RESEARCH-ARTICLE

You Don't Bring Me Flowers: Mitigating Unwanted Recommendations Through Conformal Risk Control

GIOVANNI DE TONI, Bruno Kessler Foundation, Trento, TN, Italy

ERASMO PURIFICATO, European Commission Joint Research Centre, Brussels, Belgium

EMILIA GÓMEZ, European Commission Joint Research Centre, Brussels, Belgium

ANDREA PASSERINI, University of Trento, Trento, TN, Italy

BRUNO LEPRI, Bruno Kessler Foundation, Trento, TN, Italy

CRISTIAN CONSONNI, European Commission Joint Research Centre, Brussels, Belgium

Open Access Support provided by:

European Commission Joint Research Centre

Bruno Kessler Foundation

University of Trento

Published: 22 September 2025

Citation in BibTeX format

RecSys '25: Nineteenth ACM Conference on Recommender Systems
September 22 - 26, 2025
Prague, Czech Republic

Conference Sponsors:
SIGCHI

https://doi.org/10.1145/3705328.3748054
ISBN: 9798400713644

You Don't Bring Me Flowers: Mitigating Unwanted Recommendations Through Conformal Risk Control

Giovanni De Toni
Fondazione Bruno Kessler (FBK)
Trento, Italy
gdetoni@fbk.eu

Erasmo Purificato*
European Commission, Joint Research Centre (JRC)
Ispra, Italy
erasmo.purificato@acm.org

Emilia Gomez*
European Commission, Joint Research Centre (JRC)
Seville, Spain
emilia.gomez-gutierrez@ec.europa.eu

Andrea Passerini
University of Trento
Trento, Italy
andrea.passerini@unitn.it

Bruno Lepri
Fondazione Bruno Kessler (FBK)
Trento, Italy
lepri@fbk.eu

Cristian Consonni*
European Commission, Joint Research Centre (JRC)
Ispra, Italy
cristian.consonni@acm.org

Abstract

Recommenders are significantly shaping online information consumption. While effective at personalizing content, these systems increasingly face criticism for propagating irrelevant, unwanted, and even harmful recommendations. Such content degrades user satisfaction and contributes to significant societal issues, including misinformation, radicalization, and erosion of user trust. Although platforms offer mechanisms to mitigate exposure to undesired content, these mechanisms are often insufficiently effective and slow to adapt to users' feedback. This paper introduces an intuitive, model-agnostic, and distribution-free method that uses conformal risk control to provably bound unwanted content in personalized recommendations by leveraging simple binary feedback on items. We also address a limitation of traditional conformal risk control approaches, *i.e.*, the fact that the recommender can provide a smaller set of recommended items, by leveraging implicit feedback on consumed items to expand the recommendation set while ensuring robust risk mitigation. Our experimental evaluation on data coming from a popular online video-sharing platform demonstrates that our approach ensures an effective and controllable reduction of unwanted recommendations with minimal effort. The source code is available here: https://github.com/geektoni/mitigating-harm-recsys.

CCS Concepts

• **Information systems** → **Recommender systems**; **Personalization**; • **Social and professional topics**;

Keywords

Recommender systems, Unwanted recommendations, Responsible recommendations, Conformal risk control

ACM Reference Format:
Giovanni De Toni, Erasmo Purificato, Emilia Gomez, Andrea Passerini, Bruno Lepri, and Cristian Consonni. 2025. You Don't Bring Me Flowers: Mitigating Unwanted Recommendations Through Conformal Risk Control. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3705328.3748054

1 Introduction

The widespread use of recommender systems on online platforms has significantly changed how users engage with content. These systems are integral to modern digital media, facilitating personalized navigation through extensive online content across domains such as e-commerce [33, 42], social media [30, 49], and content streaming [23, 25], and tailoring experiences based on historical behaviors and expressed preferences [12, 28]. By leveraging user-generated data, recommenders aim to enhance user engagement [62], satisfaction [45], and platform profitability [16]. However, despite their proven utility and benefits, the basic concepts that make recommendation algorithms effective, such as *similarity-based filtering* and *preference prediction*, can inadvertently contribute to the spread of unwanted content [40]. Platforms have faced increasing criticism for creating *filter bubbles* and *rabbit holes* that reinforce existing beliefs, and for spreading potentially dangerous or harmful content [56]. Such dynamics are often attributed to algorithms that prioritize *engagement*, sometimes at the expense of content quality and truthfulness [57]. Features like the "*Not Interested*" button on YouTube [40] suffer from issues like low user awareness, complexity of use, or limited effectiveness. The lack of transparency and explainability in such strategies further exacerbates these problems, making it difficult for users to understand why they are being shown certain content and how to avoid it. As a result, users develop "*folk theories*" to explain the recommendations they are getting and try to influence the recommender outcomes [11, 26, 27, 40]. Recent works have studied how to mitigate unwanted content, biases, or

*Disclaimer: The view expressed in this paper is purely that of the authors and may not, under any circumstances, be regarded as an official position of the European Commission.

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748054

ensure fairness in recommendations [1, 2, 18, 44], but it remains unclear how to grant users precise and transparent control over the systems' *risk*.

In this paper, we propose a post-hoc method based on conformal prediction [4, 52] to provably mitigate the exposure to unwanted recommendations *in expectation*. Conformal prediction is a framework for uncertainty quantification for machine learning models that constructs *prediction sets* containing the true output with a *user-specified confidence level*, assuming data exchangeability. It uses *non-conformity scores* from model outputs to provide *distribution-free*, *finite-sample* guarantees. Conformal methods have recently been applied to decision support in classification tasks [22, 54, 55], natural language processing [17], and recommendation systems [7, 61]. However, in recommender systems, current approaches often overlook key challenges, such as re-recommending unwanted items or balancing personalization with content filtering. To address this, we apply *conformal risk control* [6], which generalizes conformal prediction to control *general loss functions* (or "*risk*") rather than coverage. Our method builds recommendation sets by *filtering* potentially unsafe items and *replacing* them with previously consumed, *safe* content, while provably respecting a user-defined risk level. Specifically, our contributions include:

(1) A comprehensive analysis of real data to understand *reporting patterns* and watch time behavior related to unwanted content, providing insights into how users interact with and respond to such content.
(2) A method based on conformal risk control that provably limits unwanted content in recommendations, with theoretical guarantees, by leveraging *repeated content*.
(3) A practical heuristic to choose *safe* alternative content, balancing the need for personalization and the goal of minimizing unwanted recommendations.
(4) Experiments and ablations with a real-world dataset, highlighting our approach's effectiveness in significantly reducing unwanted content.

2 Related Work

Recommender systems have faced growing criticism for promoting unwanted content such as misinformation [28], hate speech [8], and offensive or irrelevant items [2, 63]. These issues often stem from algorithmic bias [9, 13], engagement-driven objectives [34], and feedback loops that reinforce filter bubbles [20, 58]. Further, repeated exposure to undesirable recommendations impacts user trust and satisfaction [65], contributing to what researchers describe as "algorithmic hate" [53]. Unfortunately, efforts to empower users, such as feedback tools, are often hindered by low visibility and unclear system responses [11, 40].

To address these challenges, prior work has proposed content moderation [50], fairness constraints [66], and diversity-promoting techniques [53, 57]. While effective to a degree, these methods often lack formal guarantees on limiting exposure to harmful content and seldom incorporate fine-grained user behavior [32, 36]. Our approach is closely related to [18, 44] but differs in key ways: we provide item-level control over unwanted content exposure – beyond group-level guarantees [44] – and avoid the need to learn custom policies from user feedback [18]. Recent work has introduced conformal methods to recommender systems [7, 61], focusing on controlling metrics like Normalized Discounted Cumulative Gain (nDCG) or False Discovery Rate (FDR). However, to the best of our knowledge, no prior method has used conformal techniques to control the fraction of unwanted content in recommendations.

