This Is The World (Recommender Systems Made It For You)

Capturing Risk Pathways for Children through a Pluralistic Multi-Risk Assessment of Recommenders

Robin Ungruh
Delft University of Technology
Delft, Netherlands
R.Ungruh@tudelft.nl

Maria Soledad Pera
Delft University of Technology
Delft, Netherlands
M.S.Pera@TUDelft.nl

Abstract

Recommender systems increasingly shape children's online experiences, yet the risks they might pose remain poorly understood. Formal risk assessments are rare, overlook children's experiences, treat them as a homogeneous group, and often consider risks on a population-wide level. To provide a more nuanced child-centered perspective on risks associated with recommenders, we conduct a comprehensive risk assessment that examines multiple risks simultaneously and captures individual-level exposure rather than solely aggregated metrics. We also explore behavioral characteristics that make some children more susceptible to risk exposure to grasp how individual differences shape the manifestation of risks. Our results highlight the need for risk assessments that recognize heterogeneity among children and understand when risks become 'systemic' through individualized recommendation patterns. These insights spark the discussion about child-aware recommenders and ongoing regulatory efforts for evidence-based accountability.

CCS Concepts

• **Social and professional topics** → **Children**; • **Information systems** → **Recommender systems**.

Keywords

Systemic Risks, Digital Services Act, Risk Assessment, Risk Profiling

ACM Reference Format:
Robin Ungruh and Maria Soledad Pera. 2026. This Is The World (Recommender Systems Made It For You): Capturing Risk Pathways for Children through a Pluralistic Multi-Risk Assessment of Recommenders. In *34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26), June 08–11, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3774935.3806148

1 Introduction

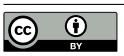
Content Warning: This work discusses online risks for children, including sensitive and potentially disturbing content.

Children (as per UNICEF [120], individuals under 18) frequently encounter content curated by recommender systems (RS), supporting how they can access information. Yet, suggestions may be unsuitable [73, 112, 118] and children's interactions with RS can reinforce problematic behaviors [14, 42, 107]. Thus, it is an ethical but also, in light of existing legislations (e.g., EU's DSA [32], UK's OSA [82], US's COPPA [122]), a legal requirement to assess RS-driven risks to ensure that RS accomodate children's best interests [121].

The Digital Services Act (DSA) [32] presents a guideline for online platforms to assess both the likelihood and severity of children's exposure to **systemic risks**. 'Systemic', however, is missing a clear definition [43, 64], often associated with risk to *large-scale* harm [43, 77] to *societal* functions [13]. This framing is valuable for ensuring that platform-level risks are recognized and mitigated. However, in practice, by centering on *society*, **individuals** are overlooked [77], including children who experience individual forms of risk and are highly heterogeneous. Although the DSA acknowledges children's evolving capacities conceptually [60], existing approaches offer limited guidance on how to account for individual differences in assessing risk. As a result, risks that do not lead to broad impacts for an entire population may be disregarded as non-systemic, although they can have consequences for particular children. Large-scale impact or not, some risks can be clearly considered unsuitable for children and are therefore typically legally prohibited (e.g., pornographic content); others are largely overlooked as they cannot be uniformly assigned impact. Such *conditional risks* may affect children depending on their individual sensitivities. Consequently, the prevailing framing of risks introduces a conceptual gap, as it constrains what counts as a risk and how risks to children are understood. Crucially, this gap leads to risk assessment that, in practice, fail to meaningfully capture the risks to individuals' vulnerabilities.

Besides the lack of consideration for children's heterogeneity and the tendency to treat risks in a binary manner—seeing them as either present or absent, harmful or harmless—current risk assessments are not suited to reflect children's lived realities. Specifically for RS, risk assessments are rarely conducted, and often prioritize compliance-oriented metrics [5] or generic user populations rather than accounting for factors that shape how *children* engage with RS. Moreover, they usually assess a predefined risk [16, 42, 71, 83, 110], rather than exploring how different risks may interact or co-occur; they, thus, neglect the multifaceted nature of RS and ignore that a recommendation can simultaneously encompass a *suite* of risks.

The current state of risk assessments offers a narrow perspective that likely underestimates the breadth and true impact of risks that RS may pose to children. RS actively shape how children see and make sense of the world; their underlying dynamics can surface disturbing content and even have tangible repercussions, especially for children as a vulnerable group [65]. We argue that when children are the primary stakeholders, risk assessments must provide a broader and nuanced account that captures (1) different risks in tandem [98], (2) the spectrum from *established* risks that are generally considered harmful to *conditional* ones, which manifest differently



UMAP '26, Gothenburg, Sweden

ACM ISBN 979-8-4007-2311-7/26/06
https://doi.org/10.1145/3774935.3806148

across children, and (3) pluralistic perspectives that recognize the diversity among children. The notion that children's heterogeneity may lead to differences in risk exposure demands understanding of **who** is more likely to encounter different kinds of risks.

To provide a comprehensive understanding of how RS impact children, we address the research questions (RQs):

RQ1. *To what degree do risks manifest in recommendations made to children?*

RQ2. *Are subgroups of children disproportionately exposed to suites of risks in their interactions with RS, and what behavioral or demographic characteristics make them more susceptible?*

To answer these RQs, we introduce a child-centered risk-assessment methodology. We detect risks across recommended items and construct **risk profiles** that capture the extent of potentially risky items in recommendations to identify a diverse range of risks and quantify their presence within recommendations made to children. By looking at individual exposure, we detect users particularly susceptible to exposure to risks through RS. These profiles *document* outcomes of RS that can be classified as a risk. We make no conclusive statements about the impact from these risk factors, as these are often conditional and depend on contextual aspects. Instead, documentations provide a foundation for relevant stakeholders to infer whether a specific child or group of children could be impacted. Further, we create **profile–risk mappings**, which draw connections between children's individual characteristics and susceptibility to exposure to certain risks, revealing individualized patterns. This allows us to identify sub-groups that are disproportionately susceptible to risks and to investigate which characteristics and behaviors may drive such elevated exposure.

Our exploration is scoped to 'content risks' [63]. With a wide range of online risks for children [63], content risks are a natural and widely discussed starting point. We probe recommendations created by a variety of different recommendation algorithms (RAs) on the Goodreads [128, 129] dataset and create risk profiles from risks inferred from the textual content of recommended books.

Our work contributes (1) a conceptual framework for child-centered risk assessments, including conditional risks, (2) a methodology to capture co-occurring risks, and (3) an empirical exploration of differences in risk exposure between users. We publish our code at https://github.com/rUngruh/2026_UMAP_risk_assessment

2 Background & Related Work

During their interactions with online media, children are subjected to risks. Here, we provide an overview of risk exposure online and how risk assessments fail to address issues comprehensively.

Risks manifest online through exposure to material or interactions that can lead to or promote harmful effects for a child. Livingstone and Stoilova [63] provide a framework of risks children may encounter; however, harmful outcomes from these depend on the child's sensitivities or the context of exposure to a risk. When asking children about unwanted behavior or content they are confronted with online, pornographic content is mentioned most frequently [62]; however, there is a wide range of unsuitable content, and new risks emerge constantly. Often, risk discussions allude to 'systemic' risks; this term comes from an economic view and describes effects on a significant part of a system [108]. While systemic risks are not universally understood when it comes to the DSA [64], systemic risks, by definition, imply large-scale impact on *society* [77]; yet, individual children may encounter risks that are deeply harmful but not widespread enough to be considered systemic. Beyond mere differences in risk exposure, treating an outcome as either harmful or harmless ignores differences in age, experiences, and behaviors; children's exposure to and perception of risks can vary substantially. For example, children with mental health difficulties are especially vulnerable to negative online content [106], and exposure to self-harm material may exacerbate vulnerabilities [94, 106]. Likewise, recommending weight-loss content to children with eating disorders can trigger harmful behaviors, such as relapses [42]. Stereotypes are another prominent risk. Yet, whether they result in harm depends on individual characteristics and contextual factors. While stereotypes are frequent in media recommendations for children [115], their actual impact varies, and stereotypical recommendations may have amplified effects on children identifying with marginalized groups [69, 131].

Risk assessments are critical tools that provide a structured means to identify, evaluate, and document potential harms, allowing auditors and designers to trace how risks manifest in recommendations. Integrating risk assessments into design and oversight processes allows moving from reactive responses to proactive accountability, ensuring that risks are neither overlooked nor dismissed as marginal [18, 123]. Current assessments, however, lack the granularity and child-specific focus needed to capture children's experiences. They rely on generalized user assumptions, ignore developmental differences, and treat risks as static or universal, leaving subtle, conditional forms of harm underexplored and differences between children insufficiently understood. While methods to measure online risks exist [80], algorithm audits are typically limited to specific risks [e.g., 53–55, 115] and general audiences [19]. Holistic assessments on broader effects of algorithmic outcomes on children's development and online experiences are rare, mostly concentrating on content risks [63] while comprehensive analyses rely on qualitative judgments of appropriateness [e.g., 36, 81].

