First, do no harm: Breaking suicidogenic echo chambers in media recommendation

Alberto Díaz-Álvarez(✉)[0000−0002−4150−9052], Raúl Lara-Cabrera[0000−0002−7959−1936], Fernando Ortega-Requena[0000−0003−4765−1479], and Víctor Ramos-Osuna[0009−0009−6139−967X]

E.T.S.I. Sistemas Informáticos (Universidad Politécnica de Madrid), C. de Alan Turing, s/n, 28031 Madrid, Spain
{alberto.diaz,raul.lara,fernando.ortega,victor.ramos.osuna}@upm.es

Abstract. Recommender systems generally optimises user engagement, but this approach is dangerous in mental health contexts. When vulnerable users show signs of suicidal ideation, standard algorithms often trap them in echo chambers of harmful content, worsening their psychological state. In response, we introduce RankAid, a re-ranking method that prioritises clinical safety alongside predictive relevance. It works as an add-on layer to existing models: it penalises risky items and boosts therapeutic content depending on the user's current level of vulnerability. We evaluated this approach using the MovieLens 1M dataset, where items were semantically annotated for clinical risk and therapeutic value using large language models. Our simulations show that our algorithm successfully blocks the recommendation of harmful content during crisis peaks, actively reshaping the feed to support emotional de-escalation. Furthermore, this safety intervention only causes a controlled, acceptable drop in standard accuracy metrics like NDCG. By using asymmetric hyperparameters, RankAid also gives system administrators the flexibility to tune the severity of the intervention based on specific clinical guidelines.

Keywords: Recommender systems · Algorithmic ethics · Crisis intervention · Human-centered AI · Machine learning for social good.

1 Introduction

Recommender systems (RSs) are a core component of digital platforms, typically optimised to maximise user engagement and retention. As almost everyone spends more time on algorithmic platforms, their impact on psychological well-being increases. While recommenders excel at finding relevant content, they usually ignore the user's emotional context. Their performance is traditionally evaluated with ranking metrics. In this work, we use Normalized Discounted Cumulative Gain (NDCG) [12] as our main accuracy metric, as it rewards placing the most relevant items at the top of the list and penalises those ranked lower.

However, because these models prioritise click probability, they assume all high-engagement content is suitable. In mental health and suicide prevention,

this optimisation is dangerous. Previous research shows that vulnerable users experiencing suicidal ideation often fall into "digital rumination" [16]. In such cases, traditional RS risk trapping individuals in echo chambers of nihilistic or self-harming content, potentially worsening their mental state [1].

While some studies use heuristics or language models to mitigate harmful content [9,15], there is currently no mathematical function that combines clinical vulnerability with psychological risk to explicitly penalise exposure.

We propose a "preventive adjustment" via RankAid, a re-ranking algorithm that modifies predictions in crisis contexts. By weighing a users latent vulnerability against an items psychological risk and therapeutic value, RankAid penalises harmful recommendations for high-risk users, regardless of engagement potential, and promotes therapeutic content to help de-escalate the crisis or reduce rumination.

2 Related work

Recent advancements in large language models (LLMs) have expanded the intersection of artificial intelligence (AI) and mental health; yet, recommendation algorithms still fall short in crisis contexts.

Suicide risk detection: Current literature mainly addresses suicide prevention with natural language processing (NLP) techniques. Many studies use deep learning (DL) models, such as *transformer* architectures, to detect suicidal ideation and depressive rumination. Analysing non-clinical texts (e.g., social media posts or search histories) is crucial for assessing psychological states in digital environments without medical records [6,4].

This work has produced triage systems that estimate intervention urgency from these signals. Metrics like hierarchical Time-Biased Gain (hTBG) help quickly identify individuals at suicide risk when time is limited [19]. However, these tools only alert clinical evaluators; they cannot assess the risk of individual content items or prevent users exposure to harmful material.

Conversational agents (i.e., crisis-intervention chatbots) have shown promise in providing early emotional support [7]. Yet both chatbots and detection systems operate separately from users' everyday content consumption and thus overlook a key issue: the RS itself may recommend harmful content that worsens the crisis.