Lastly, repeated content consumption is a well-established behavior across platforms [3, 48]. Recent approaches have leveraged this phenomenon through repeat-explore strategies [48], neural ODEs for temporal dynamics [21], and collaborative filtering models that capture item-specific repeat patterns with varying lifetimes [59]. We exploit this phenomenon to offer formal guarantees on risk control while preserving performance.

Table 1: Statistics on reported videos.

H_X (%)	N. of users	% of users	eCDF
$0 \leq H < 0.1$	4569	60.1	0.601
$0.1 \leq H < 0.3$	1909	25.1	0.852
$0.3 \leq H < 1$	827	10.9	0.961
$1 \leq H < 100$	296	3.9	1.0

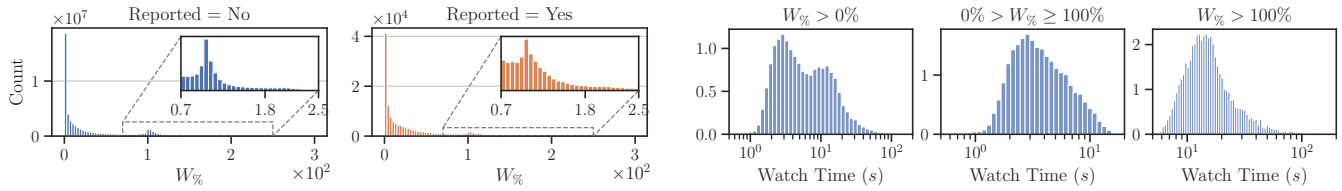
3 Case Study: A Video-Sharing Platform

We begin our study by examining data from a popular Chinese short-form video-sharing platform, *Kuaishou*[1], which has over 300 million daily users at the time of writing. We use the *KuaiRand* dataset [29], which was collected by sampling 27000 Kuaishou users who collectively watched approximately 32 million videos over one month (from April 8 to May 8, 2022). This dataset contains various user interaction signals, such as watch time for each video and whether users liked or commented on a video. Unfortunately, the dataset does not include Kuaishou's ranking scores or the explicit recommendation order of videos. However, the dataset also includes *negative feedback* signals, such as when users click the *"Do not recommend"* or *"Report"* buttons while watching a video. Similar negative feedback data is available in the *YouTube Regrets* dataset[2], but only in an aggregated format without user profiles, limiting the scope of analysis. We focus exclusively on short videos, removing advertisements and videos with zero duration (*e.g.*, dynamic photos). Notably, approximately 29% of users provided negative feedback on at least one video during the data collection period. Consequently, our analysis is restricted to this subset. After further filtering, our dataset consists of 7601 users, 56 million interactions, and 10 million unique videos. On average, a user watched 7, 417 videos during the collection period (median: 4, 623), with a large standard deviation of 9, 223.

3.1 Sparsity of Negative Feedback

We start by considering the propensity of users to give negative feedback to a recommended video seen within the interface. Thus, we consider the percentage of videos a user reported overall videos they have seen $H_X = \frac{\#\text{ of videos reported by user } U}{\#\text{ of videos seen by user } U} \cdot 100$. Table 1 shows that 60% of users report at most 1 every 1000 videos they have seen. Overall, the negative feedback is very sparse since > 95%

[1] https://en.wikipedia.org/wiki/Kuaishou
[2] https://huggingface.co/datasets/mozilla-foundation/youtube_regrets



(a) Distribution of $W_{\%}$ on reported (right) and unreported (left) videos. **(b) Reported videos watch time segmented by $W_{\%}$ (log-scale).**

Figure 1: Statistics on the user behaviour on Kuaishou data.

of users only report at most 1 every 100 videos recommended to them. The relatively low reporting rate may be attributed to the accessibility of the *"Do not recommend"* and *"Report"* options, which require multiple interactions within the interface. Indeed, many users might simply skip a video they do not like. Nevertheless, the data sparsity represents a challenge for recommender systems that want to exploit this information to minimize recommendations of unwanted content.

3.2 Analysis of the Watch Time

Next, we analyze user watch time behavior to investigate potential differences in the viewing patterns of reported videos. We define with $W \in \mathbb{R}_{\geq 0}$ the *duration of a user-video interaction* (*e.g.*, how much time the user U spent on that video I). Since video durations vary, we measure each user-video interaction as the proportion of the video watched by the user:

$$W_{\%}(U, I) = \frac{W(U, I) \coloneqq \text{View time by user U of video I}}{\text{Duration of video } I} \cdot 100 \quad (1)$$

Figure 1a shows the distribution of $W_{\%}$ for both reported and unreported videos. Both distributions peak near zero, indicating that many videos are either skipped or watched very briefly. We observe that 21% of interactions result in no watch time. There is also a peak near one, as highlighted, suggesting a different behavior closer to videos viewed in full. Additionally, some user-video interactions exceed the video's original duration, indicating users occasionally linger on videos. Similarly to prior research [64], we focus on reported videos and group them based on their duration. In our analysis, we consider videos with durations between 7 and 12 seconds, which are the most common. Since the distribution is zero-inflated (many videos are not watched), we analyze only interactions with positive watch time ($W > 0$). Fig. 1b illustrates the distribution of watch time (log scale). As seen in the left-most plot, there are peaks around zero and near 10 seconds. Consequently, we model two scenarios: (a) interactions where the video is watched at most in full ($0\% < W_{\%} \leq 100\%$), and (b) interactions where the video is watched for longer than its duration ($W_{\%} > 100\%$). Fig. 1b's centre and right-most plots show both distributions have approximately normal shapes with left and right tails, respectively. Such behavior could be modeled by considering the LogNormal or Weibull distributions, which are commonly used for describing dwell time, *e.g.*, on the page views [41, 64].

3.3 Trend of Video Reporting Patterns

We also analyze user behavior when rewatching previously viewed videos. On average, 2.6% of videos watched by users had been seen before. Furthermore, as shown by Table 2, 8% of users are shown videos they previously reported, and another 8% report previously watched videos that they had not reported initially. As argued in Section 3.2, the latter behavior suggests that users may have skipped the video on the first encounter or found the reporting process cumbersome (*e.g.*, requiring multiple clicks). This hypothesis is supported by Table 2 that shows that the 75% percentile of the watch time for previously seen videos reported anew (2nd row) resembles the 75% percentile of initially reported videos (4th row). Interestingly, reported videos that get flagged a second time show a very similar watch time (3rd row). We postulate this is stronger feedback indicating items that the user strongly considers unwanted. These findings may be influenced by the platform's continuous stream of new content, which reduces the likelihood of users encountering the same video multiple times. Nevertheless, 16% of users either see or report again at least one previously seen video. Lastly, Table 2 shows how videos seen a second time have a greater average watch time. For videos reported anew the second time (2nd row), the watch time increases significantly, suggesting that users may deliberately review undesirable content more thoroughly before reporting.

Summary of the analysis. We conclude the analysis of Kuaishou data by underlining some key insights that will be useful in designing harm-controlling recommendation systems:

- The engagement and perceived harmfulness of videos are not fully correlated. As shown by Fig. 1a, reported videos have the same $W_{\%}$ distribution as non-reported videos. Thus, *maximizing engagement does not imply providing less harmful videos to users.*
- In Kuaishou, the system suggests undesired content *despite explicit user feedback* (cf. Table 2);
- Given that over 95% of users report less than 1% of videos they watch, recommender systems should maximize the usefulness of this little explicit negative feedback.
- With 21% of interactions resulting in no watch time and a strong peak near zero in view time distribution (cf. Section 3.2), very short engagement can serve as an early indicator of harmful or undesired content.