RS users are rarely uniform; users differ in preferences, behaviors, and needs. Prior work shows that demographic attributes such as age or gender affect how RS perform for them [29, 66]; as do behavioral patterns: users who diverge from the mainstream often receive less suitable recommendations [90, 125]. This variability also extends to children: RS differ generally in how well they fare for children compared to mostly adult-dominated mainstream groups [29, 95, 116, 117]. Yet children are far from a uniform group; they progress through distinct developmental stages with evolving cognitive, emotional, and social capacities [47, 60, 124, 126]. Systems explicitly designed for children acknowledge this to some extent—mostly by tailoring to specific age groups [e.g., 34, 58, 59, 85, 88]. However, such differentiation rarely extends further, and risk assessments, in particular, tend to default to broad user categories instead of accounting for the substantial diversity of children.

3 Experimental Framework

To identify risks in top-N recommendations, we conduct an offline RS experiment and analyze risks conveyed in recommendations.

3.1 Risk Scoping

To define an actionable and measurable range of risks, we scope our study in two stages [10, 27]: **assessment goal identification** (i.e., the specific risks considered) and **risk operationalization** (i.e., how considered risks are measured). We scope our exploration to *content risks* [63]. These are frequently discussed concerns regarding RS [73] and online platforms at large [5, 96], more so for children [115, 118], as *harmful* content[1] can have a considerable impact on them [1, 15, 91, 100, 105]. Chosen risks do not provide a complete picture on risks; rather, they are a starting point. Our framework aims to support value pluralism [64], i.e., varying normative judgements of what constitutes a risk, by allowing the adaptation of the scope of risk and measures. We quantify the probability of risks in recommendations. While risk assessments are commonly supposed to also assess the severity [32, 63], for scope, and as severity characterizes consequences [6] that may depend on unknown and unquantifiable individual vulnerabilities, we omit this facet here.

3.1.1 Assessment Goals. Inspired by the DSA's aim for safety through risk assessment, and informed by studies on children's online experiences [62, 63] and a taxonomy of "undesired content"[67], we identify a risk scope reflecting children's lived realities:

- **Sexual Content** (S): Unwanted pornographic content is the most common risk reported by children [62]. Exposure to explicit pornographic content has various negative effects on children's behaviors and attitudes [37, 75]. We consider references to sexual material in erotic contexts, even if it does not depict illegal activities [67]. We also highlight depictions of sexual activities involving minors (S3) as a clearly inappropriate type of content.
- **Graphic Content** (V): Explicit depictions or detailed descriptions of violence, injury, or death that can have psychological and behavioral consequences for children, such as desensitization to violence and normalization of aggressive behavior [86, 103, 109].
- **Hate Speech** (H): Content that directly attacks or demeans people belonging to a certain group [21, 84]. It is among the most prevalent forms of harmful content detection [4, 38, 97].
- **Offensive Language** (O): Profanity, slurs, or otherwise coarse and disrespectful expressions that may not qualify as hate speech, as they are not necessarily intended to degrade a specific group, yet can—particularly when presented without context—still contribute to a toxic or unsafe environment [21, 99, 134].
- **Stereotypical Portrayals** (ST): Content that reinforces biased or reductive representations of individuals or groups [46]. Such portrayals can perpetuate inequality, reinforce social hierarchies, and shape self-perceptions and aspirations in children [11, 69]. The harmfulness of stereotypes often depends on context and user characteristics, making them a prime example of *conditional risks*. Here, we focus on gender (ST_G) and race (ST_R) stereotypes.
- **Self-Harm** (SH): Depictions, descriptions, or discussions of self-harm can normalize or glamorize self-destructive behaviors. Such content has been linked to increased self-harm ideation among vulnerable youth [70, 76, 91]. Even if aiming to raise awareness or provide support, an absence of appropriate framing or content warnings can make these materials harmful, particularly for children with preexisting mental health vulnerabilities [94, 106].

[1]The term 'harm' is not well-defined and is used variably across contexts. Here, we refer to 'harm' as a negative impact of a risk, whereas 'harmful content' relates to the common definition of material that has the *potential* to be harmful [5, 42, 63].

3.1.2 Risk Operationalization. For risk detection, we adopt pre-trained models, which we carefully evaluate for suitability for our task, data, and domain [99]. We favor open-source models (instead of popular content moderation APIs [99]) to support reproducibility.

For S, V, SH, and S3, we utilize a BERT [23]-based `content moderation` model [52]. Probabilities for risk labels are obtained by applying a sigmoid function to the model outputs. Although the content-moderation model provides additional labels (e.g., for H), we rely on dedicated, established classification models for H, O, and ST. For H, we use the *Robust Hate Speech Detection* model [51] by [4]. For O, we use a RoBERTa-based offensive-language classifier [50] by [8]. For ST, we fine-tune a RoBERTa model following [114], using their dataset, but extending it to multi-class prediction. The model outputs both a probability that a sentence contains a stereotype and probabilities for specific types of stereotypes (e.g., gender, race, religion). Besides reporting the probability that a text contains *any* kind of stereotype, we compute an aggregated stereotype probability for *Gender* and *Race* stereotypes by combining the binary stereotype probability with the conditional probability of the class. To quantify the risk likelihood for an item consistently across models, each item i receives a score l_r^i, which is the probability of item i presenting risk r as predicted by the respective model.

While the H and O models are established baselines with well-documented performance, the content-moderation model [52] lacks external validation, and the ST model is adapted by us. We therefore validate these two models to ensure they reliably detect the relevant risks in texts. To quantify model performance in predicting risks related to S, V, SH, and S3, we use the test set employed for evaluating the state-of-the-art OpenAI content moderation API [67]. The model achieves an Area Under the Curve (AUC) above 0.92 across all labels. Additional details are available in our repository. The fine-tuned ST model achieves a macro-F1 score of 0.8755 on the binary stereotype classification test set from the reference model's authors [114], aligning with their reported results. For multi-class classification, it reaches a macro-F1 of 0.8413, indicating strong performance in distinguishing stereotype types.

3.2 Dataset

We do not limit our exploration to child-only datasets to emulate a real-world setting, where children are part of the user base, but not the sole users. However, datasets with explicit information about child users are rare [116] and those available often lack comprehensive metadata about the items interacted with or do not have a sufficiently large number of children to reliably capture individual differences. We use **Goodreads (GR)**, a large-scale dataset that includes bookshelves from the Goodreads website, where we identify users showing child-oriented interaction patterns. We use the book descriptions as a proxy for the book's content to detect risks.

Pre-processing. To make GR suitable for our exploration, we only consider user-book interactions where the book was marked as 'read' and we aggregate different editions of the same book. We remove users who have rated fewer than 5 items positively (rating > 3) to ensure that we can make meaningful inferences about users' consumption behavior. We do not apply further filtering to avoid

Table 1: Dataset Characteristics of GR.

	# Users	# Items	# Interactions
Filtered dataset (non-binarized)	733,598	1,506,329	72,387,373
Training Sample	9,999	202,978	973,439
Evaluation Sample (Child-Only)	34,383	160,525	3,466,784

distorting the users' behaviors by removing interactions. Thus, although 16% of items lack descriptions that are required for risk assessment, we retain these items, as filtering would skew users' interaction patterns. We reserve 10% of each user's interactions as **holdout** for validation and testing; the remaining data serves as the **user profile**, which we explore to identify child users and which serves as the foundation for the recommendation experiment.

Identifying child user profiles. We identify users likely to be children based on their overall child-content consumption across all interactions in their user profiles. Assessing user demographics as a proxy based on previous behavior is a common approach in the absence of real demographics [102, 111, 115, 133]. From the 1,506,3295 items in GR, 659,286 are assigned at least one of 8 main genres, with 124,082 being assigned the genre 'children' . For each user, we calculate the proportion of consumed items that are tagged with the 'children'-genre. Users whose proportion of child-content consumption exceeds two standard deviations above the mean proportion across all users are classified as children. This approach captures users whose engagement with children's content is magnitudes higher than average. Using this criterion, approximately 4.75% of users are classified as children, a fraction comparable to that observed in other RS datasets across various domains [116, 117]. By using a threshold based on the mean and standard deviation rather than an arbitrary cutoff, we ensure that only users with disproportionate interest in children's content are labeled as child users. This approach enables an examination of behaviors associated with child-centered content without requiring explicit demographic data.

