You Don't Bring Me Flowers: Mitigating Unwanted Recommendations Through Conformal Risk Control (Extended Abstract)*

Giovanni De Toni[1], **Erasmo Purificato**[2†], **Emilia Gómez**[2†], **Andrea Passerini**[3], **Bruno Lepri**[1], **Cristian Consonni**[2†]
[1]Fondazione Bruno Kessler
[2]European Commission, Joint Research Centre (JRC)
[3]University of Trento
{gdetoni, lepri}@fbk.eu, {erasmo.purificato, cristian.consonni, emilia.gomez-gutierrez}@ec.europa.eu, andrea.passerini@unitn.it

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.

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 [Ghanem *et al.*, 2022; Loukili *et al.*, 2023], social media [Gao *et al.*, 2023; Sánchez-Fernández and Jiménez-Castillo, 2021], and content streaming [Deldjoo *et al.*, 2021; Elahi *et al.*, 2022], and tailoring experiences based on historical behaviors and expressed preferences [Fernández and Bellogín, 2020; Boratto *et al.*, 2021]. By leveraging user-generated data, recommenders aim to enhance user engagement [Xue *et al.*, 2023], satisfaction [Nguyen *et al.*, 2018], and platform profitability [Cai and Zhu, 2019]. 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 [Liu *et al.*, 2024]. Platforms have faced increasing criticism for creating *filter bubbles* and *rabbit holes* that reinforce existing beliefs, and for spreading potentially dangerous or harmful content [Tomlein *et al.*, 2021]. Such dynamics are often attributed to algorithms that prioritize *engagement*, sometimes at the expense of content quality and truthfulness [Tommasel *et al.*, 2021]. Features like the "*Not Interested*" button on YouTube [Liu *et al.*, 2024] 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 [Bernstein *et al.*, 2013; Eslami *et al.*, 2015; Eslami *et al.*, 2016; Liu *et al.*, 2024]. Recent works have studied how to mitigate unwanted content, biases, or ensure fairness in recommendations [Chee *et al.*, 2024; Ahn *et al.*, 2024], but it remains unclear how to grant users precise and transparent control over the systems' *risk*.

Our contributions. We propose a post-hoc method based on conformal prediction [Shafer and Vovk, 2008; Angelopoulos and Bates, 2023] 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 [Straitouri *et*

*Full version appears in the proceedings of RecSys 2025 [De Toni *et al.*, 2025b], winner of the Best Paper Award.

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

al., 2023; Straitouri and Rodriguez, 2024; De Toni *et al.*, 2025a], natural language processing [Campos *et al.*, 2024], and recommendation systems [Angelopoulos *et al.*, 2023b; Xu *et al.*, 2024]. 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* [Angelopoulos *et al.*, 2023a], 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 Case Study: A Video-Sharing Platform

We begin our study by examining data from a popular Chinese short-form video-sharing platform, *Kuaishou*, which has over 300 million daily users at the time of writing. We use the *KuaiRand* dataset [Gao *et al.*, 2022], 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. The dataset also includes *negative feedback* signals, such as when users click the *"Do not recommend"* or *"Report"* buttons while watching a video. 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 $9,223$.

Summary of the analysis. From our analysis[1], we report some key insights that will be useful in designing risk-controlling recommendation systems:

- The engagement and perceived harmfulness of videos are not fully correlated. Reported videos have the same watch time 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 1). Since 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, very short engagement can serve as an *early indicator of harmful or undesired content*.

[1]The full in-depth analysis of *KuaiRand* is available in Section 3 of the main manuscript [De Toni *et al.*, 2025b].

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)

Table 1: 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.