Table 2: User's behavior with repeated videos. Given a video I, the user can choose to report it (✗) or not (✓). We show the average watch time (1st time and 2nd time the user sees it), the standard deviation, and the third quartile in brackets.

Behaviour	I (%)	U (%)	Watch Time 1st (s)	Watch Time 2nd (s)
✓ → ✓	99.79	100.0	2.71 ±12.15 (1.39)	10.48 ±25.85 (7.95)
✓ → ✗	0.11	8.69	0.50 ±7.41 (0.0)	10.73 ±20.85 (10.13)
✗ → ✗	0.07	6.33	7.74 ±15.83 (6.64)	7.85 ±15.86 (6.72)
✗ → ✓	0.03	1.49	1.36 ±6.91 (0.0)	6.46 ±25.42 (2.5)

3.4 Unwanted, Disliked and Harmful Content

Given the results of Section 3, we want to emphasize how understanding user feedback is fundamental for the task at hand. Indeed, online platforms provide various mechanisms for users to express feedback on recommended items. Recent research shows that mobile app affordances lead to "*triggered essential reviewing*" [47], where users provide quick, focused, and often emotionally charged feedback immediately after an experience. In this context, it is crucial to recognize that users can express feedback through multiple channels, each carrying distinct and sometimes diverging meanings. For instance, *flagging an item as unwanted* and *reporting it as harmful* are typically separate actions within a service interface. Moreover, to effectively evaluate mitigation techniques in realistic settings, it is essential to consider feedback signals that distinguish between *unwanted* or *harmful items* and those that simply fail to meet user expectations. For example, our analysis of the KuaiRand dataset reveals that videos with high watch times (suggesting positive engagement) may still be explicitly flagged as unwanted (*c.f.*, Table 2). However, many studies [18, 19, 51] assume that a rating below a certain threshold (*e.g.*, below 3 on a 1–5 scale) implies the item is either unwanted or harmful. In other datasets, such as the Music Streaming Sessions Dataset (MSSD) [14], item skips indicate unwanted content for a user. These proxies, as seen, can be insufficient. To the best of our knowledge, only the KuaiRand dataset [29] provides such explicit, disaggregated feedback, despite the widespread availability of these mechanisms on modern platforms.

4 Recommender Systems With Risk Control

Given the key insights obtained from the analysis of Kuaishou data, let us now define more formally a *risk-controlling recommender system*. Consider a finite set of items $\mathcal{I} \in \{i_1, \dots, i_N\}$ and a finite set of users $\mathcal{U} \in \{u_1, \dots, u_M\}$, each represented by d-dimensional feature vectors, respectively. Our recommender system outputs a set of items $S(U) \subseteq \mathcal{I}$ based on some relevance score estimated by a ranker $f(U, I) \in \mathbb{R}^+$ predicting how much an item I is relevant for a user U based on the information within U and I (*e.g.*, in terms of watch-time). In two-stage recommender systems, we might perform further *post-processing* activities to *filter* certain items, before presenting the final recommendation to the user. Similarly to Angelopoulos et al. [7], let us consider a post-processing step where we employ some threshold $\lambda \in \mathbb{R}$ to pick only certain items, based on a score $s : U \times I \to \mathbb{R}$. For example, consider $s(U, I = i) = f(U, I = i)$, we might want to keep only items with a predicted relevance above a certain base value:

$$T_\lambda(U) = \{i \in \mathcal{I} : s(U, I = i) \geq \lambda\} \tag{2}$$

After post-processing, our objective becomes providing the best set $S_\lambda(U) \subseteq T_\lambda(U)$ maximizing a target metric computed on the recommendations (*e.g.*, *serendipity* [31]). Generally, we are interested in recommending the *most* relevant items to each user, under a budget $k \leq N$ *e.g.*, the number of slots in a YouTube main page. Let us define with $\pi(\mathcal{I}) = \{i_1, \dots, i_N\}$ the order induced by sorting the items given the ranker scores $f(U, I = i_j)$. If our ranker is reasonably good, then the optimal solution would be simply picking up to the k-th item from $\pi(T_\lambda(U))$:

$$S_\lambda(U, k) = \{i_j \in \pi(T_\lambda(U)) : j \leq k\} \tag{3}$$

In classical learning-to-rank tasks, we try to learn the optimal ranker $f^* = \text{argmax}_{f \in \mathcal{F}} \mathbb{E}[\ell(S_\lambda(U, k))]$ where $\ell : \mathcal{I} \to \mathbb{R}$ is our target metric, such as the nDCG. Besides picking the most relevant items for the user, now we want to *control* the *risk* of the set $\mathcal{S}_\lambda(X, k)$ induced by the optimal ranker f^*. Let us assume we have a risk metric $R : \mathcal{I} \to (0, 1)$ we would like to provably *bound* below a certain value $\alpha \in (0, 1)$. For example, we might want the *fraction* of unwanted content in $S_\lambda(U, k)$ to be below a user-defined value. Since we want to avoid performing costly operations such as retraining the ranker f^*, an intuitive way to reach such an objective is by acting on the *post-processing* step, by finding the optimal λ threshold ensuring our *risk* is below a user-defined level $\alpha \in (0, 1)$:

$$\lambda^* = \underset{\lambda \in \Lambda}{\text{argmax}}\, \mathbb{E}\left[\ell(S_\lambda(U, k))\right] \qquad \text{s.t.} \quad \mathbb{E}[R(S_\lambda(U, k))] \leq \alpha \tag{4}$$

In the next sections, we will show how to provably bound the risk in Eq. (4) for *any* ranker by leveraging the user's negative feedback.

4.1 Measuring Unwanted Content

In our setting, we consider an item *unwanted* for a user if she expresses distaste for it once seen within the recommendations. Users can express their distaste for certain recommendations by giving either a negative score or binary feedback (YouTube or Kuaishou *"Don't recommend"* button [43]). Similarly to [24], we focus on the latter, and we consider a setting in which we can observe only binary negative feedback, as it happens for Kuaishou data (cf. Section 3). Let us denote with $H(U, I) \in \{0, 1\}$ a binary random variable indicating if an item I has been flagged ($H = 1$) by the user U. Given a set of recommendations $S_\lambda(U, k)$, we denote its overall *risk* as the fraction of items the user has flagged once presented with it:

$$R_H(S_\lambda(U, k)) = \frac{|\{i \in S_\lambda(U, k) : H(U, I = i) = 1\}|}{|S_\lambda(U, k)|} \tag{5}$$

4.2 Conformal Risk Control

By simply plugging the previous risk definition in Eq. (4), we have our optimization objective. However, how do we pick the threshold λ to control the risk defined by Eq. (5)? We can find the optimal threshold via *conformal risk control* [6] by employing a held-out calibration set $\{(u, i, h))_j\}_{i=1}^n$, where for each user-item pairs we record also if it was flagged as *unwanted* ($h = 1$). We first state the main theorem from [6], and then we explain its practical implications.

Theorem 4.1. *Assume that $R_H(S_\lambda(X, k))$ is non-increasing in λ, right-continuous, $R_H(S_{\lambda_{max}}(U, k)) \leq \alpha$ and $\sup_\lambda R_H(S_\lambda(U, k)) \leq$*

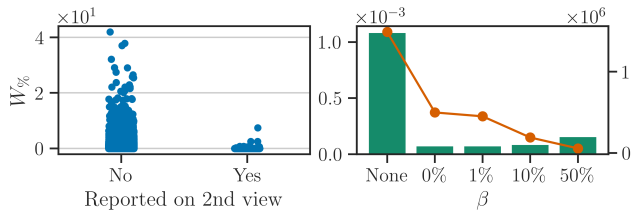


Figure 2: Validity of Property 1 in Kuaishou. (Left) Frequency of reported videos seen the second time based on their first watch time $W_{\%}$. (Right) Empirical $P(H = 1 \mid H_{1st} = 0, W_{\%} > \beta)$ with varying β (green) and total remaining repeated videos after thresholding over the watch-time (orange).