Echo chambers: The formation of echo chambers [14,2] has recently been mathematically modelled. For example, [5] shows that when user preferences evolve, static penalties on problematic content can fail or even increase long-term harm. To address this, the "trustworthy recommender systems" paradigm [10] requires systems to be fair, explainable, and safe, prompting proposals such as filters for negative experiences [9] and using LLMs to re-rank feeds and hide self-harm material [15]. Our work also uses the semantic capabilities of LLMs.

Unlike models that delegate re-ranking directly to a LLM (which is computationally expensive and opaque), we first use the LLM to build a clinical

knowledge matrix. Instead of static penalties, our approach treats LLM assessments as variables in a mathematical function that sets rewards or penalties based on user vulnerability, aligning protection with digital wellbeing principles.

Ethical modification of algorithms: A common way to avoid echo chambers is to make algorithms also optimise content diversity, often by combining diversity and accuracy metrics [22]. In mental health, however, mathematical diversity alone is insufficient: it maximises list heterogeneity without modelling clinical risk, so a list can be diverse yet still contain explicit, implicit, or ambivalent self-harm content [21,18]. Position bias worsens this, as high-risk items at the top receive disproportionate attention [13]. Even when diverse items are harmless, variety without guaranteed relevance or clinical suitability is often treated as noise [23].

Given these risks, recommendation objectives must be adapted for clinical contexts. While formula changes for clinical reasons are rare, fairness research in RSs treats recommendations as a scarce resource to be fairly allocated under constraints [20]. Several authors modify classic metrics with penalties that ensure visibility for minority groups [8,3], effectively redefining the objective: some predictive accuracy is sacrificed to meet ethical criteria.

The tech industry has begun to acknowledge these algorithmic harms and has introduced mechanisms to limit exposure to negative user experiences (Negative User Experiences (NUXs)) [9]. However, social networks usually implement these as reactive *a posteriori* filters (heuristic *down-ranking*), rather than embedding them directly into the training and guiding formulas.

3 Methodology

Instead of building a new predictive model, we use a two-stage pipeline. First, a standard model generates relevance scores. Second, our algorithm recalculates each items value using clinical parameters before display. This model-agnostic strategy adds only a re-ranking layer to any recommendation system. The workflow used in our experiment (Figure 1) consists of the following steps:

1. **Candidate item generation**. We use a standard Neural Collaborative Filtering (NCF)-based algorithm [11], formulated as an implicit-feedback task predicting user interaction. As detailed in section 3, binarized ratings are fed to a multilayer perceptron (MLP) that assigns an affinity score $rel_{i,u} \in [0, 1]$ to each item. User and item IDs are mapped to 128-dimensional embeddings, concatenated, and passed through Dense layers (64 and 32 units) with Dropout, followed by a sigmoid output layer.
2. **Re-ranking** (RankAid algorithm). Given the $rel_{i,u}$ scores, our algorithm adjusts each item by applying penalties or rewards based on the user's mental state.
3. **Final top-N generation**. Items are sorted in descending order by the adjusted scores and presented to the user.

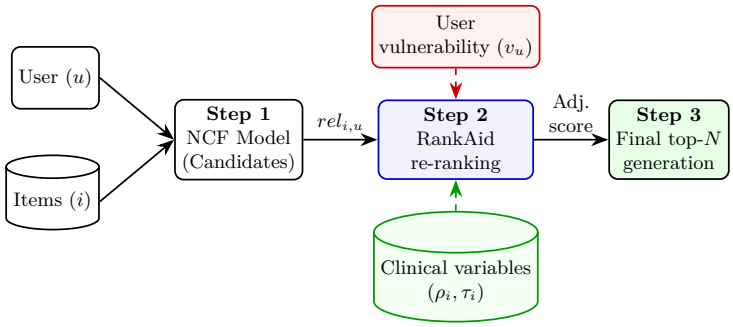


Fig. 1: General architecture of the pipeline. The system consists of a baseline predictive model and a safety re-ranking layer that actively adjusts item scores based on the dynamic vulnerability of the user before generating the final feed.

