

(a) Optimized Training Hyperparameters

Hyperparameter	Binary	Multi-class
Learning rate	2.66e-6	1.94e-5
Weight decay	0.0054	0.0078
Scheduler type	Cosine	Cosine
Warmup ratio	0.145	0.135
Per-device batch size	8	8
Grad. Accum. Steps	8	8
Training epochs	15	15
Optimizer	AdamW with 8-bit quantization	
Threshold	0.8	0.8

(b) Evaluation Performance

Metric	Binary Classifier	Multi-class Classifier
Accuracy	93.9%	92.5%
Precision	93.9%	92.6%
Recall	93.9%	92.5%
F1-score	93.9%	92.5%

Table 2: Confusion matrices for binary and multi-class classification.

Binary Classification

	Pred. H	Pred. F
True H	271	9
True F	22	258

Multi-class Classification

	HH	SXL	PH
True HH	61	2	4
True SXL	5	60	2
True PH	1	1	65

To maximize training data reliability, We selected only instances where the binary harmful label was unanimously agreed across all three annotator types (MTurk workers, GPT-4, and domain experts). For multi-class classification, only harmful samples were used, where the multi-class label was agreed by majority vote from three annotator types.

As a result, our dataset consists 2,240 examples for training and and 560 for validation and testing. We utilized title, description and transcript to train the model.

B.3 Training Configuration and Evaluation Results

Table 1 summarizes the optimized hyperparameters and evaluation performance of the RoBERTa classifiers.

The confusion matrices for the binary and multi-class settings are shown in Table 2.

B.4 Integration with Main Experimental Framework

In our experimental pipeline, the RoBERTa classifiers provide real-time harmful content detection and harm category identification during sock puppet interaction sessions, enabling dynamic intervention implementation as harmful content is encountered in new recommendation outcomes. To optimize computational efficiency and response times, we implemented a Redis-based caching system that stores classification results for previously classified videos.

C User Interaction Model

Formally, for a video v located at rank i (where $i = 1$ is the top position) within the modified recommendation list $\mathcal{R}^{(s)}(u,t)$, we define its selection probability $P(v)$ as:

$$P(v \mid \text{rank} = i) = \frac{\lambda^{i-1}}{\sum_{j=1}^{|\mathcal{R}'|} \lambda^{j-1}}$$

where $\lambda \in [0,1]$ is a decay factor controlling the steepness of the user's attention curve, and the denominator serves as the normalization constant to ensures $\sum P(v) = 1$.

At each time step t, the sock puppet samples exactly one video to watch according to this probability distribution. We set $\lambda = 0.9$ across all of our experimental conditions.

D Additional Analysis: Category-Level Effects

D.1 Pre-Post Category Composition Changes

Pre-Intervention Harms Composition The proportion of physical harm instances in our sock puppet training dataset is relatively lower compared to Hate and Sexual content categories (27.7% vs. 37.5% and 34.8% respectively). Given our random sampling methodology from the harmful training dataset, the category composition retained in individual sock puppet training sequences should theoretically mirror this distribution.

Post-Intervention Harms Composition Canges Our results show physical harm content consistently comprises 50-70% of harmful content encountered in post-training recommendation outcomes, This gap suggests that existing recommendation algorithms may amplify physical harm content, or moderation systems may more effectively detect and filter hate and sexual content, while physical harm content may be overlooked. Table 3 shows the complete pre-post intervention effects results in category-specific perspective.

E Full Results Across All Experimental Conditions

E.1 Complete Statistics of Overall Intervention Effectiveness

Table 4 and 5 provide the complete results underlying Figure 2. The first table reports the final-state harmful video percentages, standard errors, and sample sizes for each plotted group. The second table reports the corresponding pairwise statistical comparisons, including raw and Holm-adjusted p values.

E.2 Complete Descriptive Statistics of Pre-Post Harmful Content Exposure

Table 6 presents comprehensive descriptive statistics for harmful content exposure across all 18 experimental conditions, comparing Round 0 (pre-intervention) and Round 30 (post-intervention) measurements. Table 7 reports the direct

Table 3: Category-level pre–post intervention effects across all experimental conditions. Values report the proportional composition of harm categories within harmful content before intervention (Pre; Round 0) and after intervention (Post; Round 30). Physical harm generally shows the largest proportional shifts across conditions.