Splitting. As commonly done due to GR's size [30], we sample 10,000 users to reduce GR to a manageable but meaningful size for training RAs. As random sampling tends to fail in yielding generalizable results [89] and may overlook subgroups, we conduct stratified sampling to ensure the sample remains representative across key behavioral facets (number of interactions, mean rating, rating standard deviation). The interactions of sampled users are used for hyperparameter tuning and model training. The remaining child profiles—not included in the sampled subset—serve as evaluation set, following strong generalization principles [61, 68]. Remaining adult profiles are excluded from evaluation, as our risk assessment centers on children. Table 1 shows dataset characteristics.

3.3 Characterizing Risks in Recommendations

For our study, we consider a range of RAs: `Popularity` as a **non-personalized** RA which recommends the most popular unseen items; the nearest neighbours approach `ItemKNN` [93], the linear model `SLIM` [78], and the neural model `RecVAE` [104], as **personalized** RAs. We deploy all models using the RecPack library [72].

Recommendation Pipeline. We binarize interactions on GR, only treating items rated > 3 as positive interactions. For hyperparameter tuning, we train RAs on the sampled user profiles and evaluate them on their respective holdout interactions. We use Tree-Parzen Estimator targeting *nDCG*@10 with early stopping after five stagnant evaluations [35]. Explored hyperparameter ranges, along with the selected configurations, are available in the repository. Once the optimal hyperparameters are identified, we train the model on the sampled users and evaluate it for the held-out children. For testing, we use the evaluation sample of child users as a fold-in set (unseen during training) to generate recommendations with the tuned model. We evaluate the RAs' performance in recommending items from their holdout interactions, following "strong generalization" principles during testing [61, 68]. This ensures that RAs are trained on a representative sample, while testing and conducting risk assessment is done on a previously unseen sample of children. We predict the top 25 recommendations per user for each RA. For non-deterministic RAs, we repeat each training procedure 5 times with different random seeds and compute metrics and risk profiles as the mean across runs. This ensures that results are not artifacts of a single stochastic training run [9, 28, 132].

Risk Profile Construction. We create risk profiles of top-N recommendations R_u created by a RA for each user u. Each risk profile is a vector, where each component reflects the extent of one of the risks (Section 3.1) across all recommendations. For this, we gather the risk likelihood l_r^i per risk r of each item $i \in R_u$ by applying the risk detection models outlined in Section 3.1.2 to item descriptions. We aggregate l_r^i values for all $i \in R_u$ per risk r to create a vector component. We employ 4 **aggregation strategies** to capture different perspectives on how individual items contribute to the risk likelihood of an entire recommendation list, thereby providing a comprehensive and robust understanding of risks in recommendations. Unless reported otherwise, we report the mean across child-users (marked by omitting superscript u).

- **Average Risk (AR_r^u):** The mean likelihood of r across all items recommended to u, providing an overall measure of exposure intensity: $AR_r^u = \frac{1}{|R_u|} \sum_{i \in R_u} l_r^i$.
- **Quadratic-average Risk (QR_r^u):** The quadratic mean likelihood across all recommendations of u, emphasizing items with high risk likelihood: $QR_r^u = \sqrt{\frac{1}{|R_u|} \sum_{i \in R_u} (l_r^i)^2}$.
- **Presence of Risk (PR_r^u):** Indicates whether r appears in the recommendations of u, capturing the highest likelihood of that risk among all recommended items, reflecting worst-case exposure—the most risky item shown. Formally, $PR_r^u = \max_{i \in R_u} l_r^i$.
- **Weighted Risk (WR_r^u):** Since children are more likely to interact with items appearing at the top of the list [25, 45, 74], this metric accounts for the ranking position of recommendations, giving higher weight to risks appearing near the top of the list analogous to the relevance weighting used in *nDCG* [25, 45, 74]. Here, $\text{rank}(i)$ denotes the position of i in the recommendation list of u (1 for the top item). It is defined as:

$$WR_r^u = \frac{\sum_{i \in R_u} \frac{l_r^i}{\log_2(1+\text{rank}(i))}}{\sum_{j=1}^{|R_u|} \frac{1}{\log_2(1+j)}}$$

3.4 Characterizing Susceptible Users

We create *profile–risk mappings*, which connect exposure to specific risks with the users' behaviors. To identify characteristics of users particularly exposed to certain risks, i.e., those highly **susceptible**

to a risk, we analyze differences in exposure according to QR_r^u, as it balances the high impact of individual risks and the overall presence of risks throughout a recommendation list. We z-score normalize QR_r^u, denoted as $\tilde{QR}_r^u$, which expresses each user's risk exposure in terms of standard deviations from the mean across users per risk, enabling comparability across risks. We categorize a user as *susceptible* to risk r if $\tilde{QR}_r^u > 2$, meaning the user's recommendations receive a risk score that deviates more than two standard deviations from the mean. We describe users through certain **behavioral features** derived from their user profiles:

- **Interaction features:** We gather interaction data such as the number of consumed items (`NumInts`) and the popularity of each item (`Pop`), measured as the min-max normalized logarithm of the number of interactions with an item in the user profiles of users sampled for training, i.e., the popularity in the train set.
- **Genre consumption:** Genre coverage (`Cov`) measures the proportion of different genres out of all 8 in a user profile, `Gini` measures the balance between different genres. Further, we compute the proportion of each genre among all items consumed by a user that were tagged with a genre; we denote these as $P(\text{Children})$, $P(\text{Fantasy})$, etc.
- **Risk consumption:** Using risk detection methods from Section 3.1.2, we identify to what degree children have consumed items that posed risks. We compute the risk likelihood for each item in a child's user profile as before and determine the average risk likelihood of consumed items per user: e.g., $\bar{C}(H)$, $\bar{C}(ST)$.

4 Results & Analysis

Here, we present results from our experiments.

4.1 RQ1: Analysis of Risk Profiles

Risk profiles in Table 2 show the degree to which RAs differ based on aggregated risk likelihoods. AR_r, QR_r, and WR_r scores are consistently low, indicating that recommendations generally pose little risk, except for notably higher `O` and `ST` scores. Across all RAs, PR_r values (indicating the highest likelihood for r across recommended items) exceed the scores of other aggregation metrics, evidencing that while individual items rarely contain risks, the probability that

Table 2: Average risk profiles across all users per RA.

	Algorithm	H	O	S	S3	SH	ST	ST_G	ST_R	V
AR_r	Popularity	.003	.223	.000	.000	.000	.031	.003	.000	.002
	ItemKNN	.006	.220	.001	.001	.000	.035	.010	.001	.002
	SLIM	.004	.227	.001	.001	.000	.033	.009	.000	.002
	RecVAE	.003	.227	.001	.000	.000	.043	.015	.000	.001
QR_r	Popularity	.005	.249	.001	.000	.000	.055	.011	.000	.008
	ItemKNN	.013	.248	.004	.004	.000	.080	.044	.002	.009
	SLIM	.008	.251	.004	.006	.000	.080	.039	.001	.009
	RecVAE	.004	.251	.002	.002	.000	.099	.066	.001	.005
WR_r	Popularity	.003	.233	.000	.000	.000	.033	.002	.000	.001
	ItemKNN	.006	.222	.001	.001	.000	.034	.009	.000	.002
	SLIM	.004	.228	.001	.001	.000	.032	.008	.000	.002
	RecVAE	.003	.228	.001	.000	.000	.043	.013	.000	.001
PR_r	Popularity	.021	.483	.004	.001	.000	.185	.053	.002	.040
	ItemKNN	.053	.490	.017	.018	.001	.335	.211	.010	.041
	SLIM	.035	.486	.018	.028	.001	.336	.192	.004	.043
	RecVAE	.013	.482	.009	.009	.000	.412	.314	.002	.025

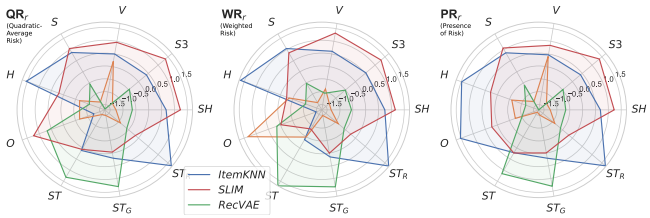


Figure 1: Risk scores (z-score normalized across RAs).

at least one risky item appears in top-N recommendations is substantially higher. Ultimately, risks do not emerge uniformly across recommendations, but specific items in the list may surface risks.