Our proposal adjusts the original recommendation by reducing the weight of harmful content and increasing the weight of therapeutic content based on the user's current sensitivity. We define this as a scoring function (Equation 1):

$$RankAid(i, u) = \max(0, rel_{i,u} - \alpha \cdot (v_u \cdot \rho_i) + \beta \cdot (v_u \cdot \tau_i)) \tag{1}$$

Where:

- $v_u \in [0, 1]$: **User vulnerability**, from stability (0) to extreme crisis (1). Following *Fluid Vulnerability Theory* [17], we model risk as a dynamic state inferred externally (e.g., via NLP).
- $\rho_i \in [0, 1]$: **Risk factor**, indicating an item's toxicity or danger, assumed pre-labelled by automatic classifiers or specialists.
- $\tau_i \in [0, 1]$: **Therapeutic rescue factor**, measuring an item's potential for emotional de-escalation (e.g., shifting focus away from trauma).
- α and β: **Intervention hyperparameters** controlling risk penalty (α) and therapeutic reward (β). Using separate parameters instead of a convex pair (e.g., λ, $1 - \lambda$) enables asymmetric calibration, letting clinicians tighten protection (e.g., strict censorship) without limiting therapeutic content.

If a user shows no clinical risk ($v_u \approx 0$), the penalty and reward terms vanish, leaving the original relevance $rel_{i,u}$ and preserving standard recommender behavior. As vulnerability increases ($v_u \to 1$), the term $-\alpha(v_u \cdot \rho_i)$ sharply lowers scores for harmful items, even when $rel_{i,u}$ is high. Unlike traditional interventions (e.g., *fairness*), which only penalize, our method also adds $+\beta(v_u \cdot \tau_i)$ to boost items with therapeutic value.

When a film has both high risk and high rescue factors ($\rho_i \approx 1$, $\tau_i \approx 1$), the formula balances them via α and β, keeping the final score close to its original relevance and avoiding over-promotion of highly emotional content. Although ρ_i and τ_i are currently fixed per film, the formulation can be extended to user-specific values (e.g., $\rho_{i,u}$) if personal triggers or medical history become available.

We evaluated our algorithm using a standard 80/20 traintest split. Before splitting, we applied a 20-core filter (at least 20 interactions), ensuring all users and movies in the test set had sufficient training history and avoiding cold-start issues. We then converted explicit ratings to an implicit task by binarizing labels: ratings ≥ 4 were treated as positive interactions. Unobserved items were added via negative sampling at a 1:4 ratio (a standard value [11]) to train the baseline NCF model. Using a fixed ratio removed the need for a validation set and prevented test-set data leakage.

Accuracy is measured against the explicit ground truth of the 20% test set using NDCG (see Equation 2). This ensures that any measured degradation reflects the actual loss in ranking quality relative to the user's real preferences.

$$NDCG_p = \frac{DCG_p}{IDCG_p} \qquad (2) \qquad\qquad DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \qquad (3)$$

Ideal DCG (IDCG) is calculated by sorting test items by their original explicit ratings (1 to 5) rather than the implicit labels used for training. This prevents ties and provides a more precise ideal baseline. If items have identical ratings (e.g., several 5s), their relative order does not affect the cumulative sum. This normalisation allows for consistent performance comparison across all users.

4 Experiments and results

To validate this methodology, we simulated it on the MovieLens 1M dataset. Because no public dataset links viewing histories with real clinical diagnoses, we used movie metadata as proxy variables. To capture catalogue variability and assign consistent content values, we built a semantic enrichment pipeline that matched dataset identifiers with TMDBs API to retrieve each movies full synopsis. To support open science and reproducibility, we released the experiment code and the extended, semantically annotated dataset on Codeberg[1].

We then performed automated clinical inference with a locally hosted, open-source 35Bparameter Qwen 3.5 LLM. To minimise randomness, we used the APIs native javascript object notation (JSON) mode with strict formatting and set the temperature to zero. The code also parsed the models internal reasoning steps to extract structured data. A different LLM might slightly change the exact risk and rescue scores, but the clinical labels should remain stable given the specific instructions. Through prompt engineering, we cast the model as an expert psychologist in suicide risk assessment and asked it to jointly analyse each movies title, synopsis, and tags. For each item, the LLM produced an evaluation with three structural variables (see Equation 1):

[1] https://codeberg.org/knodis/paper-caepia-2026-first-do-no-harm

- **Risk factor** ($\rho_i \in [0, 1]$): A continuous value quantifying the toxicity level or the movie's potential to induce negative rumination.
- **Therapeutic rescue factor** ($\tau_i \in [0, 1]$): A continuous value reflecting the content's potential to facilitate cognitive emotional de-escalation.
- **Clinical label**: A categorical classification assigned to one of three classes: Harmful, Therapeutic, or Neutral.