Focus	Intervention	Category	Percentage (%)					
			0%		25%		50%	
			Pre	Post	Pre	Post	Pre	Post
Homepage	None	Hate	18.34	22.28	16.39	22.82	16.32	22.36
		Physical	69.19	53.86	70.34	51.98	64.37	50.81
		Sexual	12.47	23.86	13.27	25.21	19.31	26.83
	Replace	Hate	15.70	22.01	15.00	21.69	16.48	22.27
		Physical	70.78	53.99	70.29	55.09	65.50	51.72
		Sexual	13.52	24.00	14.71	23.22	18.03	26.01
	Downrank	Hate	16.30	20.25	17.15	21.24	15.73	22.49
		Physical	71.53	56.07	67.61	53.76	66.89	50.55
		Sexual	12.17	23.67	15.23	25.00	17.38	26.97
Up-Next	None	Hate	19.49	15.22	22.37	16.23	27.11	18.22
		Physical	57.59	63.77	56.05	58.44	50.83	58.41
		Sexual	22.92	21.01	21.59	25.33	22.06	23.36
	Replace	Hate	23.65	31.55	21.25	23.46	23.33	18.68
		Physical	57.71	52.38	52.17	56.17	55.06	52.20
		Sexual	18.63	16.07	26.58	20.37	21.62	29.12
	Downrank	Hate	16.81	21.21	24.05	16.02	24.04	20.14
		Physical	65.80	55.30	52.00	56.90	51.41	50.69
		Sexual	17.40	23.48	23.95	27.08	24.55	29.17

Table 4: Final-state harmful video percentage used in Figure 2. Means and standard errors (SE) are reported in percentage points. N denotes the number of puppets in each plotted group. Panel labels correspond to Figure 2.

Panel	Surface	Strategy	Mean (%)	SE	N
A	All	None	4.70	0.085	6000
A	All	Intervention	4.16	0.054	12000
B	Homepage	None	8.48	0.124	3000
B	Homepage	Intervention	7.57	0.081	6000
B	Up-Next	None	0.93	0.065	3000
B	Up-Next	Intervention	0.74	0.035	6000
C	All	None	4.70	0.085	6000
C	All	Replacement	4.49	0.080	6000
C	All	Downranking	3.82	0.072	6000
D	Homepage	None	8.48	0.124	3000
D	Homepage	Replacement	8.13	0.118	3000
D	Homepage	Downranking	7.00	0.110	3000
D	Up-Next	None	0.93	0.065	3000
D	Up-Next	Replacement	0.84	0.053	3000
D	Up-Next	Downranking	0.64	0.047	3000

Table 5: Pairwise statistical comparisons shown in Figure 2. Reported p values are Holm-adjusted values from the final Figure 2 analysis pipeline. Significance labels match the plotted annotations.

Panel	Surface	Comparison	p_{Holm}	Sig.
A	All	None vs Intervention	6.27e-08	***
B	Homepage	None vs Intervention	1.03e-09	***
B	Up-Next	None vs Intervention	0.0115	*
C	All	None vs Replacement	0.0694	n.s.
C	All	Replacement vs Downranking	3.00e-04	***
C	All	Downranking vs None	3.00e-04	***
D	Homepage	None vs Replacement	0.0419	*
D	Homepage	Replacement vs Downranking	3.00e-04	***
D	Homepage	Downranking vs None	3.00e-04	***
D	Up-Next	None vs Replacement	0.3111	n.s.
D	Up-Next	Replacement vs Downranking	0.0088	**
D	Up-Next	Downranking vs None	0.0012	**

pairwise comparisons between Replacement and Downranking for net change.

E.3 Complete Descriptive Statistics of Cumulative Harmful Content Exposure

Tables 8 and 9 present comprehensive descriptive statistics for cumulative harmful content exposure across all 18 experimental conditions, including intervention effectiveness against the baseline. Table 10 further reports the direct pairwise comparisons between Replacement and Downranking, showing that Downranking achieves significantly lower cumulative harmful exposure across all Homepage conditions and most Up-Next conditions.

E.4 Additional Results for Category-Specific Analysis: Cumulative Harmful Exposure

Figure 6 presents the distribution of cumulative harmful video percentage over the 30-round intervention for each harm type. Box plots show the median and interquartile range; while diamonds indicate means. Overall, these patterns do not indicate a clear and uniform intervention advantage across harm categories. Physical Harm appears more responsive to intervention, whereas Sexual and Hate content exhibit weaker and less consistent shifts across conditions.