As shown in Fig. 1, the tendency to risk exposure varies across RAs. For each r, PR_r^u scores differ significantly between RAs according to one-way ANOVAs followed by pairwise Welch's t-tests with Bonferroni correction ($p < .05$); except for `SLIM` and `ItemKNN` for `ST`, `Popularity` and `RecVAE` for `O`, and `ItemKNN` and `Popularity` for `V`. The effect sizes of the statistical differences, however, remain small: η^2 does not exceed 0.13 for any r and Cohen's d stays smaller than ±0.4 for all r except `ST` and ST_G, where it remains smaller than ±1.25 for all RA pairs. For these two risks, `RecVAE` shows significantly higher scores than any other RA, despite it typically receiving lower risk scores than the other personalized RAs for other r. These results suggest that, although RAs differ in how frequently risks appear in their top-N recommendations, risk exposure is a consistent and systematic pattern of all RAs. This trend is evident across all aggregation strategies: significant differences exist; however, effect sizes are typically negligible. As RAs vary in the degree to which they present risks, they also show trade-offs in traditional evaluation criteria without showing a clear connection to risk exposure: according to the performance metrics *nDCG* and *HIT*, personalized approaches outperform `Popularity`, with `ItemKNN` and `SLIM` yielding the highest scores (Table 3).

Ultimately, the average prevalence of risks is low, but risk exposure is highly uneven across users. Most users encounter items with near-zero risk likelihoods, but a small fraction experience disproportionate exposure. The cumulative exposure curve of PR_r^u scores across users in Fig. 2 depicts this pattern—users are sorted from lowest to highest scores, and the curve illustrates the proportion of cumulative PR_u^r accounted for as more users are considered. For most risks, it stays nearly flat for the majority of users and rises sharply for the top 10–20%, indicating that these users are collectively responsible for most of the cumulative exposure and therefore experience disproportionate levels of risk relative to the rest of the user base. Notable exceptions are `O` and `ST`; their curves reflect a more even distribution of risk scores. Individual differences

Table 3: Performance scores per RA.

	nDCG		*HIT*	
	Mean	Std	Mean	Std
`Popularity`	0.08	0.14	0.39	0.49
`ItemKNN`	0.17	0.20	0.63	0.48
`SLIM`	0.19	0.21	0.64	0.49
`RecVAE`	0.13	0.18	0.54	0.49

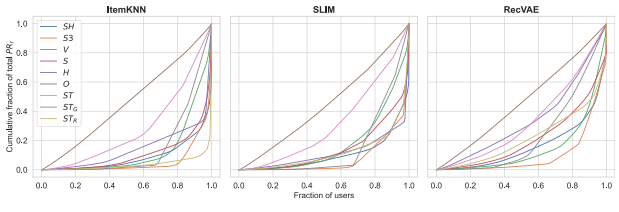


Figure 2: Cumulative PR_r^u scores across users; PR_r^u denotes the maximum likelihood of r in recommendations to u.

and disproportionate exposure are not statistical outliers; they reflect a consistent trend across risks, RAs, and aggregation strategies. Even when a risk is statistically rare, it can produce substantial individual-level exposure for a small subset of users. In other words, low overall prevalence does not imply negligible exposure.

To investigate how risks accumulate at the individual level, we probe whether users exposed to one risk are concurrently exposed to others with correlations between PR_r^u scores for different risks within user profiles (Fig. 3). Correlations are generally low per RAs, with moderate correlations between O and H and strong ones between S and S3, as well as ST and ST_G. Low correlations underscore the independence of risks: top-N lists do not pose uniform risk across categories; exposure to one risk by an RA does not imply exposure to others. Moreover, correlation analyses of average PR_r^u across all r show moderate correlations among personalized RAs ($.3 < \rho < .4$) and weak ones with Popularity ($\rho \leq .2$), indicating that RAs do not follow a universal exposure pattern. This trend is further reinforced when comparing risk-specific exposure.

Overall, these findings highlight that while some prominent risks emerge in recommendations, individualized assessment is crucial, as aggregate metrics may underestimate the exposure experienced by individuals more strongly affected by particular risks. Risks do not manifest uniformly, and there is no such thing as a 'typical' child profile. Risk exposure to one risk cannot be reliably extrapolated and be predictive for others. Instead, risks emerge in diverse ways, and children are exposed to risks through unique patterns.

To illustrate why looking at individual differences in risk exposure matters, consider two children from GR (C1 and C2) with similar exposure to sexual content by ItemKNN ($PR_S^{C1} = 0.976$, $PR_S^{C2} = 0.912$). The RA recommends to both at least one item that is highly likely to include sexual remarks. Closer inspection of risk profiles reveals nuanced differences. C1 has mostly consumed *Children* or *Romance* books; their recommendations reflect this pattern, resulting in the highest PR_S^u among all users. Exposure to sexual remarks in Romance may reinforce problematic portrayals—like

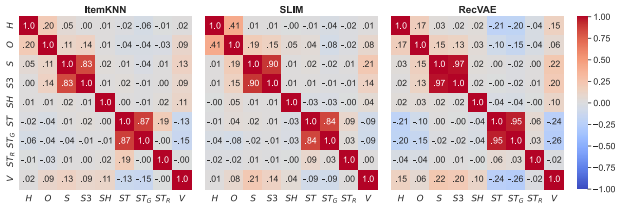


Figure 3: Correlation of PR_r^u (presence of risk) between risks.

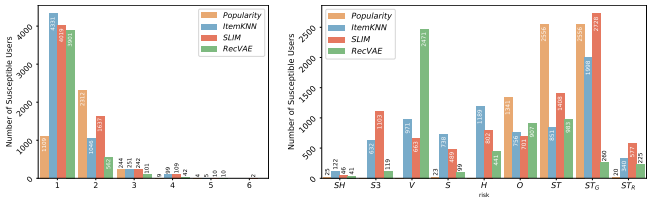
traditional power dynamics in relationships [12, 113]—but whether such exposure has an impact on the child is conditional, depending on the child's maturity, and prior experiences. By contrast, C2's interests extend beyond children's books to *History & Biography*. However, they receive the highest PR_{S3}^u score ($PR_S^{C2} = 0.97$; for comparison, $PR_S^{C1} = 0.069$). This is driven by the recommendation of *'The Waiting'* by Cathy LaGrow, a historical account containing explicit references to the rape of a 16-year-old. Although the impact of such content will vary across children, such themes may lead to stronger psychological or emotional distress than comparatively mild sexual remarks in C1's recommendations. This example showcases the importance of capturing a range of risks for *meaningful* risk assessments. Broad categories, such as S alone, could not account for the differences between the recommendations.

4.2 RQ2: Capturing Profile–Risk Mappings

Individual differences in risk exposure underscore the need to understand not only *what* risks are present in recommendations, but also *who* they affect. To determine those highly exposed, we identify *susceptible* users as described in Section 3.4: 16.7%, 17.5%, and 13.5% of users are susceptible to at least one risk for ItemKNN, SLIM, and RecVAE, respectively; many of them to more than one (Figure 4a). Figure 4b captures the specific risks users are susceptible to. To compare differences in *behavioral features* (§ 3.4) between susceptible and non-susceptible users for each risk, we conduct independent t-tests on every feature (with Bonferroni correction) per RA. Most behaviors are significantly different ($p < .05$), yet most effects are small in magnitude. On SLIM, only 17 differences have an effect size (Cohen's d) exceeding 0.5 (RecVAE: 29; ItemKNN: 23).

Table 4 reports the significant associations with largest effect sizes for personalized RAs, highlighting behaviors most linked to risk exposure. For ItemKNN, previous consumption of risky items is predictive of exposure to the respective risk in recommendations. Users susceptible to H also consume more H content. Similar effects emerge for ST, ST_R, and O. Exposure to ST_R, ST, and SH is associated with consumption of **less** popular items, signaling that engagement with niche content may inadvertently amplify risks in this neighborhood-based approach. Niche items containing risks may cluster together in item neighborhoods, increasing the likelihood of propagating these risks to users engaging with them.

In contrast to the simpler neighborhood-based strategy, for SLIM and RecVAE, thematic associations tend to predict future risk exposure. Most saliently, children susceptible to SH showed significantly greater consumption of *young adult* content, suggesting that RAs



(a) Count of users susceptible to a given number of risks.

(b) Number of users susceptible to each r.

Figure 4: Susceptible users for $\tilde{QR}_r^u$ (quadratic-average risk).

Table 4: Most salient behavioral differences between susceptible and non-susceptible users per r.