$1 < \infty$ *almost surely. Consider* $\hat{R}(\lambda) = \frac{1}{n}\sum_{j=1}^{n} R_H(S_\lambda(U = u_j, k))$, *given a held-out calibration set. Given any desired risk upper bound* $\alpha \in [0, 1]$, *if we choose a threshold as:*

$$\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\hat{R}(\lambda) + \frac{1}{n+1} \leq \alpha\right\} \tag{6}$$

then, we have $\mathbb{E}[R(S_{\hat{\lambda}}(U, k))] \leq \alpha$.

In practice, as long as our risk function decreases as λ increases, by choosing $\hat{\lambda}$ as in Theorem 4.1 we are guaranteed the risk will be *provably* bound in expectation. Practically speaking, if we consider as score the ranker output $s(U, I) = f(U, I)$, as we increase the threshold, fewer items will be contained within $T_\lambda(U)$. Thus, the likelihood of picking harmful items within $S_\lambda(U, k)$ decreases monotonically since by removing items, we cannot increase $R_H(S_\lambda(U, k))$. Therefore, we can enjoy the risk-control guarantees provided by Theorem 4.1[3]. More importantly, the guarantees offered by Theorem 4.1 are model-agnostic and distribution-free. Therefore, they can be readily applied to *any trained recommender*, given historical interaction data.

5 Risk Control via Item Replacement

Recent applications of conformal prediction to recommender systems consider only the simple filtering approach as detailed in Eq. (2) [7, 39, 61]. The approach ensures the monotonicity requirements of Theorem 4.1 but at the price of losing control of the size of the resulting set $S_\lambda(U, k)$. First, notice that the sets are nested by λ since $\lambda \geq \lambda' \Rightarrow T_{\lambda'}(U) \subseteq T_\lambda(U)$. If we use the ranker outputs as a scoring function, we will be able to influence the content of $S_\lambda(U, k)$ only when the threshold is greater than the score $f(U, I)$ of the k-th item in $\pi(I)$. Hence, in order to reduce the risk within $S_\lambda(U, k)$, we will have to choose a threshold λ for which we will necessarily end up with fewer items than k. Indeed, we can state the following proposition:

Proposition 5.1. *Given a risk* $\alpha \in [0, 1]$ *and* $s(U, I) = f(U, I)$, *there exist data distributions for which a thresholding rule, as in Eq. (2), cannot provide the set with* $H(S_\lambda(U, k)) \leq \alpha$ *unless* $|T_\lambda(U)| = 0$.

[3]Our metric does not take into consideration the *order* in which we present the item to the user, which might make $R_H(S_\lambda(U, k))$ non-monotone. In practice, we can monotonize any risk by taking $R_H(S_\lambda(U, k)) = \max_{\lambda' \geq \lambda} S_{\lambda'}(U, k)$ [6].

Proof. Consider the items $\mathcal{I} = \{A, B, C, D\}$ where the ranker scores are $f(u, D) = 5$ and $f(u, i) = 1$ for $i \in \{A, B, C\}$. Thus, the ranker induces the following ordering $\pi(I) = \{D, A, B, C\}$ where we broke ties at random. Further, let us pick as a risk level $\alpha = 0.1$. Then, it is easy to see how, for any $k \in [1, 4]$, only with $\lambda > 5$ we have $H(S_\lambda(U, k) \leq 0.1$. Unfortunately, any threshold where $\lambda > 5$ implies $T_\lambda(I) = \varnothing$. □

Practically, it means that to provably control the associated risk of a recommendation, we might have to remove too many items from the recommendation pool. Therefore, in our recommendation, we might not be able to return k items to the user. Proposition 5.1 applies to *any* score function. Consequently, this strategy is not optimal since, for example, in the case of a video streaming platform, we are essentially "wasting" items, thus potentially reducing engagement.

5.1 A Property for Safe Alternatives

Rather than simply removing items, we take an alternative approach, which consists of *replacing* potentially unwanted items with safer ones, thus sidestepping the effects of Proposition 5.1. However, how do we choose these *safe items*? Our approach exploits a simple fact: *users consume again previously seen items* [3, 21, 48]. Furthermore, as seen in Kuaishou data (Table 2), users often dwell longer on already-consumed items, leading to greater engagement. Inspired by the analysis in Section 3, if we consider items the user has seen before, *that were not reported as unwanted*, we might as well suggest them again in place of potentially unsafe items. As shown in Table 2, there exist *rare* items (0.11%) which are reported only the second time a user watches them. This poses a challenge since we might lose the monotonicity guarantees needed by Theorem 4.1. Therefore, we need a sensible way to select previously seen videos for which we are reasonably sure the user will not report anew. We formalize this desideratum in a simple property that can be tested by employing historical interaction patterns:

Property 1. *Denote with* $H_{1st} \in \{0, 1\}$ *the event that the user has reported an item the first time they saw it. There exists a variable* C *and a value* $\beta \in supp(C)$ *such that the probability of an item being flagged a second time is zero* $P(H = 1 \mid H_{1st} = 0, C > \beta) = 0$.

In the case of Kuaishou, we can pick as C the watch time of videos. Then, we can simply pick videos on which the users have spent time $W_{\%}(U, I)$ above a certain threshold. For example, Fig. 2 shows the distribution of $P(H = 1 \mid H_{1st} = 0, W_{\%})$. Indeed, the plot on the left shows that the videos that are reported in the second view have a much shorter watch time. Moreover, the right plot of Fig. 2 shows that by filtering repeated videos with a thresholding β, we can reduce the likelihood of picking a harmful video almost to zero, globally. For $\beta = 50\%$, the probability slightly increases, since we have fewer repeated videos, and mostly because of a few outliers (*e.g.*, reported videos with a very high watch time). Practically, we provide an ablation study in Section 7 showing the effect of the choice of β on risk control guarantees.

Thus, if Property 1 holds, we can simply replace items with a score below the risk control threshold, with items the user *has previously* seen and not flagged as harmful, for which we have

Algorithm 1 Risk-controlling pipeline for a recommender system

Require: u, user, $\mathcal{I}$, items, α, risk level, $f : U \times I \to \mathbb{R}$, ranker, $s : U \times I \to \mathbb{R}^+$, score function, k recommendation size

1: $\hat{\lambda} \leftarrow \textsc{RiskControl}(\alpha)$ {Theorem 4.1}
2: $T_{\hat{\lambda}}(U) \leftarrow \{y \in Y : s(U, I = i) \geq \hat{\lambda}\}$
3: $T^{(\text{safe})} \leftarrow \{y' \in Y_X : H(U, I = i') = 0 \wedge \ W_{\%}(U, i') > \beta\}$
4: $T^{(\text{replace})}_{\hat{\lambda}}(U) \leftarrow T_{\hat{\lambda}}(U) \cup T^{(\text{safe})}(U)$
5: **return** $\{i_j \in \pi(T^{(\text{replace})}_{\hat{\lambda}}(U)) : j \leq k\}$

$C > \beta$. Let us denote the pool of *safe* videos with the following:

$$T^{(\text{safe})} = \{i' \in \mathcal{I}_U : H(U, I = i') = 0 \wedge \ C(I = i') > \beta\} \quad (7)$$

where $\mathcal{I}_U \subseteq \mathcal{I}$ indicates the items previously shown to the user U. Here, we are assuming that the set $\mathcal{I}_U$ is not empty, and more importantly, that there exist videos $i \in \mathcal{I}_U$ such that $C(I = i) > \beta$. A new user might not satisfy the conditions above, akin to *cold start* issues [46]. In such a scenario, we might want to *ask* the user to choose those videos for us. Thus, the new pool of items will become:

$$T^{(\text{replace})}_{\lambda}(U) = T_{\lambda}(U) \cup T^{(\text{safe})}(U) \quad (8)$$

It is straightforward to show that the new set $T^{(\text{replace})}_{\lambda}(U)$ satisfies the monotonicity requirements of conformal risk control, and it ensures we can always pick k items to show to the user.