E.5 Aggregate Harm Reduction Is Dominated by Physical Harm

We show that the observed aggregate reduction in harmful recommendations is mathematically dominated by decline in Physical harm. Let the overall harmful rate be defined as a weighted sum of category-specific rates. The net change in overall harm can be written as:

$$\Delta_{overall} = \sum_{c} w_c \dot{\Delta}_c$$

where w_c is the proportion of category c at Round 0, and Δ_c is the net change for that category.

Across all user baseline preferences and recommendation surfaces, Physical harm constitutes the largest share of initial recommendations (typically 45–60%), while Hate and Sexual harms together account for a smaller fraction. At the same time, Physical harm exhibits large negative net changes (e.g., -6 to -18 pp), whereas Hate and Sexual harms show positive or much smaller-magnitude changes.

As a result, the weighted contribution of Physical harm is consistently negative and substantially larger in magnitude than the combined contributions of the other categories. Consequently, the aggregate reduction in harmful recommendations is driven primarily by declines in Physical harm rather than uniform reductions across categories.

E.6 Factors of Intervention Effectiveness

Training Exposure Dependencies. Our results reveal relationships between sock puppet harmful exposure during training phase and intervention effectiveness. Overall, users with higher baseline harm preferences during training demonstrate greater susceptibility to intervention effects, particularly on the Homepage, where both intervention strategies produce significant cumulative reductions across all three training groups. Notably, even under the minimal training condition (0%), baseline harmful content remained at roughly 5% on Homepage and about 2% on Up-Next across conditions. This finding provides evidence that social media recommendation systems inherently introduce some level of harmful content despite platform moderation and user-side efforts.

Recommendation Focus Dependencies. We examined intervention effects across two distinct recommendation contexts: Homepage and Up-Next recommendations. These contexts shows different response patterns to interventions. Homepage recommendations demonstrated consistent and robust intervention effects, with reductions in both net change and cumulative exposure varying systematically. By contrast, Up-Next recommendations showed weaker effects in net change. This difference may be related to the sequential structure of Up-Next recommendations, which may limit the magnitude of observable shifts in final-state outcomes.

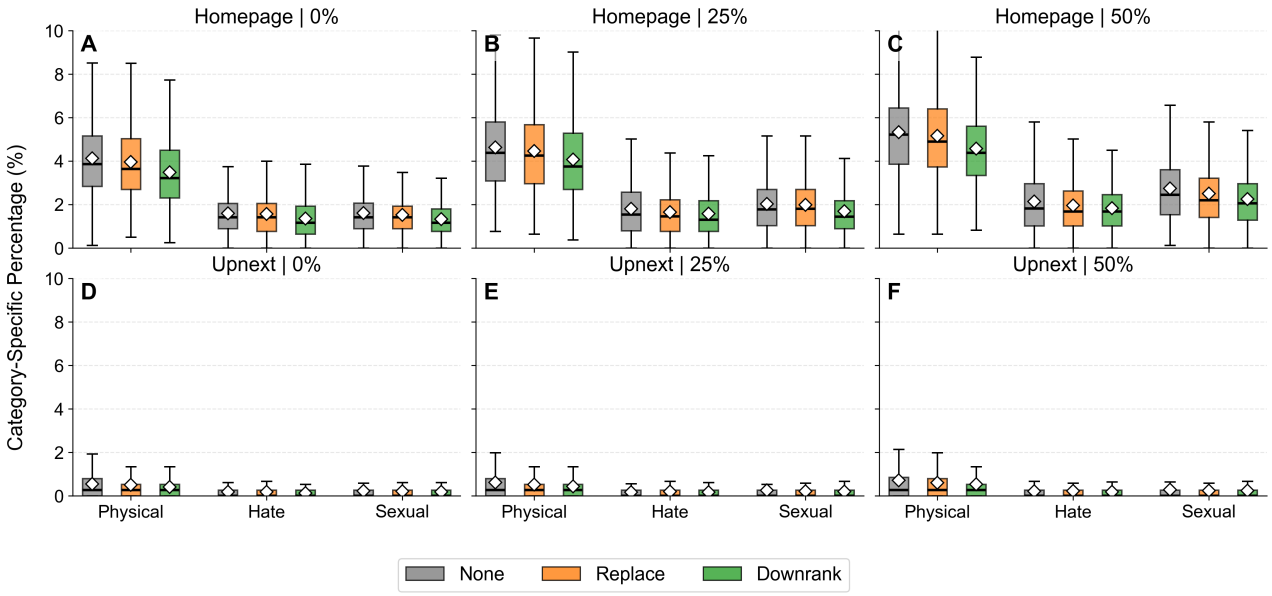

Figure 6: Cumulative harmful percentage for category-specific analysis over the 30-round intervention. Box plots show the median and interquartile range; while diamonds indicate means. The figure provides a complementary descriptive view of category-specific cumulative exposure across intervention conditions.