ItemKNN					SLIM					RecVAE				
Risk	**Mean Risk Diff.**	**Behav.**	**Mean Behav. Diff.**	**Eff. Size**	**Risk**	**Mean Risk Diff.**	**Behav.**	**Mean Behav. Diff.**	**Eff. Size**	**Risk**	**Mean Risk Diff.**	**Behav.**	**Mean Behav. Diff.**	**Eff. Size**
H	0.22	$\bar{C}$(H)	0.03	1.97	SH	0.05	P(Young adult)	0.23	1.84	SH	0.01	P(Young adult)	0.30	2.45
ST	0.18	$\bar{C}$(ST)	0.06	1.76	ST_R	0.02	P(history/biography)	0.15	1.16	S	0.12	P(Mistery/Thriller/Crime)	0.15	1.74
ST_R	0.15	$\bar{C}(ST_R)$	0.01	1.30	S3	0.10	P(Mistery/Thriller/Crime)	0.072	0.83	V	0.03	P(Young adult)	0.18	1.56
ST_R	0.15	Pop	-0.19	-1.24	H	0.19	$\bar{C}$(O)	0.03	0.75	S3	0.13	P(Mistery/Thriller/Crime)	0.13	1.48
O	0.09	$\bar{C}$(O)	0.04	1.16	H	0.19	P(Children)	0.13	0.70	ST_R	0.02	NumInts	225.30	1.10
ST	0.18	Pop	-0.17	-1.11	S	0.10	P(Mistery/Thriller/Crime)	0.06	0.69	ST_G	0.15	Cov	0.27	1.09
ST	0.18	$\bar{C}$(O)	-0.04	-1.06	ST_R	0.02	P(Fantasy/Paranormal)	-0.09	-0.66	ST_G	0.15	NumInts	206.58	1.01
SH	0.02	Pop	-0.16	-1.03	SH	0.05	Cov	-0.16	-0.66	H	0.01	P(Fantasy/Paranormal)	0.14	1.01
ST	0.18	Cov	-0.21	-0.87	SH	0.05	P(Fantasy/Paranormal)	-0.09	-0.66	ST_R	0.02	Cov	0.25	1.00
ST_R	0.15	$\bar{C}$(ST)	0.03	0.79	S	0.10	P(Romance)	0.05	0.64	H	0.01	P(Young adult)	0.11	0.94

For each RA, the table presents the behavioral features that most saliently differentiate susceptible users from non-susceptible ones across risks. "Mean Risk Diff." reports the average difference in standardized exposure ($\tilde{QR}_r^u$) between the two groups for risk r. "Behav." shows the behavioral feature significantly differing and "Mean Behav. Diff." the average magnitude of the difference between the groups. "Eff. Size" reports the strength of the difference (Cohen's d). Absolute values and results for all features available in the repository.

learn associations between thematic themes in young adult books and those present in self-harm-themed material. Although this association may emerge from shared developmental or emotional themes—e.g., identity struggles, mental health, and relationship dynamics [49, 57]—prior consumption of SH is **not** predictive of the system's subsequent self-harm exposure. Cohen's d for the difference in $\bar{C}$(SH) shows that, despite significance, the effect is trivial in magnitude (SLIM: $d = -0.11$; RecVAE: $d = 0.33$), indicating minimal behavioral difference between susceptible and non-susceptible users in terms of SH content consumption. This suggests that the RA may expose children to self-harm content based on general thematic affinities rather than explicit interest in or engagement with such material, potentially broadening exposure beyond those who actively seek it. Similarly, racial stereotypes (ST_R) are disproportionate in SLIM's recommendations to users who read History books. As history books often depict real-world events and social hierarchies, related books may surface content reflecting racial stereotypes [39], creating a link that RAs detect and amplify. Such thematic associations between prior consumption and risk exposure in recommendations are found for various risks for these RAs.

Overall, disproportionate risk exposure is associated with certain characteristics of the user profiles. For ItemKNN, the RA "correctly" identifies these preferences and exposes the users to similar content. However, risk amplification is a broader pattern that not only encompasses RAs mimicking what the user consumed and trapping them in a loop of risky content. RAs actively introduce these risks through implicit thematic associations, potentially kicking off a spiral of risks and harmful outcomes.

5 Discussion

Together, outcomes from RQ1 and RQ2 (§4) reveal that there is no such thing as a uniform child risk profile, challenging the view of the abstract entity of the 'user'; instead, risk assessments must consider individual and unique experiences [56]. Some risks appear consistently in recommendations, others are amplified due to the RA exploiting specific preferences and previous consumption behavior. Through thematic associations, RAs can suggest content that seems fitting but inadvertently propagates risks. This is alarming when disproportionately affecting those most vulnerable or navigating difficult experiences, e.g., as is the case for teenagers. Population-level summaries alone are insufficient: who is exposed to what kind of content matters just as much as the presence of the risk itself.

Risk propagation does not arise from deliberate RAs decisions. Instead, it reflects *hidden* biases [20] that emerge as the RAs learn associations present in the underlying data. The only information implicitly available to the RAs is derived from user–item interactions (§3.4); all other information is item-level features that the RAs cannot access. Still, interaction patterns can encode relationships between items. Due to the large number of items in the dataset, collaborative filtering RAs must operate on sparse overlaps between user profiles, leading them to rely on the few co-occurring signals they observe. Risky items within sparse intersections can become disproportionately influential, as RAs lacks alternative evidence, thereby reinforcing exposure to risks. Surprisingly, Popularity suggests markedly less risky content, as it draws from a small pool of globally high-frequency, mainstream items. This concentration limits exposure to the broader catalogue, including niche or thematically specific books where risks may be more prevalent. Hence, the lack of personalization acts as a buffer against risk exposure. However, this effect may be domain-dependent: popularity in social media or news is often driven by engagement rather than inherent quality or safety, potentially allowing popularity-based RAs to amplify sensational or misinformation content [41, 83].

Building on individualized risk profiles to form evaluation lenses that complement (and potentially challenge) traditional relevance metrics, the profile–risk mappings we introduce offer a technique for *diagnosing* why risk exposure emerges for individuals. Crucially, our approach works without model introspection, making it applicable even to black-box RS. The result is a substantive methodological contribution: by capturing both users who are disproportionately affected by risks and the facets that shape exposure, design properties of RAs driving these outcomes can be better understood, enabling comprehensive evaluation protocols, mitigation strategies, and design principles that explicitly incorporate user-specific safety considerations. While risk profiles and profile–risk mappings serve as analytical tools here, the framework provides a meaningful foundation for many stakeholders. For researchers, it enables systematic probing of algorithmic mechanisms underlying risks.

For practitioners, individualized risk diagnostics offer actionable signals for targeted interventions. For policymakers and auditors, our approach offers a transparent and reproducible methodology to document risks in deployed RS and act and regulate accordingly.

Toward Impact Assessments. Some children are more susceptible to exposure to recommendations conveying certain risks (*RQ2*). While susceptibility alone does not imply harm, its intersection with a child's vulnerabilities can magnify potential impacts. Those affected are not IDs in a dataset but real children in a critical developmental phase. Children's experiences and vulnerabilities are shaped by intersectional factors such as gender, ethnicity, disability, and socioeconomic background [17]. These identities and lived experiences may affect how risks are perceived and experienced [87], leading to unique experiences that broad assessments overlook. Assessing these vulnerabilities is inherently challenging, and intersectional factors are often not represented in datasets with sufficient granularity to support comprehensive studies [130]. Understanding vulnerabilities is a key first step toward **impact assessments**, which aim to predict a system's effects on its users and are required to argue that a risk can lead to actual harm. These assessments require not only understanding the RS and the risks it generates, but also *who* is affected and in *what (social) context*.

The Role of the DSA. The DSA provides a legal foundation to establish accountability, with Article 28 explicitly requiring online platforms *accessible* to minors—not only those designed specifically for children—to put appropriate and proportionate measures in place to ensure privacy, safety, and security for children [32]. However, in risk assessments reports, a persistent gap between regulatory intent and platform practice remains [24, 31, 79]. Platforms recognize risks stemming from RS—e.g., as per internal documents, TikTok is aware of systemic risks to children [3]—yet these findings are often omitted [24] or downplayed [31] in official reports.

The DSA faces criticism given the current requirements to assess risks, which remain non-standardized and open to selective interpretation [26, 43, 79]. Missing definition of 'systemic risk' [64] and lack of practical guidance may limit the ability to measure real-world impacts and hold platforms accountable [26]. Our work highlights that not all potentially consequential risks would be conventionally considered systemic. Risks may be conditional and only affect users with specific vulnerabilities or arise from particular usage. Recognizing these 'localized' risks is essential for comprehensive evaluation and informed regulatory oversight, since when certain users are susceptible to specific risks mediated by an RA, then this is, in fact, a systemic effect. The DSA has the potential to shape the relationship between users and online platforms [77], empowering users and moving toward sociotechnical considerations in the design of RS [33, 48]. For this, however, value pluralism, acknowledging the diversity of what may constitute a risk and criticizing their outcomes from varying lenses, is necessary [64]. Based on our exploration, we call for a rethinking of what is considered a systemic risk—from society to individuals; our framework is a starting point for this that can accommodate various normative definitions of risk and interpretations of risk measures.