3 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 $\mathcal{I} \in \{i_1, \ldots, i_N\}$ and a finite set of users $\mathcal{U} \in \{u_1, \ldots, u_M\}$, 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.*, 2023b], 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:

$$T_\lambda(U) = \{i \in \mathcal{I} : s(U, I = i) \geq \lambda\} \tag{1}$$

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* [Ge *et al.*, 2010]). 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 $\pi(\mathcal{I}) = \{i_1, \ldots, i_N\}$ as 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 $\pi(T_\lambda(U))$:

$$S_\lambda(U, k) = \{i_j \in \pi(T_\lambda(U)) : j \leq k\} \tag{2}$$

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 $\mathcal{S}_\lambda(U,k)$ 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 re-training 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 $\alpha \in (0,1)$:

$$\lambda^* = \underset{\lambda \in \Lambda}{\text{argmax}}\, \mathbb{E}\left[\ell(S_\lambda(U,k))\right] \text{ s.t. } \mathbb{E}[R(S_\lambda(U,k))] \leq \alpha \quad (3)$$

In our setting, we consider an item *unwanted* for a user if she expresses distaste for it once seen within the recommendations. 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:

$$R_H(S_\lambda(U,k)) = \frac{|\{i \in S_\lambda(U,k) : H(U, I=i) = 1\}|}{|S_\lambda(U,k)|} \quad (4)$$

3.1 Conformal Risk Control

By simply plugging the previous risk definition in Eq. (3), we have our optimization objective. However, how do we pick the threshold λ to control the risk defined by Eq. (4)? We can find the optimal threshold via *conformal risk control* [Angelopoulos *et al.*, 2023a] by employing a held-out calibration set $\{(u,i,h))_j\}_{i=1}^n$, where for each user-item pair we also record if it was flagged as *unwanted* ($h = 1$). We first state the main theorem from [Angelopoulos *et al.*, 2023a], and then we explain its practical implications.

Theorem 1. *Assume that $R_H(S_\lambda(U,k))$ is non-increasing in λ, right-continuous, $R_H(S_{\lambda_{max}}(U,k)) \leq \alpha$ and $\sup_\lambda R_H(S_\lambda(U,k)) \leq 1 < \infty$ 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:*

$$\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\hat{R}(\lambda) + \frac{1}{n+1} \leq \alpha\right\} \quad (5)$$

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 1, we are guaranteed the risk will be *provably* bound in expectation. If we consider the ranker output $s(U,I) = f(U,I)$ as a score, 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 $R_H(S_\lambda(U,k))$.

4 Risk Control via Item Replacement

Recent applications of conformal prediction to recommender systems consider only the simple filtering approach as detailed in Eq. (1) [Angelopoulos *et al.*, 2023b; Xu *et al.*, 2024]. The approach ensures the monotonicity requirements of Theorem 1 but at the price of losing control of the size of the resulting set $S_\lambda(U,k)$ [De Toni *et al.*, 2025b, Prop. 5.1]. Rather than simply removing items, we take an alternative approach, which consists of *replacing* potentially unwanted items with safer ones. However, how do we choose these *safe items*? Our approach exploits a simple fact: *users consume previously seen items* [Anderson *et al.*, 2014; Dai *et al.*, 2024]. Inspired by the analysis in Section 2, 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 1, there exist *rare* items (*cf.* ✓ → ✗, 0.11%) that are reported only the second time a user watches them. 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:

Algorithm 1 Risk-controlling 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 set size
1: $\hat{\lambda} \leftarrow \text{RiskControl}(\alpha)$ {Theorem 1}
2: $T_{\hat{\lambda}}(U) \leftarrow \{i \in \mathcal{I} : s(U,i) \geq \hat{\lambda}\}$
3: $T^{(\text{safe})} \leftarrow \{i' \in \mathcal{I}_U : H(U,i') = 0 \wedge W_\%(U,i') > \beta\}$
4: $T_{\hat{\lambda}}^{(\text{replace})}(U) \leftarrow T_{\hat{\lambda}}(U) \cup T^{(\text{safe})}(U)$
5: **return** $\{i_j \in \pi(T_{\hat{\lambda}}^{(\text{replace})}(U)) : j \leq k\}$

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 choose videos on which the users have spent time $W_\%(U,I)$ above a certain threshold. If Property 1 holds, we can 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 $C > \beta$. Let us denote the pool of *safe* videos as:

$$T^{(\text{safe})} = \{i' \in \mathcal{I}_U : H(U,i') = 0 \wedge\ C(i') > \beta\} \quad (6)$$