To ensure reproducibility, Table 1 details the structural instructions (prompt) provided to the language model for clinical inference. This approach allows us to obtain ρ_i and τ_i values based on a semantic analysis of the items' content, reflecting the variability found in a real recommendation catalogue.

Table 1: Prompt provided to the LLM for semantic annotation

```
You are an expert clinical psychologist specializing in media impact on
vulnerable individuals, specifically those experiencing suicidal
ideation or acute psychological crises.

Analyze the following movie metadata and assess two metrics:
1. rho (risk factor, float 0.0 to 1.0): How harmful, triggering, or
   toxic the movie could be. High risk includes themes of hopelessness,
   self-harm, severe trauma, nihilism, or explicit graphic violence.
2. tau (therapeutic rescue factor, float 0.0 to 1.0): The potential to
   generate emotional de-escalation, comfort, or hope. High rescue
   includes lighthearted comedy, positive overcoming of adversity,
   relaxing pacing, or deep emotional validation.

Based on rho and tau, assign a clinical label which MUST strictly be one
of: "Harmful", "Therapeutic", or "Neutral".

Title: {title}
Genres: {genres}
Synopsis: {overview}

OUTPUT FORMAT:
You must return ONLY a raw, valid JSON object. Do not include markdown
formatting, code blocks (do not use ```json), introductions, or
explanations. Adhere strictly to this schema:

{{
  "rho": <float>,
  "tau": <float>,
  "clinical_label": "<string>"
}}
```

Table 2 shows the general distribution of these clinical labels across the entire dataset.

Additionally, Table 3 provides concrete examples of the LLM's semantic annotation. This demonstrates how the model interprets specific content to assign risk and rescue factors.

To establish a baseline for this experiment, we initialised the hyperparameters α and β at 0.2. As detailed in section 3, these soft weights act as a starting "nudge" for the intervention, so they are designed to be adjusted based on the required domain/clinical criteria.

Table 2: Distribution of clinical labels

Clinical label	Count	Percentage	Avg. risk (ρ)	Avg. Rescue (τ)
Harmful	1096	28.6600	0.8411	0.1964
Neutral	811	21.2100	0.6395	0.4065
Therapeutic	1917	50.1300	0.2324	0.7430

Table 3: Examples of LLM-based semantic annotation

Movie title	Risk (ρ)	Rescue (τ)	Clinical label
Nixon (1995)	0.8500	0.2000	Harmful
Casino (1995)	0.7500	0.2000	Harmful
Toy Story (1995)	0.1000	0.9000	Therapeutic
Jumanji (1995)	0.3000	0.7000	Therapeutic
Heat (1995)	0.6500	0.3500	Neutral
Sudden Death (1995)	0.6500	0.3500	Neutral

4.1 Crisis escalation simulation

With this structure, the experiment consists of simulating a progressive escalation in the user's crisis (increasing the v_u variable from 0 to 1). The goal is to evaluate how the function changes the exposure of items in the top ranking positions compared to a traditional recommender.

Exposure control Initially, we evaluated the clinical safety of the recommended items. Exposure is defined as the average ratio of items belonging to a specific clinical label (e.g., "Harmful" or "Therapeutic") present within the top-10 recommendations for a user. As shown in Figure 2, a classic RS maintains a constant exposure to harmful content (around 10%) regardless of the user's state. In contrast, RankAid significantly changes this dynamic.

As the crisis intensifies, the recommender engages two protective behaviours. First, it reduces harm: as vulnerability rises, harmful content in the top-10 declines until it disappears. Second, it injects "rescue" content: items with high therapeutic value (τ_i) grow from about 55% of recommendations at emotional stability ($v_u = 0$) to nearly 95% at maximum crisis ($v_u = 1$). Together, these mechanisms show that the re-scoring function prevents ruminative echo chambers and turns the recommendation space into a de-escalation environment.

(Controlled) degradation of accuracy Any deviation from an algorithm's original predictions inevitably reduces its accuracy. By prioritising clinical criteria, the system stops optimising exclusively for historical engagement. To quantify this impact, we evaluated the resulting rankings using the NDCG metric.

As shown in Figure 3, RankAid causes a slight downward trend in accuracy as user vulnerability increases. However, the results demonstrate that this decrease remains within acceptable margins, preserving a stable NDCG index throughout the process.