The Path Ahead. As online platforms perform a public function, their users have the right that RS respect their interests and fundamental rights [77]. For children, this means *access to information* [119] and ensuring that the *best interests of the child* are a primary consideration [121]. RS can support this by protecting children from harm and supporting their developmental, educational, and psychological well-being and needs. In practice, this dual responsibility requires balancing informational rights with proactive measures to mitigate risks in recommendations. Protecting children while accommodating their evolving capacities and supporting their agency [126] may require re-thinking mitigation goals and recognizing the conditional nature of risks. Rather than treating recommendations as a binary 'suitable or unsuitable' construct, mitigation strategies should adopt a user-centered perspective. For example, De Toni et al. [22] address exposure to 'unwanted' recommendations, illustrating a shift from broad 'inappropriateness' toward understanding recommendations as outcomes that may be misaligned with user needs or preferences. From this perspective, age assurance and age restrictions are only one component of safeguarding strategies, and even these must be handled carefully [101]. They can help set baseline protections, but they cannot capture the diversity of children. Thus, risk assessments accounting for individual sensitivities to risks and diverse users must be a foundation for meaningful protection.

Limitations. Our risk scope provides an initial overview; expanding this scope and creating reliable risk detection methods demands interdisciplinary efforts, including NLP research and social sciences. However, building high-quality datasets to train such methods is complex and raises ethical concerns [127]. We draw on models trained on social-media data [44, 92], reflecting the focus of risk-detection research and widespread use of off-the-shelf models for reproducibility [5, 7, 99]. Although not explicitly validated on item descriptions like the ones used by us, these models offer a practical and established starting point [40]. Still, metrics may behave differently when applied to book descriptions. Hence, comparisons across metrics should be interpreted with caution. Further, descriptions are only a proxy for full book content, an established practice in prior work on modeling book characteristics [e.g., 2, 56], which may result in a conservative estimate of risk, as further risk exposure could emerge throughout the full narrative. We also retain items with empty descriptions to preserve all behavioral facets, but they may distort aggregated risk scores, as they default to predicting no risk. Our risk scores may thus be underestimates.

Our study does not capture real user demographics. To our knowledge, no dataset that includes demographics is suitable for this study, as they either do not include enough child-users to detect individual differences, or as item risks cannot be assessed due to a lack of item descriptions or other multimedia content. In reality, the absence of demographics is a common issue; using behavioral features as proxies is oftentimes the best alternative [102, 111, 133].

6 Concluding Remarks

RS shape what children see, learn, and normalize. Narrow relevance metrics for evaluation are no longer sufficient. We must shift toward pluralistic human-aligned assessments recognizing diverse risks, individual patterns, and real-world consequences of algorithmic choices, so that RS support, rather than undermine, children's rights and well-being. Our work lays the groundwork for this shift by demonstrating how risk assessment can uncover the mechanisms through which RS propagate risks affecting children.

Acknowledgments

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-14461.

The authors used generative AI tools for grammar and spelling checks and other light editing. The authors take full responsibility for the publication's content.

References

[1] Vivek Agarwal and Saranya Dhanasekaran. 2012. Harmful effects of media on children and adolescents. *Journal of Indian Association for Child and Adolescent Mental Health* 8, 2 (2012), 38–45.

[2] Charu C Aggarwal. 2016. Content-based recommender systems. In *Recommender systems: The textbook*. Springer, 139–166.

[3] Bobby Allyn, Sylvia Goodman, and Dara Kerr. 2024. *TikTok executives know about app's effect on teens, lawsuit documents allege.* https://www.npr.org/2024/10/11/g-s1-27676/tiktok-redacted-documents-in-teen-safety-lawsuit-revealed

[4] Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics, Toronto, Canada, 231–242. https://aclanthology.org/2023.woah-1.25

[5] Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, et al. 2023. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *Comput. Surveys* 56, 3 (2023), 1–17.

[6] Terje Aven and Ortwin Renn. 2009. On risk defined as an event where the outcome is uncertain. *Journal of risk research* 12, 1 (2009), 1–11.

[7] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the fourth workshop on online abuse and harms*. 125–137.

[8] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

[9] Christine Bauer, Alan Said, and Eva Zangerle. 2024. Evaluation Perspectives of Recommender Systems: Driving Research and Education (Dagstuhl Seminar 24211). *Dagstuhl Reports* 14, 5 (2024), 58–172.

[10] Lex Beattie, Dan Taber, and Henriette Cramer. 2022. Challenges in translating research to practice for evaluating fairness and bias in recommendation systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 528–530.

[11] Rebecca S Bigler and Lynn S Liben. 2006. A developmental intergroup theory of social stereotypes and prejudice. *Advances in child development and behavior* 34 (2006), 39–89.

[12] Amy E Bonomi, Emily M Nichols, Christin L Carotta, Yuya Kiuchi, and Samantha Perry. 2016. Young women's perceptions of the relationship in Fifty Shades of Grey. *Journal of women's health* 25, 2 (2016), 139–148.

[13] Sally Broughton Micova and Andrea Calef. 2023. Elements for effective systemic risk assessment under the DSA. *Available at SSRN 4512640* (2023).

[14] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*. 160–169.

[15] Hrishita Chakrabarti, Diletta Micol Tobia, Monica Landoni, and Maria Soledad Pera. 2025. Online Information Disorder & Children. *Proceedings of ROMCIR* (2025), 24–38.

[16] 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.

[17] Patricia Hill Collins. 2015. Intersectionality's definitional dilemmas. *Annual review of sociology* 41, 1 (2015), 1–20.

[18] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.

[19] Shakked Dabran-Zivan, Ayelet Baram-Tsabari, Roni Shapira, Miri Yitshaki, Daria Dvorzhitskaia, and Nir Grinberg. 2023. "Is COVID-19 a hoax?": auditing the quality of COVID-19 conspiracy-related information and misinformation in Google search results in four languages. *Internet Research* 33, 5 (2023), 1774–1801.

[20] Savvina Daniil, Mirjam Cuper, Cynthia Liem, Jacco van Ossenbruggen, and Laura Hollink. 2022. Hidden Author Bias in Book Recommendation. In *Proceedings of FAccTRec: Workshop on Responsible Recommendation 2022*. Available at: https://doi.org/10.48550/arXiv.2209.00371.

[21] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 512–515.

[22] 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*. 492–502.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[24] DSA Civil Society Coordination Group. 2025. *Initial Analysis on the First Round of Risk Assessment Reports under the EU Digital Services Act.* Technical Report. 5Rights Foundation. https://5rightsfoundation.com/wp-content/uploads/2025/03/Initial-analysis-on-the-first-round-of-Risk-Assessment-Reports-under-the-DSA.pdf [Accessed 30-09-2025].

[25] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. 2010. Query log analysis in the context of information retrieval for children. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 847–848.

[26] Niklas Eder. 2024. Making systemic risk assessments work: how the DSA creates a virtuous loop to address the societal harms of content moderation. *German Law Journal* 25, 7 (2024), 1197–1218.

[27] Michael D Ekstrand, Lex Beattie, Maria Soledad Pera, and Henriette Cramer. 2024. Not just algorithms: Strategically addressing consumer impacts in information retrieval. In *European Conference on Information Retrieval*. 314–335.

[28] Michael D Ekstrand, Ben Carterette, and Fernando Diaz. 2024. Distributionally-informed recommender system evaluation. *ACM Transactions on Recommender Systems* 2, 1 (2024), 1–27.

[29] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on fairness, accountability and transparency*. PMLR, 172–186.

[30] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM conference on recommender systems*. 242–250.

[31] Eurochild. 2025. *Online platforms' risk assessment reports under the DSA fall short on children's protection.* https://eurochild.org/news/online-platforms-risk-assessment-reports-under-the-dsa-fall-short-on-childrens-protection/

[32] European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065

[33] Matteo Fabbri and Ludovico Boratto. 2025. Auditing Recommender Systems for User Empowerment in Very Large Online Platforms under the Digital Services Act. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 51–61.

[34] Jerry Alan Fails, Maria Soledad Pera, Oghenemaro Anuyah, Casey Kennington, Katherine Landau Wright, and William Bigirimana. 2019. Query formulation assistance for kids: What is available, when to help & what kids want. In *Proceedings of the 18th ACM international conference on interaction design and children*. 109–120.

[35] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2021. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–49.

[36] Flavio Figueiredo, Felipe Giori, Guilherme Soares, Mariana Arantes, Jussara M Almeida, and Fabricio Benevenuto. 2020. Understanding Targeted Video-Ads in Children's Content. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 151–160.