Table 6: Pre- and post-intervention net change ($\widehat{\Delta\mu}$) in harmful video percentage (pp) by focus, training harm level, and strategy. Negative values indicate harmful-content reductions. Δ 95% CI are bootstrap confidence intervals for $\widehat{\Delta\mu}$. Diff denotes for Net-change difference vs. baseline, which is the difference between a strategy's $\widehat{\Delta\mu}$ and the None baseline at the same training level (negative values favor the strategy), with bootstrap 95% confidence intervals. P values come from two-sided permutation tests, with Holm correction applied within each focus across intervention-versus-baseline comparisons. N is the number of puppets per condition. In this metric, Downranking on the Homepage produces larger, significant reductions across training levels, whereas Replacement shows smaller and non-significant reductions. On Up-Next, net changes are small and non-significant across all strategies.

Focus	Training	Strategy	$\widehat{\Delta\mu}$ (pp)	Δ 95% CI (pp)	N	Diff. (pp)	Diff. 95% CI	P
Homepage	0%	None	2.489	[2.016, 2.978]	1000			
		Replacement	2.251	[1.791, 2.705]	1000	-0.238	[-0.898, 0.437]	0.830
		Downranking	0.817	[0.373, 1.252]	1000	-1.672	[-2.320, -1.005]	0.001
	25%	None	0.601	[0.084, 1.111]	1000			
		Replacement	0.304	[-0.204, 0.793]	1000	-0.298	[-0.999, 0.423]	0.830
		Downranking	-0.896	[-1.389, -0.402]	1000	-1.497	[-2.208, -0.755]	0.001
	50%	None	-1.741	[-2.313, -1.159]	1000			
		Replacement	-2.249	[-2.845, -1.646]	1000	-0.508	[-1.325, 0.316]	0.663
		Downranking	-3.146	[-3.714, -2.576]	1000	-1.405	[-2.207, -0.585]	0.003
Up-Next	0%	None	-1.363	[-1.709, -1.021]	1000			
		Replacement	-1.411	[-1.803, -1.029]	1000	-0.048	[-0.539, 0.469]	1.000
		Downranking	-1.330	[-1.682, -1.012]	1000	0.033	[-0.436, 0.522]	1.000
	25%	None	-2.411	[-2.908, -1.930]	1000			
		Replacement	-1.901	[-2.317, -1.501]	1000	0.510	[-0.108, 1.149]	0.724
		Downranking	-2.183	[-2.562, -1.811]	1000	0.227	[-0.364, 0.859]	1.000
	50%	None	-3.112	[-3.593, -2.616]	1000			
		Replacement	-3.286	[-3.822, -2.768]	1000	-0.174	[-0.916, 0.547]	1.000
		Downranking	-3.239	[-3.745, -2.759]	1000	-0.126	[-0.821, 0.557]	1.000

Table 7: Direct two-sided comparisons between **Replacement** and **Downranking** for net change in harmful-video percentage. $\widehat{\Delta\mu}$ is the mean net change from the initial to the final round (pp), with bootstrap 95% confidence intervals. Diff. is Downranking minus Replacement in percentage points (negative values favor Downranking). p values are Holm-adjusted from two-sided permutation tests within each surface.