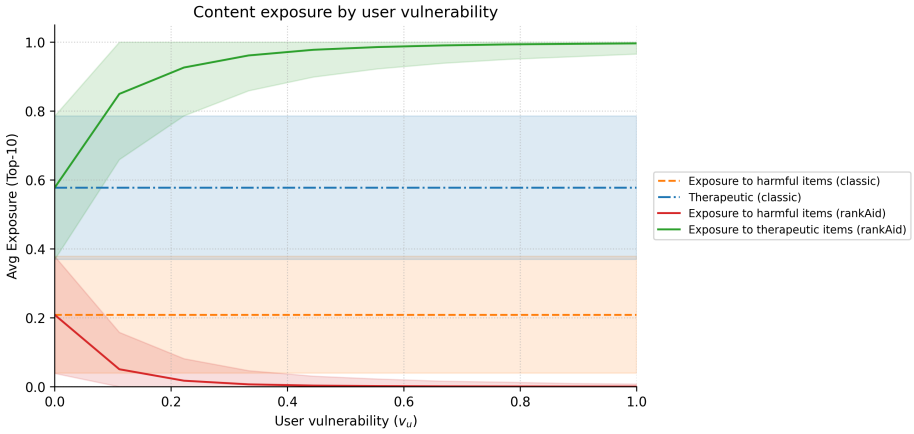


Fig. 2: Average content exposure in the top-10 recommendations as user vulnerability (v_u) increases. While the classic model maintains constant exposure to harmful (orange) and therapeutic items (blue), RankAid adapts dynamically. As the crisis escalates, it eliminates harmful content (red) and boosts therapeutic recommendations (green). Shaded areas show the standard deviation across all users.

Ethically, this reduction is not a model flaw but a deliberate and acceptable "sacrifice". The system accepts being slightly less accurate in predicting impulsive (and potentially ruminative) preferences in exchange for a safer digital environment that protects the user's psychological integrity. A comparison of this trade-off at key vulnerability thresholds is presented in Table 4.

Table 4: Performance comparison at key thresholds

Model	**Vulnerability** (v_u)	**NDCG**	**Harmful exp.** (%)	**Rescue exp.** (%)
Classic Top-N	0.00	0.216	20.9	-
RankAid	0.00	0.216	20.9	57.7
Classic Top-N	0.56	0.216	20.9	-
RankAid	0.56	0.152	0.2	98.6
Classic Top-N	1.00	0.216	20.9	-
RankAid	1.00	0.141	0.0	99.6

The absolute NDCG values of the base model are relatively low compared to the state of the art. However, the goal of this experiment is not to maximise predictive accuracy, but to illustrate degradation dynamics and the RankAid safety mechanism. Thus, the key metric is the relative drop between the baseline and intervened models, which shows that our proposal is mathematically viable.

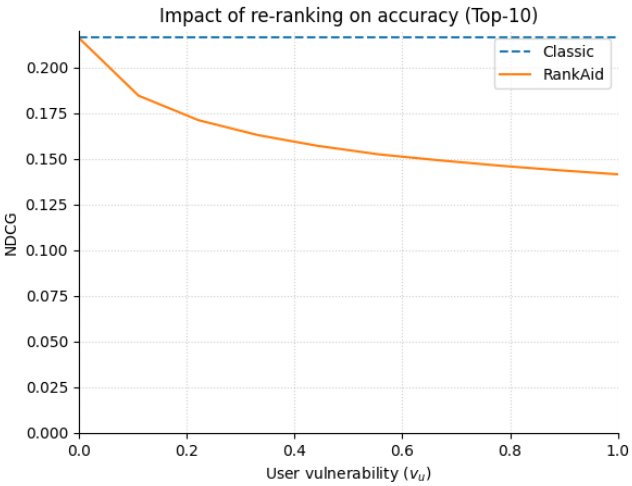


Fig. 3: Controlled degradation of the NDCG metric as the system prioritizes clinical safety. As the individuals state requires more aggressive de-escalation, the algorithm moves away from recommendations based purely on interaction.

α and β calibration: To understand how hyperparameters affect the recommender, we performed a grid search varying the risk penalty α and therapeutic reward β between 0.0 and 1.0. Since the baseline probability $rel_{i,u}$ also falls within this range, these values are sufficient to either neutralize or maximally boost any item. The results are summarized in Figure 4.