Proposition 5.2. *Given any $\lambda, \lambda' \in \Lambda$ such that $\lambda < \lambda'$. Consider the set constructed with Eq.(8). Then, if Property 1 holds, we have the following $R_H\left(T^{(replace)}_{\lambda'}(U)\right) \leq R_H\left(T^{(replace)}_{\lambda}(U)\right)$.*

We omit the proof since it readily follows from the fact that any repeated item has $H(U, I = i) = 0$ if Property 1 holds.

6 A Simple Post-Hoc Pipeline

We consolidate the previous findings into a simple post-hoc algorithm applicable to *any* pre-trained recommender system – requiring *no retraining* and the simple conditions outlined in Theorem 4.1. Algorithm 1 presents the pseudo-code of the procedure. Given a user-defined risk level $\alpha \in (0, 1)$, the algorithm provably controls the fraction of unwanted content in the final recommendation list (*c.f.*, Eq.(5)). For a given user u and risk level α, we first compute the threshold $\hat{\lambda}$ using the RiskControl procedure from Theorem 4.1 (line 1). We then remove from the item pool all candidates with scores below $\hat{\lambda}$ (line 2) and identify potential replacements (line 3)—previously seen, non-flagged videos with watch-time exceeding a threshold β (*c.f.*, Property 1). These replacements are added to the remaining item pool (line 4), and the top-k items are selected using the ranker (line 5). Note that the final recommendation set may contain fewer than k items if no suitable replacements are available (*i.e.*, if $T^{(\text{safe})}$ is empty). Indeed, Algorithm 1 reduces to classical risk control via removal only if we *discard* line 3, and set $T^{(\text{safe})} = \{\}$. Lastly, the scoring function used for filtering may differ from the one used for reordering. For instance, risk control guarantees would still hold in a two-stage system where the score is the probability of *not reporting* an item, and the ranker predicts watch time.

Computational complexity. The threshold $\hat{\lambda}$ can be precomputed for any risk level α and user U with a complexity of $O(Q)$, where Q is the size of the calibration set. RiskControl (line 1) can be run only once and can be cached for efficient retrieval at inference time. Constructing the filtered set $T_{\hat{\lambda}}(U)$ and the replacement set $T^{(\text{safe})}(U)$ (lines 2–3) requires a single pass over the item pool, with complexity $O(|I|)$. Re-ranking and selecting the top-k items (line 5) has a complexity of $O(|I| \log |I|)$. Thus, the overall complexity of Algorithm 1 is dominated by the sorting step, resulting in $O(|I| \log |I|)$. Algorithm 1 is compatible with standard optimization techniques used in recommender systems (*e.g.*, efficient indexing, caching, or approximate nearest neighbours for lines 2–4). Naturally, the threshold $\hat{\lambda}$ should be periodically updated as new interaction data becomes available.

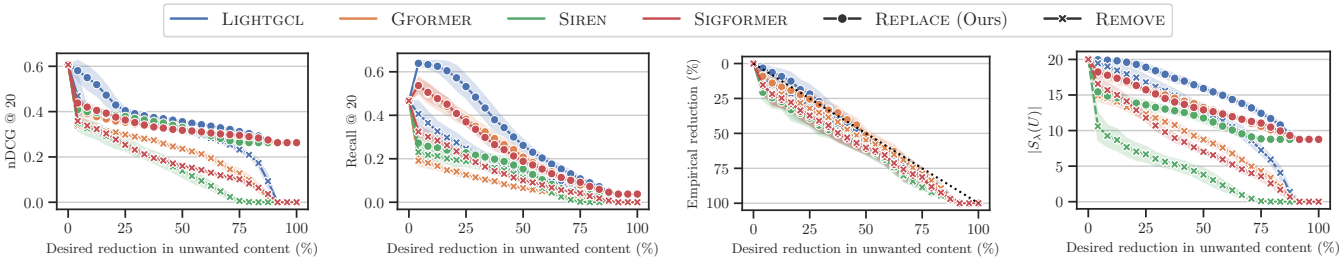
7 Experiments With Real Data

In this section, we evaluate the effects of the risk control strategies (Eqs. (2) and (8)) on a simple two-stage recommender setup for a realistic video recommendation task. In the following experiments, we aim to answer the following research questions:

- ***RQ1*** What is the impact of the conformal risk control on standard evaluation metrics (*e.g.*, *Recall@k* and *nDCG@k*)?
- ***RQ2*** What are the effects of employing different strategies for risk control (Eq. (2) and Eq. (8))?
- ***RQ3*** How many items do we have to replace from $S_{\hat{\lambda}}(U, k)$ to achieve the desired risk level?
- ***RQ4*** How does approximately satisfying Property 1 impact the desired unwanted content reduction?
- ***RQ5*** If we divide the users according to their reporting habits in *low-reporting* and *high-reporting*, is Algorithm 1 reducing the fraction of unwanted content also for *high-reporting* users?

Dataset. We use the KuaiRand dataset [29], which contains both positive and negative user feedback, along with information about interactions with previously seen videos. As discussed in Section 3.4, to the best of our knowledge, it is the only publicly available dataset that includes realistic feedback signals and a target metric such as video watch time. For our experiments, we focus on videos viewed at least twice by at least one user, resulting in a dataset that includes both repeated and single interactions. This leaves us with 5657 unique users, 117695 unique items, and more than 3 million interactions. Then, we first partition the full interaction set into two datasets based on whether the interaction is repeated or not. We adopt a 10-core setting, randomly splitting the subset with only single interactions into training (70%), validation (15%), and test (15%) sets. The validation set also serves as a calibration set for determining the threshold λ as defined in Theorem 4.1. To ensure evaluation fidelity, we exclude all test set videos from the repeated interaction set. Thus, the repeated interaction set only considers videos previously seen by users — *i.e.*, those appearing in either training or validation.

Metrics. We compare the effect of the risk control and various strategies on two widely used metrics: *Recall@k* and *nDCG@k*. In our experiments, we set $k = 20$ as usually done in the literature [19]. We also measure the empirical *risk*, *i.e.*, the fraction of unwanted content in the recommendations, as defined in Eq. (5).



(a) nDGC @ 20 **(b) Recall @ 20** **(c) Reduction in $R_H(S_\lambda(U,k))$** **(d) Size of $S_\lambda(U)$**

Figure 3: Results over *KuaiRand* ($k = 20$ and $\beta = 0$). The results and standard deviation (shaded areas) are over 10 runs.

Models. Similarly to previous studies [60], we evaluate the effectiveness of risk control using a two-stage recommender architecture: (a) the first stage filters items using the conformal risk control procedure described in Eq.(8); (b) the second stage re-ranks the filtered items using a model trained to predict watch-time $W(U,Y)_{\%}$. We consider two risk-control strategies: Remove (Algorithm 1 without line 3 and $T^{(\text{safe})} = \{\}$) and Replace (Algorithm 1). Remove is the classical risk-control strategy for recommender systems [7, 61]. For simplicity, under the Replace strategy, repeated items are not re-scored; instead, their original watch time is reused. To evaluate the impact of different scoring functions $s(U,Y)$ on the effectiveness of risk control, we experiment with several state-of-the-art methods. These include both sign-aware and unsigned graph-based models, depending on whether they incorporate negative user feedback (*e.g.*, reported videos): **LightGCL** [15] and **GFormer** [38], state-of-the-art graph-based methods leveraging contrastive learning and the transformer architecture, respectively. **SiReN** [51], the classic sign-aware method for recommendation learning two embeddings from positive and negative interaction graphs, and **SIGFormer** [19], a state-of-the-art sign-aware method employing the transformer architecture. For the re-ranking stage, we adopt a simple Neural Collaborative Filtering (NCF) model [35], as our focus is not on maximizing performance on the Kuaishou dataset, but rather on isolating the effects of risk control. Model parameters were selected according to the original papers and implemented using the authors' source code. Our full implementation is openly available[4] to support reproducibility and further research.