Focus	Training	Replacement $\widehat{\Delta\mu}$ (95% CI)	Downranking $\widehat{\Delta\mu}$ (95% CI)	Diff. (pp)	p_{Holm}	Sig.
Homepage	0%	2.25 [1.79, 2.70]	0.82 [0.37, 1.25]	-1.43	0.0003	***
Homepage	25%	0.30 [-0.20, 0.79]	-0.90 [-1.39, -0.40]	-1.20	0.0014	**
Homepage	50%	-2.25 [-2.84, -1.65]	-3.15 [-3.71, -2.58]	-0.90	0.0323	*
Up-Next	0%	-1.41 [-1.80, -1.03]	-1.33 [-1.68, -1.01]	0.08	1.0000	n.s.
Up-Next	25%	-1.90 [-2.32, -1.50]	-2.18 [-2.56, -1.81]	-0.28	0.9521	n.s.
Up-Next	50%	-3.29 [-3.82, -2.77]	-3.24 [-3.74, -2.76]	0.05	1.0000	n.s.

Table 8: Cumulative harmful-video percentage on **Homepage** by training and strategy (condensed). Means and 95% CIs are reported in %. Δ is in percentage points (pp). p values are Holm-adjusted from two-sided permutation tests comparing each intervention against the baseline within surface and training condition.

Training	Intervention	Baseline mean (95% CI)	Interv. mean (95% CI)	Δ vs. ctrl (pp)	Rel. change (%)	p_{Holm}	Sig.
0%	Replacement	7.33 [7.15, 7.52]	7.07 [6.90, 7.24]	-0.26	-3.55	3.8696e-02	*
	Downranking	7.33 [7.15, 7.52]	6.19 [6.02, 6.35]	-1.14	-15.56	2.9997e-04	***
25%	Replacement	8.48 [8.26, 8.70]	8.12 [7.91, 8.32]	-0.36	-4.25	1.6898e-02	*
	Downranking	8.48 [8.26, 8.70]	7.37 [7.17, 7.55]	-1.11	-13.09	2.9997e-04	***
50%	Replacement	10.23 [10.00, 10.46]	9.64 [9.43, 9.86]	-0.59	-5.76	3.9996e-04	***
	Downranking	10.23 [10.00, 10.46]	8.68 [8.47, 8.88]	-1.55	-15.21	2.9997e-04	***

Table 9: Cumulative harmful-video percentage on **Up-Next** by training and strategy (condensed). Means and 95% CIs are reported in %. Δ is in percentage points (pp). p values are Holm-adjusted from two-sided permutation tests comparing each intervention against the baseline within surface and training condition.

Training	Intervention	Baseline mean (95% CI)	Interv. mean (95% CI)	Δ vs. ctrl (pp)	Rel. change (%)	p_{Holm}	Sig.
0%	Replacement	0.96 [0.89, 1.03]	0.91 [0.84, 0.98]	-0.05	-5.21	3.5176e-01	n.s.
	Downranking	0.96 [0.89, 1.03]	0.75 [0.69, 0.80]	-0.21	-21.88	2.9997e-04	***
25%	Replacement	1.06 [0.97, 1.14]	0.94 [0.87, 1.01]	-0.12	-11.32	3.8296e-02	*
	Downranking	1.06 [0.97, 1.14]	0.80 [0.74, 0.87]	-0.26	-24.53	2.9997e-04	***
50%	Replacement	1.24 [1.14, 1.33]	1.08 [1.00, 1.16]	-0.16	-12.90	2.3998e-02	*
	Downranking	1.24 [1.14, 1.33]	0.97 [0.89, 1.06]	-0.27	-21.77	2.9997e-04	***

Table 10: Direct two-sided comparisons between **Replacement** and **Downranking** for cumulative harmful-video exposure. Means and 95% CIs are reported in %. Δ is Downranking minus Replacement in percentage points (negative values favor Downranking). p values are Holm-adjusted within each surface.

Surface	Training	Replacement mean (95% CI)	Downranking mean (95% CI)	Δ (pp)	p_{Holm}	Sig.
Homepage	0%	7.07 [6.90, 7.24]	6.19 [6.02, 6.35]	-0.88	3.0000e-04	***
Homepage	25%	8.12 [7.91, 8.32]	7.37 [7.17, 7.55]	-0.75	3.0000e-04	***
Homepage	50%	9.64 [9.43, 9.86]	8.68 [8.47, 8.88]	-0.96	3.0000e-04	***
Up-Next	0%	0.91 [0.84, 0.98]	0.75 [0.69, 0.80]	-0.17	1.2000e-03	**
Up-Next	25%	0.94 [0.87, 1.01]	0.80 [0.74, 0.87]	-0.14	7.1993e-03	**
Up-Next	50%	1.08 [1.00, 1.16]	0.97 [0.89, 1.06]	-0.11	5.8594e-02	n.s.