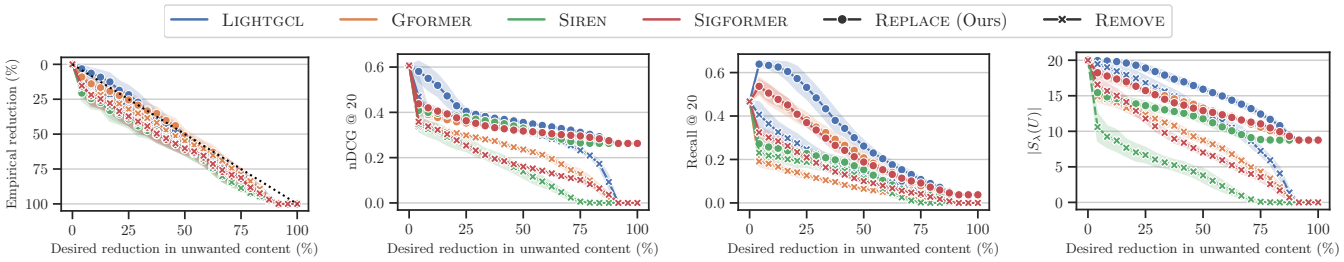
where $\mathcal{I}_U \subseteq \mathcal{I}$ indicates the items previously shown to the user U. Thus, the new pool of items will become:

$$T_\lambda^{(\text{replace})}(U) = T_\lambda(U) \cup T^{(\text{safe})}(U) \quad (7)$$

It is straightforward to show that the new set $T_\lambda^{(\text{replace})}(U)$ satisfies the monotonicity requirements of conformal risk control, and it ensures we can always pick k items to show to the user [De Toni *et al.*, 2025b, Proposition 2]. We consolidate the previous findings into Algorithm 1, a post-hoc algorithm applicable to *any* pre-trained recommender system.

5 Experiments With Real Data

We evaluate the effects of the risk control strategies (Eqs. (1) and (7)) on a simple two-stage recommender setup for a realistic video recommendation task, towards understanding what



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

Figure 1: Results over *KuaiRand* ($k = 20$ and $\beta = 0$). The results and standard deviation (shaded areas) are over 10 runs.

the impact of the conformal risk control on standard evaluation metrics (*e.g.*, *Recall@k* and *nDCG@k*) is and what the effects of employing different strategies for risk control are (Eq. (1) and Eq. (7)). In the full paper, we further analyze other aspects, such as how many items we have to replace to achieve the desired risk level, or how approximately satisfying Property 1 impacts the desired unwanted content reduction [De Toni *et al.*, 2025b, Sec. 7].

Dataset & Metrics. We use the *KuaiRand* dataset [Gao *et al.*, 2022], which contains both positive and negative user feedback, along with information about interactions with previously seen videos. 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 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 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. 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 is usually done in the literature [Chen *et al.*, 2024]. We also measure the empirical *risk*, *i.e.*, the fraction of unwanted content in the recommendations, as defined in Eq. (4).

Models. Similar to previous studies [Wang and Joachims, 2023], we evaluate the effectiveness of risk control using a two-stage recommender architecture. 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 [Angelopoulos *et al.*, 2023b]. For simplicity, under the REPLACE strategy, repeated items are not re-scored; instead, their original watch time is reused. We experiment with several state-of-the-art methods, including both sign-aware and unsigned graph-based models, depending on whether they incorporate negative user feedback (*e.g.*, reported videos): **LightGCL** [Cai *et al.*, 2023], **GFormer** [Li *et al.*, 2023], **SiReN** [Seo *et al.*, 2022], and **SIGFormer** [Chen *et al.*, 2024]. For the re-ranking stage, we adopt a simple Neural Collaborative Filtering (NCF) model [He *et al.*, 2017], as our focus is not on maximizing performance on the Kuaishou dataset, but rather on isolating the effects of risk control. Our full implementation is openly available[2] to support further research.

6 Results & Discussion

Fig. 1a shows how 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. 1b and 1c show how risk control affects nDCG@20 and Recall@20. Removing items sharply degrades nDCG as the desired reduction level (α in Eq. (3)) increases, due to the shrinking pool of candidate items (Eq. (1)). 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. 1d illustrates that, 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 fewer than k items at test time, since we have finite replacement options.