[37] Michael Flood. 2009. The harms of pornography exposure among children and young people. *Child Abuse Review: Journal of the British Association for the Study and Prevention of Child Abuse and Neglect* 18, 6 (2009), 384–400.

[38] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)* 51, 4 (2018), 1–30.

[39] Stuart J Foster. 1999. The struggle for American identity: Treatment of ethnic groups in United States history textbooks. *History of Education* 28, 3 (1999), 251–278.

[40] Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. 2024. Hate speech detection: A comprehensive review of recent works. *Expert Systems* 41, 8 (2024), e13562.

[41] Fabrizio Germano, Vicenç Gómez, and Francesco Sobbrio. 2026. Ranking for engagement: How social media algorithms fuel misinformation and polarization. *Journal of Public Economics* 255 (2026), 105589.

[42] Jennifer Golbeck. 2025. Recommender System-Induced Eating Disorder Relapse: Harmful Content and the Challenges of Responsible Recommendation. *ACM*

Transactions on Intelligent Systems and Technology 16, 1 (2025), 1–13.
[43] Rachel Griffin. 2025. Governing platforms through corporate risk management: the politics of systemic risk in the Digital Services Act. *European Law Open* 4, 2 (2025), 223–253.
[44] Isuru Gunasekara and Isar Nejadgholi. 2018. A review of standard text classification practices for multi-label toxicity identification of online content. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*. 21–25.
[45] Jacek Gwizdka and Dania Bilal. 2017. Analysis of children's queries and click behavior on ranked results and their thought processes in google search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 377–380.
[46] David L. Hamilton and Tina K. Trolier. 1986. Stereotypes and stereotyping: An overview of the cognitive approach. In *Prejudice, discrimination, and racism*. Academic Press, San Diego, CA, US, 127–163.
[47] David J Hargreaves, Adrian C North, and Mark Tarrant. 2015. How and why do musical preferences change in childhood and adolescence. *The child as musician: A handbook of musical development* (2015), 303–322.
[48] Natali Helberger, Max Van Drunen, Sanne Vrijenhoek, and Judith Möller. 2021. Regulation of news recommenders in the Digital Services Act: Empowering David against the very large online Goliath. *Internet Policy Review* 26, 26 February (2021).
[49] Sierra Holmes. 2014. Mental health matters: Addressing mental illness in young adult fiction. *Language Arts Journal of Michigan* 30, 1 (2014), 14.
[50] Huggingface. 2020. *cardiffnlp/twitter-roberta-base-offensive*. https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive
[51] Huggingface. 2023. *cardiffnlp/twitter-roberta-base-hate-latest*. https://huggingface.co/cardiffnlp/twitter-roberta-base-hate-latest
[52] Huggingface. 2024. *ifmain/ModerationBERT-En-02*. https://huggingface.co/ifmain/ModerationBERT-En-02
[53] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (2020), 1–27.
[54] Prerna Juneja, Md Momen Bhuiyan, and Tanushree Mitra. 2023. Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22.
[55] Prerna Juneja and Tanushree Mitra. 2021. Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–27.
[56] Nityaa Kalra and Savvina Daniil. 2025. Reading Between the Lines: A Study of Thematic Bias in Book Recommender Systems. In *Proceedings of FAccTRec 2025*. Available at: https://doi.org/10.48550/arXiv.2508.15643.
[57] Jessica Kokesh and Miglena Sternadori. 2015. The good, the bad, and the ugly: A qualitative study of how young adult fiction affects identity construction. *Atlantic Journal of Communication* 23, 3 (2015), 139–158.
[58] Natalia Kucirkova. 2019. The learning value of personalization in Children's Reading recommendation systems: what can we learn from constructionism? *International Journal of Mobile and Blended Learning (IJMBL)* 11, 4 (2019), 80–95.
[59] Monica Landoni, Davide Matteri, Emiliana Murgia, Theo Huibers, and Maria Soledad Pera. 2019. Sonny, Cerca! evaluating the impact of using a vocal assistant to search at school. In *International conference of the cross-language evaluation forum for European languages*. Springer, 101–113.
[60] Gerison Lansdown. 2005. *The evolving capacities of the child*. Technical Report. UNICEF.
[61] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.
[62] Sonia Livingstone, Lucyna Kirwil, Cristina Ponte, and Elisabeth Staksrud. 2014. In their own words: What bothers children online? *European Journal of Communication* 29, 3 (2014), 271–288.
[63] Sonia Livingstone and Mariya Stoilova. 2021. *The 4Cs: Classifying Online Risk to Children*. Report. Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI); CO:RE - Children Online: Research and Evidence, Hamburg. doi:10.21241/ssoar.71817
[64] Michele Loi, Matteo Fabbri, and Andrea Ferrario. 2025. Regulating the Undefined: Addressing Systemic Risks in the Digital Services Act (with an Appendix on the AI Act) M. Loi et al. *Philosophy & Technology* 38, 2 (2025), 82.
[65] Gianclaudio Malgieri and Jędrzej Niklas. 2020. Vulnerable data subjects. *Computer Law & Security Review* 37 (2020), 105415.
[66] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2145–2148.
[67] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 15009–15018.
[68] Benjamin Marlin. 2004. *Collaborative filtering: A machine learning perspective*.
[69] Allison Master. 2021. Gender stereotypes influence children's STEM motivation. *Child Development Perspectives* 15, 3 (2021), 203–210.
[70] Aksha M Memon, Shiva G Sharma, Satyajit S Mohite, and Shailesh Jain. 2018. The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. *Indian journal of psychiatry* 60, 4 (2018), 384–392.
[71] Anna-Katharina Meßmer and Martin Degeling. 2023. Auditing Recommender Systems: Putting the DSA into Practice with a Risk-Scenario-Based Approach. https://www.interface-eu.org/publications/auditing-recommender-systems Policy brief.
[72] Lien Michiels, Robin Verachtert, and Bart Goethals. 2022. Recpack: An (other) experimentation toolkit for top-n recommendation using implicit feedback data. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 648–651.
[73] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *Ai & Society* 35, 4 (2020), 957–967.
[74] Ashlee Milton and Maria Soledad Pera. 2020. Evaluating Information Retrieval Systems for Kids. In *Proceedings of the 4th International and Interdisciplinary Perspectives on Children & Recommender and Information Retrieval Systems (KidRec '20), co-located with ACM IDC 2020*. Available at: https://doi.org/10.48550/arXiv.2005.12992.
[75] Camille Mori, Julianna Park, Nicole Racine, Heather Ganshorn, Cailey Hartwick, and Sheri Madigan. 2023. Exposure to sexual content and problematic sexual behaviors in children and adolescents: A systematic review and meta-analysis. *Child abuse & neglect* 143 (2023), 106255.
[76] Carla Moss, Christopher Wibberley, and Gary Witham. 2023. Assessing the impact of Instagram use and deliberate self-harm in adolescents: A scoping review. *International journal of mental health nursing* 32, 1 (2023), 14–29.
[77] Laurens Naudts, Natali Helberger, Michael Veale, and Marijn Sax. 2025. A Right to Constructive Optimization: A Public Interest Approach to Recommender Systems in the Digital Services Act. *Journal of Consumer Policy* (2025), 1–28.
[78] Xia Ning and George Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *2011 IEEE 11th international conference on data mining*. IEEE, 497–506.
[79] Megan Nyhan, Susan Leavy, Pranav Narula, Kevin Doherty, Barry O'Sullivan, Ruihai Dong, Izzy Fox, and Josephine Griffith. 2025. Algorithmic Transparency: The EU Digital Services Act and Young People's Experiences of Online Platforms. In *European Workshop on Algorithmic Fairness*. PMLR, 452–456.
[80] Cecilia Panigutti, Delia Fano Yela, Lorenzo Porcaro, Astrid Bertrand, and Josep Soler Garrido. 2025. How to investigate algorithmic-driven risks in online platforms and search engines? A narrative review through the lens of the EU Digital Services Act. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 828–839.
[81] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 522–533.
[82] Parliament of the United Kingdom. 2023. Online Safety Act 2023: An Act to make provision for and in connection with the regulation by Ofcom of certain internet services; for and in connection with communications offences; and for connected purposes. https://www.legislation.gov.uk/ukpga/2023/50/contents
[83] Royal Pathak, Francesca Spezzano, and Maria Soledad Pera. 2023. Understanding the contribution of recommendation algorithms on misinformation recommendation and misinformation dissemination on social networks. *ACM Transactions on the Web* 17, 4 (2023), 1–26.
[84] María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open* 10, 4 (2020), 2158244020973022.
[85] Maria Soledad Pera and Yiu-Kai Ng. 2017. Using online data sources to make query suggestions for children. In *Web Intelligence*, Vol. 15. SAGE Publications Sage UK: London, England, 303–323.
[86] Anjali Popat and Carolyn Tarrant. 2023. Exploring adolescents' perspectives on social media and mental health and well-being–A qualitative literature review. *Clinical child psychology and psychiatry* 28, 1 (2023), 323–337.
[87] Barbara Rogoff, Audun Dahl, and Maureen Callanan. 2018. The importance of understanding children's lived experience. *Developmental Review* 50 (2018), 5–15.
[88] Almudena Ruiz-Iniesta, Luis Melgar, Alejandro Baldominos, and David Quintana. 2018. Improving children's experience on a mobile EdTech platform through a recommender system. *Mobile Information Systems* 2018, 1 (2018), 1374017.
[89] Noveen Sachdeva, Carole-Jean Wu, and Julian McAuley. 2022. On sampling collaborative filtering datasets. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 842–850.
[90] Alan Said and Alejandro Bellogín. 2018. Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction* 28, 2 (2018), 97–125.
[91] Arianna Sala, Lorenzo Porcaro, and Emilia Gómez. 2024. Social media use and adolescents' mental health and well-being: an umbrella review. *Computers in Human Behavior Reports* 14 (2024), 100404.