The graph shows a steep, non-linear surface peaking when no risk or rescue factors are applied. Yet small changes to α and β reduce risk while preserving acceptable accuracy, indicating that risk suppression and rescue promotion act independently and can be controlled flexibly.

Limitations: While our results show the algorithm is mathematically effective, they are limited by the data. As noted, no public datasets combine recommendation histories with clinical assessments. To compensate, we use an LLM to generate risk and treatment labels from film synopses, which serve only as proxies. **These labels have no real clinical or medical validity** until validated by qualified personnel. In practice, risk and rescue scores must be assigned or validated strictly by healthcare professionals, not solely by automated semantic annotations.

5 Conclusions

To reduce harmful echo chambers in psychologically vulnerable contexts, we propose RankAid, a re-ranking mechanism that balances predictive relevance and clinical safety. Our simulations show four main results:

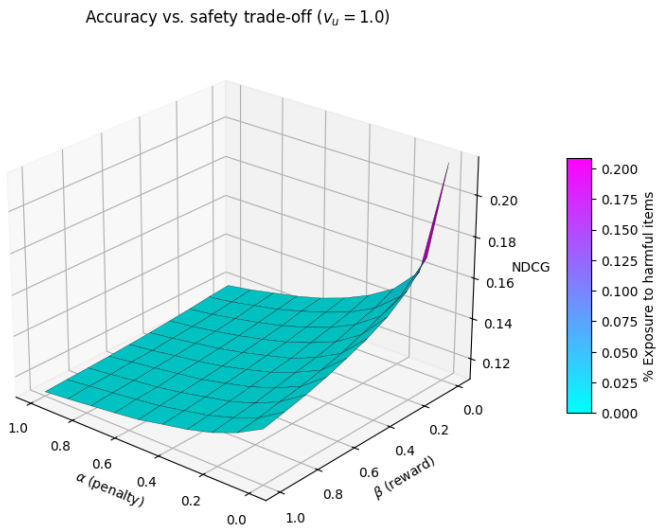


Fig. 4: Pareto front showing the trade-off between accuracy (NDCG) and safety (exposure to harm) for different combinations of α and β during maximum crisis ($v_u = 1.0$).

- **Overcoming the limitations of previous approaches**. The simulations confirm that merely injecting diversity or using traditional fairness metrics is insufficient in clinical contexts. Unlike these methods, RankAid does not just vary the content; it actively intervenes by neutralising risk and prioritising de-escalation.
- **Active de-escalation is viable (and efficient)**. The design of our scoring function is capable of autonomously intercepting echo chambers. By incorporating user vulnerability (v_u) as a dynamic variable, the system manages to reduce exposure to harmful content to zero at the peak of a crisis, replacing it with a set (nearly 100%) of alternatives with therapeutic value.
- **Viable accuracy-safety trade-off**. The slight predictive degradation (measured in NDCG) is confirmed as an acceptable operational cost to ensure a psychologically safe digital environment.
- **Asymmetric hyperparameters**: The functional independence of α and β provides the flexibility to calibrate the recommender's aggressiveness according to the clinical context, overcoming the limitations of zero-sum game approaches.

Our approach shows that recommenders can be modified to stop passively observing user crises. Instead of general content filtering, our algorithm provides personalised interventions. Under stable conditions, the system follows the users usual preferences; when clinical risk indicators appear, RankAid adapts to act as a protective agent.

Although this study focuses on movie recommendations, the method is easily transferable to other domains. In fitness and nutrition apps, for instance, standard algorithms may push vulnerable users toward extreme habits, such as overtraining or severe calorie restriction. Using our method, the system can hide high-risk content (e.g., extreme diets) and promote safer alternatives (e.g., restorative yoga or healthy eating guides) when a user shows signs of distress. Adapting the system to a new platform requires only a reliable way to estimate the risk (ρ_i) and therapeutic value (τ_i) of its items, such as workout plans or nutritional advice.

Acknowledgments. This study is supported by the Dirección General de Investigación e Innovación Tecnológica de la Comunidad de Madrid (Orden 3177/2024) through the I+D Technological activities program (TEC-2024/COM-224).

AI-assisted technologies disclosure During the preparation of this work, the authors used a LLM to assist with the translation and editing of the manuscript to improve language readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Author Contributions The authors are listed in alphabetical order. All authors contributed equally to the research, development, and writing of this article.