Ultimately, our experimental results demonstrate that our framework can controllably reduce exposure to unwanted recommendations by providing a trade-off between the recommendation set size and utility. These findings point toward developing safer and user-centric recommender systems.

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

Acknowledgments

The work of GDT, BL, and AP 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

[Ahn *et al.*, 2024] Yongsu Ahn, Quinn K Wolter, Jonilyn Dick, Janet Dick, and Yu-Ru Lin. Interactive counterfactual exploration of algorithmic harms in recommender systems. In *2024 IEEE Visualization in Data Science (VDS)*, pages 35–39. IEEE, 2024.

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

[Angelopoulos and Bates, 2023] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591, March 2023.

[Angelopoulos *et al.*, 2023a] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *ICLR*, 2023.

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

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

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

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

[Cai *et al.*, 2023] Xuheng Cai, Chao Huang, Lianghao Xia, and Xubin Ren. LightGCL: Simple yet effective graph contrastive learning for recommendation. In *The Eleventh International Conference on Learning Representations*, 2023.

[Campos *et al.*, 2024] Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 11 2024.

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

[Chen *et al.*, 2024] Sirui Chen, Jiawei Chen, Sheng Zhou, Bohao Wang, Shen Han, Chanfei Su, Yuqing Yuan, and Can Wang. Sigformer: Sign-aware graph transformer for recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1274–1284, New York, NY, USA, 2024. Association for Computing Machinery.

[Dai *et al.*, 2024] Sunhao Dai, Changle Qu, Sirui Chen, Xiao Zhang, and Jun Xu. Recode: Modeling repeat consumption with neural ode. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2599–2603, New York, NY, USA, 2024. Association for Computing Machinery.

[De Toni *et al.*, 2025a] Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez-Rodriguez. Towards human-ai complementarity with prediction sets. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc.

[De Toni *et al.*, 2025b] Giovanni De Toni, Erasmo Purificato, Emilia Gomez, Andrea Passerini, Bruno Lepri, and Cristian Consonni. 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, page 492–502, New York, NY, USA, 2025. Association for Computing Machinery.

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

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

[Eslami *et al.*, 2015] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. " 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*, pages 153–162, 2015.

[Eslami *et al.*, 2016] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman,

Kevin Hamilton, and Alex Kirlik. 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*, pages 2371–2382, 2016.

[Fernández and Bellogín, 2020] Miriam Fernández and Alejandro Bellogín. Recommender systems and misinformation: The problem or the solution? In *OHARS@RecSys*, pages 40–50, 2020.

[Gao *et al.*, 2022] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 3953–3957, New York, NY, USA, 2022. Association for Computing Machinery.

[Gao *et al.*, 2023] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, and Yong Li. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Trans. Recomm. Syst.*, 1(1), March 2023.

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

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

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

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

[Liu *et al.*, 2024] Alexander Liu, Siqi Wu, and Paul Resnick. How to train your youtube recommender to avoid unwanted videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 930–942, 2024.

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

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

[Sánchez-Fernández and Jiménez-Castillo, 2021] Raquel Sánchez-Fernández and David Jiménez-Castillo. 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):1123–1147, 2021.

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

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

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

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

[Tomlein *et al.*, 2021] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. An audit of misinformation filter bubbles on youtube: Bubble bursting and recent behavior changes. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 1–11, New York, NY, USA, 2021. Association for Computing Machinery.

[Tommasel *et al.*, 2021] Antonela Tommasel, Daniela Godoy, and Arkaitz Zubiaga. Ohars: Second workshop on online misinformation- and harm-aware recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 789–791, New York, NY, USA, 2021. Association for Computing Machinery.

[Wang and Joachims, 2023] Lequn Wang and Thorsten Joachims. Uncertainty quantification for fairness in two-stage recommender systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 940–948, New York, NY, USA, 2023. Association for Computing Machinery.

[Xu *et al.*, 2024] Yunpeng Xu, Wenge Guo, and Zhi Wei. Two-stage conformal risk control with application to ranked retrieval. *arXiv preprint arXiv:2404.17769*, 2024.

[Xue *et al.*, 2023] Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. 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*, 2023.