(RQ1 & RQ2) Replacing unwanted items ensures risk control and mitigates performance degradation. We start by evaluating the impact of risk control on standard ranking metrics, *i.e.*, nDCG@20 and Recall@20. As shown in Fig. 3c, the level of harmfulness is tightly controlled across all strategies and ranking models. For any target reduction in unwanted content (α in Algorithm 1), the empirical reduction on the test set is equal to or better (*i.e.*, all points lie on or below the diagonal) than the desired target, confirming that Algorithm 1 reliably enforces risk control. In general, the Remove strategy tends to exceed the target reduction, removing more items than strictly necessary. This proves that Algorithm 1 provides a provable link between user actions (*e.g.*, flagging content) and the resulting changes in recommendations. Indeed, by specifying a desired reduction level, users can provably influence their experience with guarantees on the expected reduction in unwanted content. Figs. 3a and 3b show how risk control affects nDCG@20 and Recall@20. Removing items sharply degrades nDCG as the desired reduction level (α in Eq. (4)) increases, due to the shrinking pool of candidate items (Eq. (2)). In contrast, the Replace strategy retains a higher nDCG, though some degradation still occurs as previously seen items are reintroduced into the ranking. This performance drop is partly explained by our experimental setup: we do not re-score repeated items but instead reuse their original watch time, which may not reflect user interest upon second exposure. Recall decreases under both strategies, as removing or replacing items may discard potentially relevant, unseen content. Finally, Fig. 3d illustrates the consequences of Proposition 5.1. When full removal of unwanted content is required ($\alpha = 0$), the sole option is to eliminate or replace all items in the recommendation set. Moreover, for the Replace strategy, we might recommend less than k items at test time, since we have finite replacement options.

(RQ3) The choice of score function impacts the number of previously seen items in the recommendation set. We examine the average fraction of previously seen items recommended to users. While the risk control guarantees of Theorem 4.1 hold independently of the underlying ranking function, in practice, we favor models that minimize item replacements at an equal level of risk control. Figure 4 shows that sign-aware models (SiReN and SIGFormer) lead to more replaced items compared to their unsigned counterparts (LightGCL and GFormer) for equivalent reductions in unwanted content. This is somewhat unexpected, as

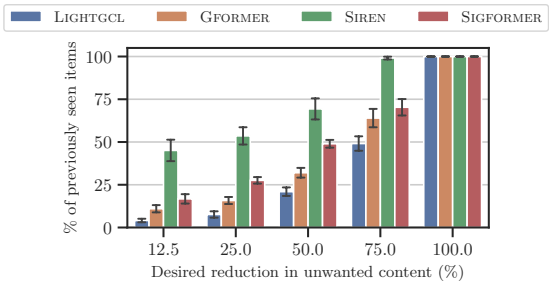
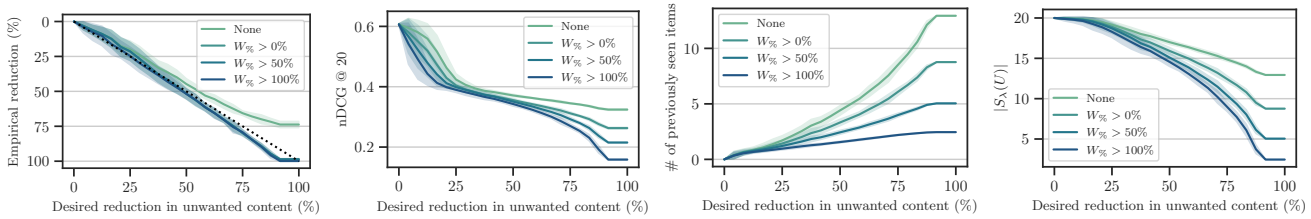


Figure 4: Fraction of previously seen items in the top-20 recommendations for Kuaishou ($k = 20$ and $\beta = 0\%$).

[4] https://github.com/geektoni/mitigating-harm-recsys



(a) Reduction in $R_H(S_\lambda(U,k))$ **(b) nDCG @ 20** **(c) # of previously seen items** **(d) Size of $S_\lambda(U)$**

Figure 5: Ablation analysis on β for LightGCL ($k = 20$). We consider only previously seen videos with a watch-time greater than $W_\% > \{0\%, 50\%, 100\%\}$, or without filtering (None). The results and standard deviation (shaded areas) are computed over 10 runs.

sign-aware models incorporate both positive and negative feedback, whereas the unsigned models do not. However, due to the low ratio of negative to positive feedback in the Kuaishou dataset, the additional information appears insufficient to improve ranking. For example, SiReN may struggle to learn meaningful embeddings from the highly sparse negative interaction graphs. As a result, sign-aware models exhibit higher uncertainty and require replacing or removing more items to meet risk control guarantees. In contrast, LightGCL and GFormer produce larger recommendation sets (*c.f.*, Fig. 3d) while minimizing item modifications. In the next experiments, we focus on LightGCL since it emerged as the best scoring function.

(RQ4) There exists a trade-off between ensuring the safety of replacement items and the recommendation set size. Further, we investigate the impact of relaxing Property 1, which is required to satisfy Proposition 5.2. Specifically, we conduct an *ablation study* to assess the effect of the approximate satisfaction of this assumption. Fig. 5 shows the effect of increasing the filtering threshold β to select safer replacement items. As shown in Fig. 5a, when no filtering is applied (None), risk control fails — indeed, some previously seen items, when recommended again, are revealed to be unwanted and reported by users (as shown in Table 2). This behavior reflects a form of distribution shift between calibration and test phases [6]. Conversely, filtering items based on their initial watch-time ($W_\% > 0\%, 50\%, 100\%$) restores risk control, as seen by results falling on or below the diagonal in Fig. 5a. However, stricter filtering reduces the pool of candidate items, approaching the behavior of the Remove strategy, as indicated in Figs. 5c and 5d. This reveals a trade-off between ensuring risk control and maintaining a sufficiently large recommendation set under finite replacement options.

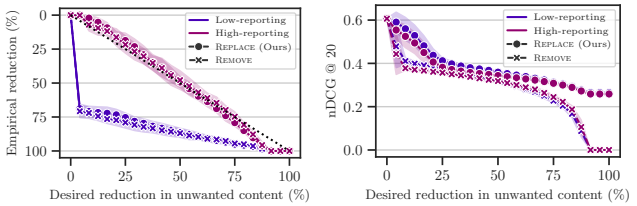
(RQ5) Risk control is biased toward high-reporting users. Lastly, we analyze the impact of the risk control procedure across user groups. We can categorize users as being *low-reporting*, who report a smaller fraction of videos ($H_X < 0.1$, cf. Table 1) versus *high-reporting*, who report a higher fraction of videos ($H_X \geq 0.1$). Fig. 6 presents the empirical reduction of unwanted content for both groups under different strategies. Notably, *low-reporting* users exhibit a greater reduction at lower α values, while other performance metrics (*e.g.*, nDCG in Fig. 6b) remain stable across groups and strategies. This suggests that *high-reporting* users disproportionately influence the calibration of the risk threshold λ, potentially leading to overly conservative behavior for everybody. In practice, equivalent risk guarantees for *low-reporting* users could be achieved by modifying or removing *fewer items*.