Disclosure of Interests. The authors have no relevant financial or non-financial interests to disclose.

References

1. Arendt, F., Scherr, S., Romer, D.: Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. New Media & Society **21**(11-12), 2422–2442 (2019)
2. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on facebook. Science **348**(6239), 1130–1132 (2015)
3. Balagopalan, A., Wang, K., Salaudeen, O., Biega, A., Ghassemi, M.: What's in a query: Polarity-aware distribution-based fair ranking. In: Proceedings of the ACM on Web Conference 2025. pp. 3716–3730 (2025)
4. Calvo, R.A., Milne, D.N., Hussain, M.S., Christensen, H.: Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering **23**(5), 649–685 (2017)
5. Chee, J., Kalyanaraman, S., Ernala, S.K., Weinsberg, U., Dean, S., Ioannidis, S.: Harm mitigation in recommender systems under user preference dynamics. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 255–265 (2024)
6. Coppersmith, G., Ngo, K., Leary, R., Wood, A.: Exploratory analysis of social media prior to a suicide attempt. In: Proceedings of the third workshop on computational linguistics and clinical psychology. pp. 106–117 (2016)
7. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. JMIR mental health **4**(2), e7785 (2017)

8. Gao, R., Ge, Y., Shah, C.: Fair: Fairness-aware information retrieval evaluation. Journal of the Association for Information Science and Technology **73**(10), 1461–1473 (2022)
9. Garcia-Pueyo, L., Kumar Sunkara, V., Senthil Kumar, P., Diwan, M., Ge, Q., Javaherian, B., Verroios, V.: Detecting and limiting negative user experiences in social media platforms. In: Proceedings of the ACM Web Conference 2023. pp. 4086–4094 (2023)
10. Ge, Y., Liu, S., Fu, Z., Tan, J., Li, Z., Xu, S., Li, Y., Xian, Y., Zhang, Y.: A survey on trustworthy recommender systems. ACM Transactions on Recommender Systems **3**(2), 1–68 (2024)
11. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S.: Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. pp. 173–182 (2017)
12. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002)
13. Klimashevskaia, A., Jannach, D., Elahi, M., Trattner, C.: A survey on popularity bias in recommender systems: A. klimashevskaia et al. User Modeling and User-Adapted Interaction **34**(5), 1777–1834 (2024)
14. Nguyen, T.T., Hui, P.M., Harper, F.M., Terveen, L., Konstan, J.A.: Exploring the filter bubble: the effect of using recommender systems on content diversity. In: Proceedings of the 23rd international conference on World wide web. pp. 677–686 (2014)
15. Oak, R., Haroon, M., Jo, C.W., Wojcieszak, M., Chhabra, A.: Re-ranking using large language models for mitigating exposure to harmful content on social media platforms. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 894–908 (2025)
16. Rogers, M.L., Joiner, T.E.: Rumination, suicidal ideation, and suicide attempts: A meta-analytic review. Review of General Psychology **21**(2), 132–142 (2017)
17. Rudd, M.D.: Fluid vulnerability theory: A cognitive approach to understanding the process of acute and chronic suicide risk. In: Ellis, T.E. (ed.) Cognition and suicide: Theory, research, and therapy, pp. 355–368. American Psychological Association, Washington, DC (2006). https://doi.org/10.1037/11377-016
18. Scherr, S.: Social media, self-harm, and suicide. Current opinion in psychology **46**, 101311 (2022)
19. Shing, H.C., Resnik, P., Oard, D.W.: A prioritization model for suicidality risk assessment. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 8124–8137 (2020)
20. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2219–2228 (2018)
21. Valentine, L., DAlfonso, S., Lederman, R.: Recommender systems for mental health apps: advantages and ethical challenges. AI & society **38**(4), 1627–1638 (2023)
22. Wang, H., Sun, Z., Du, Y., Zhang, L., He, T., Ong, Y.S.: Uncertain multi-objective recommendation via orthogonal meta-learning enhanced bayesian optimization. arXiv preprint arXiv:2502.13180 (2025)
23. Ziarani, R.J., Ravanmehr, R.: Serendipity in recommender systems: a systematic literature review. Journal of Computer Science and Technology **36**(2), 375–396 (2021)