8 Discussion and Limitations

In this section, we discuss some assumptions and limitations of our work, which open up interesting avenues for future work.

Methodology. The risk control procedure (Algorithm 1) computes a global threshold that minimizes expected risk across users. Although users can select and reliably achieve a target level of unwanted content reduction, a single global threshold affects user groups differently (*e.g.*, *low-* vs. *high-reporting* behavior). In practice, it may be preferable to personalize the threshold, akin to *group-balanced* conformal prediction [4], by calibrating $\hat{\lambda}$ using data from a single user, provided sufficient interactions are available. We also assume that user preferences over unwanted content remain stable over time; future work could relax this assumption by modeling dynamic user profiles [18]. Moreover, the current risk function (Eq. (5)) does not account for item position in the ranking and treats each item as contributing equally to risk. Extensions could incorporate user models such as *discrete choice models* [37], enabling more realistic risk functions. Finally, the guarantees in Theorem 4.1 hold *in expectation*; further extensions could include probabilistic guarantees [7, 10] or support for non-monotonic risk [5].

Evaluation. Our analysis and experiments rely on a single dataset, as discussed in Section 3.4, which currently offers the only realistic basis for studying this phenomenon. This reliance may limit the external validity of our findings. Further evaluation, particularly of assumptions such as Property 1, will require additional datasets.



(a) Reduction in $R_H(S_\lambda(U,k))$ **(b) nDCG @ 20**

Figure 6: Ablation analysis on *low-* or *high*-reporting users (LightGCL, $k = 20$ and $\beta = 0\%$). We report the results over 10 runs, where the shaded area indicates the standard deviation.

However, acquiring such data remains a significant challenge. Regulatory frameworks may help address this gap. For example, Article 40 of the European Digital Services Act[5][6] permits vetted researchers to access data from very large online platforms (*e.g.*, Facebook, Twitter) for scientific purposes, potentially enabling validation of Algorithm 1. Nonetheless, privacy and ethical considerations remain crucial, requiring appropriate data handling and anonymization.

User experience. Lastly, replacing certain items in the recommendation pool with previously seen content may lead to unbalanced recommendations, repeatedly exposing users to the same content. This could have unintended negative effects, such as increased polarization [20, 58]. While our chosen replacement strategy (see Section 5.1) is relatively conservative, future work could explore alternatives based on curated item lists from the community or user-provided signals, such as liked videos or watch-later queues. Additionally, recent research on repeated consumption [21, 48, 59] offers targeted methods that may also address the balance between risk control and recommendation fairness [60]. Interestingly, alternative approaches could adopt a more systemic perspective. For example, a video-sharing platform might aggregate user feedback to identify content to downrank, allowing individuals to share "unwanted" content signals within trusted groups, such as friends or communities. Users with limited feedback activity could opt into these shared signals, effectively leveraging collective judgments to inform their own recommendations. This could help address cold-start issues [46] in the context of safe repeated items. Finally, user studies could yield more robust empirical evidence on how repetition and replacement strategies affect satisfaction, guiding further development in this area.

9 Conclusions

This work introduces a simple but effective user-centric procedure (*c.f.*, Algorithm 1) grounded in conformal risk control to mitigate the fraction of unwanted content suggested by any recommender systems. Unlike prior approaches that offer limited guarantees, our method proactively bounds the presence of unwanted content in recommendations by ensuring a connection between user feedback and change in the recommender behavior. The procedure is model-agnostic, easy to integrate with existing systems, and backed by distribution-free theoretical guarantees. We provide a detailed analysis of real-world interaction patterns (*e.g.*, reporting behavior and watch time) to inform the design of our approach. We further introduce a practical heuristic for selecting repeated content that balances safety and relevance. Experiments on a real-world dataset demonstrate that our framework can controllably reduce exposure to unwanted recommendations by providing a trade-off between the recommendation set size and utility. Finally, these findings point toward developing safer and user-centric recommender systems.

[5]https://eur-lex.europa.eu/eli/reg/2022/2065/oj
[6]https://data-access.dsa.ec.europa.eu/home

Acknowledgments

The work was partially supported by the following projects: Horizon Europe Programme, grants #101120237-ELIAS and #101120763-TANGO. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Yongsu Ahn, Quinn K Wolter, Jonilyn Dick, Janet Dick, and Yu-Ru Lin. 2024. Interactive Counterfactual Exploration of Algorithmic Harms in Recommender Systems. In *2024 IEEE Visualization in Data Science (VDS)*. IEEE, 35–39.

[2] Muhammad Ali. 2021. Measuring and Mitigating Bias and Harm in Personalized Advertising. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 869–872. doi:10.1145/3460231.3473895

[3] Ashton Anderson, Ravi Kumar, Andrew Tomkins, and Sergei Vassilvitskii. 2014. The dynamics of repeat consumption. In *Proceedings of the 23rd international conference on World wide web*. 419–430.

[4] Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* 16, 4 (March 2023), 494–591. doi:10.1561/2200000101

[5] Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. 2021. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052* (2021).

[6] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2023. Conformal risk control. *ICLR* (2023).

[7] Anastasios N Angelopoulos, Karl Krauth, Stephen Bates, Yixin Wang, and Michael I Jordan. 2023. Recommendation systems with distribution-free reliability guarantees. In *Conformal and Probabilistic Prediction with Applications*. PMLR, 175–193.

[8] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*. 759–760.

[9] Ricardo Baeza-Yates. 2020. Bias in search and recommender systems. In *Proceedings of the 14th ACM conference on recommender systems*. 2–2.

[10] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)* 68, 6 (2021), 1–34.

[11] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 21–30.

[12] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management* 58, 1 (2021), 102387.

[13] Ludovico Boratto and Mirko Marras. 2021. Advances in Bias-aware Recommendation on the Web. In *Proceedings of the 14th ACM international conference on web search and data mining*. 1147–1149.

[14] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. 2019. The Music Streaming Sessions Dataset. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2594–2600. doi:10.1145/3308558.3313641

[15] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. 2023. LightGCL: Simple Yet Effective Graph Contrastive Learning for Recommendation. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=FKXVK9dyMM

[16] Yuanfeng Cai and Dan Zhu. 2019. Trustworthy and profit: A new value-based neighbor selection method in recommender systems under shilling attacks. *Decision Support Systems* 124 (2019), 113112.

[17] Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal Prediction for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics* 12 (11 2024), 1497–1516. doi:10.1162/tacl_a_00715 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00715/2480372/tacl_a_00715.pdf

[18] Jerry Chee, Shankar Kalyanaraman, Sindhu Kiranmai Ernala, Udi Weinsberg, Sarah Dean, and Stratis Ioannidis. 2024. Harm mitigation in recommender systems under user preference dynamics. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 255–265.

[19] Sirui Chen, Jiawei Chen, Sheng Zhou, Bohao Wang, Shen Han, Chanfei Su, Yuqing Yuan, and Can Wang. 2024. SIGformer: Sign-aware Graph Transformer for Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA)

(SIGIR '24). Association for Computing Machinery, New York, NY, USA, 1274–1284. doi:10.1145/3626772.3657747

[20] Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. 2022. The effect of people recommenders on echo chambers and polarization. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 90–101.

[21] Sunhao Dai, Changle Qu, Sirui Chen, Xiao Zhang, and Jun Xu. 2024. ReCODE: Modeling Repeat Consumption with Neural ODE. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2599–2603. doi:10.1145/3626772.3657936

[22] Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. 2025. Towards human-AI complementarity with prediction sets. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 986, 30 pages.

[23] Yashar Deldjoo, Markus Schedl, Balázs Hidasi, Yinwei Wei, and Xiangnan He. 2021. Multimedia recommender systems: Algorithms and challenges. In *Recommender systems handbook*. Springer, 973–1014.

[24] Elizabeth Dwoskin. 2020. Facebook content moderators are suffering from PTSD symptoms. Now they're suing. *The Washington Post* (2020). https://www.washingtonpost.com/technology/2020/05/12/facebook-content-moderator-ptsd/

[25] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, et al. 2022. Towards responsible media recommendation. *AI and Ethics* (2022), 1–12.

[26] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. 2371–2382.

[27] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. " I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 153–162.

[28] Miriam Fernández and Alejandro Bellogín. 2020. Recommender Systems and Misinformation: The Problem or the Solution?. In *OHARS@RecSys*. 40–50. https://ceur-ws.org/Vol-2758/OHARS-paper3.pdf

[29] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. 2022. KuaiRand: An Unbiased Sequential Recommendation Dataset with Randomly Exposed Videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 3953–3957. doi:10.1145/3511808.3557624

[30] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, and Yong Li. 2023. A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. *ACM Trans. Recomm. Syst.* 1, 1, Article 3 (March 2023), 51 pages. doi:10.1145/3568022

[31] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*. 257–260.

[32] Yingqiang Ge, Shuchang Liu, Zuohui Fu, Juntao Tan, Zelong Li, Shuyuan Xu, Yunqi Li, Yikun Xian, and Yongfeng Zhang. 2024. A survey on trustworthy recommender systems. *ACM Transactions on Recommender Systems* 3, 2 (2024), 1–68.

[33] Nada Ghanem, Stephan Leitner, and Dietmar Jannach. 2022. Balancing consumer and business value of recommender systems: A simulation-based analysis. *Electronic Commerce Research and Applications* 55 (2022), 101195.

[34] Hanna Hauptmann, Alan Said, and Christoph Trattner. 2022. Research directions in recommender systems for health and well-being: A Preface to the Special Issue. *User Modeling and User-Adapted Interaction* 32, 5 (2022), 781–786.

[35] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[36] Atoosa Kasirzadeh and Charles Evans. 2023. User tampering in reinforcement learning recommender systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 58–69.

[37] Thorsten Krause, Alina Deriyeva, Jan H. Beinke, Gerrit Y. Bartels, and Oliver Thomas. 2024. Mitigating Exposure Bias in Recommender Systems—A Comparative Analysis of Discrete Choice Models. *ACM Trans. Recomm. Syst.* 3, 2, Article 19 (Nov. 2024), 37 pages. doi:10.1145/3641291

[38] Chaoliu Li, Lianghao Xia, Xubin Ren, Yaowen Ye, Yong Xu, and Chao Huang. 2023. Graph transformer for recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. 1680–1689.

[39] Ziyi Liang, Tianmin Xie, Xin Tong, and Matteo Sesia. 2024. Structured Conformal Inference for Matrix Completion with Applications to Group Recommender Systems. *arXiv preprint arXiv:2404.17561* (2024).

[40] Alexander Liu, Siqi Wu, and Paul Resnick. 2024. How to Train Your YouTube Recommender to Avoid Unwanted Videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 930–942.

[41] Chao Liu, Ryen W White, and Susan Dumais. 2010. Understanding web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 379–386.

[42] Manal Loukili, Fayçal Messaoudi, and Mohammed El Ghazi. 2023. Machine learning based recommender system for e-commerce. *IAES International Journal of Artificial Intelligence* 12, 4 (2023), 1803–1811.

[43] Jesse McCrosky and Brandi Geurkink. 2021. YouTube Regrets: A crowdsourced investigation into YouTube's recommendation algorithm. *Mozilla Foundation. Retrieved November* 15 (2021), 2021.

[44] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2020-07-25) *(SIGIR '20)*. Association for Computing Machinery, 429–438. doi:10.1145/3397271.3401100

[45] Tien T Nguyen, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2018. User personality and user satisfaction with recommender systems. *Information systems frontiers* 20 (2018), 1173–1189.

[46] Seung-Taek Park and Wei Chu. 2009. Pairwise preference regression for cold-start recommendation. In *Proceedings of the third ACM conference on Recommender systems*. 21–28.

[47] Gabriele Piccoli. 2016. Triggered essential reviewing: the effect of technology affordances on service experience evaluations. *European Journal of Information Systems* 25, 6 (2016), 477–492. doi:10.1057/s41303-016-0019-9 arXiv:https://doi.org/10.1057/s41303-016-0019-9

[48] Pengjie Ren, Zhumin Chen, Jing Li, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019. RepeatNet: A Repeat Aware Neural Recommendation Machine for Session-Based Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 4806–4813. doi:10.1609/aaai.v33i01.33014806

[49] Raquel Sánchez-Fernández and David Jiménez-Castillo. 2021. How social media influencers affect behavioural intentions towards recommended brands: the role of emotional attachment and information value. *Journal of Marketing Management* 37, 11-12 (2021), 1123–1147.

[50] Philipp J Schneider and Marian-Andrei Rizoiu. 2023. The effectiveness of moderating harmful online content. *Proceedings of the National Academy of Sciences* 120, 34 (2023), e2307360120.

[51] Changwon Seo, Kyeong-Joong Jeong, Sungsu Lim, and Won-Yong Shin. 2022. SiReN: Sign-aware recommendation using graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 35, 4 (2022), 4729–4743.

[52] Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research* 9, 3 (2008).

[53] Jessie J. Smith, Lucia Jayne, and Robin Burke. 2022. Recommender Systems and Algorithmic Hate. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 592–597. doi:10.1145/3523227.3551480

[54] Eleni Straitouri and Manuel Gomez Rodriguez. 2024. Designing decision support systems using counterfactual prediction sets. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 1901, 23 pages.

[55] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. 2023. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*. PMLR, 32633–32653.

[56] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. 2021. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3460231.3474241

[57] Antonela Tommasel, Daniela Godoy, and Arkaitz Zubiaga. 2021. OHARS: Second Workshop on Online Misinformation- and Harm-Aware Recommender Systems. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 789–791. doi:10.1145/3460231.3470941

[58] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. 2021. I want to break free! recommending friends from outside the echo chamber. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 23–33.

[59] Chenyang Wang, Min Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2019. Modeling item-specific temporal dynamics of repeat consumption for recommender systems. In *The world wide web conference*. 1977–1987.

[60] Lequn Wang and Thorsten Joachims. 2023. Uncertainty Quantification for Fairness in Two-Stage Recommender Systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (Singapore, Singapore) *(WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 940–948. doi:10.1145/3539597.3570469

[61] Yunpeng Xu, Wenge Guo, and Zhi Wei. 2024. Two-stage Conformal Risk Control with Application to Ranked Retrieval. *arXiv preprint arXiv:2404.17769* (2024).

[62] Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. 2023. Prefrec: Recommender systems with human preferences for reinforcing long-term user engagement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2874–2884.

[63] Zixuan Yi and Iadh Ounis. 2024. A Unified Graph Transformer for Overcoming Isolations in Multi-modal Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) *(RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 518–527. doi:10.1145/3640457.3688096

[64] Peifeng Yin, Ping Luo, Wang-Chien Lee, and Min Wang. 2013. Silence is also evidence: interpreting dwell time for recommendation from psychological perspective. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 989–997.

[65] Brita Ytre-Arne and Hallvard Moe. 2021. Folk theories of algorithms: Understanding digital irritation. *Media, Culture & Society* 43, 5 (2021), 807–824. doi:10.1177/0163443720972314 arXiv:https://doi.org/10.1177/0163443720972314

[66] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in ranking, part i: Score-based ranking. *Comput. Surveys* 55, 6 (2022), 1–36.