[92] Semiu Salawu, Yulan He, and Joanna Lumsden. 2017. Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing* 11, 1 (2017), 3–24.

[93] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[94] Jenna Saul, Rachel F Rodgers, and McKenna Saul. 2022. Adolescent eating disorder risk and the social online world: An update. *Child and Adolescent Psychiatric Clinics* 31, 1 (2022), 167–177.

[95] Markus Schedl and Christine Bauer. 2017. Online Music Listening Culture of Kids and Adolescents: Listening Analysis and Music Recommendation Tailored to the Young. In *Proceedings of the 1st International Workshop on Children and Recommender Systems (KidRec), co-located with ACM RecSys 2017*. Available at: https://doi.org/10.48550/arXiv.1912.11564.

[96] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[97] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*. 1–10.

[98] Anna Schmitz, Michael Mock, Rebekka Görge, Armin B Cremers, and Maximilian Poretschkin. 2025. A global scale comparison of risk aggregation in AI assessment frameworks. *AI and Ethics* 5, 2 (2025), 1407–1432.

[99] Angela Schöpke-Gonzalez, Siqi Wu, Sagar Kumar, and Libby Hemphill. 2025. Using off-the-shelf harmful content detection models: Best practices for model reuse. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–27.

[100] Rosemary Sedgwick, Sophie Epstein, Rina Dutta, and Dennis Ougrin. 2019. Social media, internet use and suicide attempts in adolescents. *Current opinion in psychiatry* 32, 6 (2019), 534–541.

[101] Mohammed Raiz Shaffique and Simone van der Hof. 2025. Behavioural profiling for age assurance: do the ends justify the means? *International Data Privacy Law* (2025), ipaf012.

[102] Jin Shang, Mingxuan Sun, and Kevyn Collins-Thompson. 2018. Demographic inference via knowledge transfer in cross-domain recommender systems. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1218–1223.

[103] Michael Shankleman, Linda Hammond, and Fergal W Jones. 2021. Adolescent social media use and well-being: A systematic review and thematic meta-synthesis. *Adolescent Research Review* 6, 4 (2021), 471–492.

[104] Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I. Nikolenko. 2020. RecVAE: A New Variational Autoencoder for Top-N Recommendations with Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 528–536.

[105] Andrew Shtulman. 2024. Children's susceptibility to online misinformation. *Current opinion in psychology* 55 (2024), 101753.

[106] Amy Singleton, Paul Abeles, and Ian C Smith. 2016. Online social networking and psychological experiences: the perceptions of young people with mental health difficulties. *Computers in Human Behavior* 61 (2016), 394–403.

[107] Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, and János Kertész. 2019. Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model. *PloS one* 14, 3 (2019), e0213246.

[108] Paweł Smaga. 2014. The concept of systemic risk. *Systemic Risk Centre Special Paper* 5 (2014).

[109] David Smahel, Michelle F Wright, and Martina Cernikova. 2015. The impact of digital media on health: children's perspectives. *International journal of public health* 60, 2 (2015), 131–137.

[110] Ivan Srba, Robert Moro, Matus Tomlein, Branislav Pecher, Jakub Simko, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Adrian Gavornik, et al. 2023. Auditing YouTube's recommendation algorithm for misinformation filter bubbles. *ACM Transactions on Recommender Systems* 1, 1 (2023), 1–33.

[111] Mingxuan Sun, Changbin Li, and Hongyuan Zha. 2017. Inferring private demographics of new users in recommender systems. In *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*. 237–244.

[112] Tiffany Ya Tang and Pinata Winoto. 2016. I should not recommend it to you even if you will like it: the ethics of recommender systems. *New Review of Hypermedia and Multimedia* 22, 1-2 (2016), 111–138.

[113] Jessica Taylor. 2014. Romance and the female gaze obscuring gendered violence in the twilight saga. *Feminist Media Studies* 14, 3 (2014), 388–402.

[114] Aditya Tomar, V. Rudra Murthy, and Pushpak Bhattacharyya. 2025. Stereotype Detection as a Catalyst for Enhanced Bias Detection: A Multi-Task Learning Approach. In *Findings of the Association for Computational Linguistics, ACL 2025*. 17304–17317.

[115] Robin Ungruh, Murtadha Al Nahadi, and Maria Soledad Pera. 2025. Mirror, Mirror: Exploring Stereotype Presence Among Top-N Recommendations That May Reach Children. *ACM Transactions on Recommender Systems* (2025).

[116] Robin Ungruh, Alejandro Bellogín, Dominik Kowald, and Maria Soledad Pera. 2025. Impacts of Mainstream-Driven Algorithms on Recommendations for Children Across Domains: A Reproducibility Study. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 783–791.

[117] Robin Ungruh, Alejandro Bellogín, and Maria Soledad Pera. 2025. The Impact of Mainstream-Driven Algorithms on Recommendations For Children. In *European Conference on Information Retrieval*. Springer.

[118] Robin Ungruh and Maria Soledad Pera. 2024. Ah, That's the Great Puzzle: On the Quest of a Holistic Understanding of the Harms of Recommender Systems on Children. In *Proceedings of Designing for Children's Digital Well-being: A Research, Policy and Practice Agenda (DCDW '24), co-located with ACM IDC 2024*. Available at: https://doi.org/10.48550/arXiv.2405.02050.

[119] UNICEF. 1989. Convention on the Rights of the Child. https://www.unicef.org/child-rights-convention/convention-text

[120] UNICEF. 2019. The convention on the rights of the child: The children's version. https://www.unicef.org/child-rights-convention/convention-text-childrens-version#:~:text=A%20child%20is%20any%20person%20under%20the%20age%20of%2018.

[121] UNICEF. 2021. General comment No. 25 (2021) on children's rights in relation to the digital environment. https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation

[122] United States Congress. 2000. Children's Online Privacy Protection Act of 1998 (COPPA). https://www.law.cornell.edu/uscode/text/15/chapter-91

[123] Aleksandra Urman, Ivan Smirnov, and Jana Lasser. 2024. The right to audit and power asymmetries in algorithm auditing. *EPJ Data Science* 13, 1 (2024), 19.

[124] Patti M Valkenburg and Joanne Cantor. 2001. The development of a child into a consumer. *Journal of Applied Developmental Psychology* 22, 1 (2001), 61–72.

[125] Hanne Vandenbroucke, Ulysse Maes, Lien Michiels, and Annelien Smets. 2025. Not One News Recommender To Fit Them All: How Different Recommender Strategies Serve Various User Segments. In *19th ACM Conference on Recommender Systems Proceedings*. 643–648.

[126] Sheila Varadan. 2019. The Principle of Evolving Capacities under the UN Convention on the Rights of the Child. *The International Journal of Children's Rights* 27, 2 (2019), 306–338.

[127] Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one* 15, 12 (2020), e0243300.

[128] Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*. 86–94.

[129] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*. 2605–2610.

[130] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 336–349.

[131] Niobe Way, María G Hernández, Leoandra Onnie Rogers, and Diane L Hughes. 2013. "I'm not going to become no rapper" stereotypes as a context of ethnic and racial identity development. *Journal of Adolescent Research* 28, 4 (2013), 407–430.

[132] Lukas Wegmeth, Tobias Vente, Lennart Purucker, and Joeran Beel. 2023. The Effect of Random Seeds for Data Splitting on Recommendation Accuracy.. In *Perspectives@ RecSys*.

[133] Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, and Nina Taft. 2012. BlurMe: Inferring and obfuscating user gender based on ratings. In *Proceedings of the sixth ACM conference on Recommender systems*. 195–202.

[134] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018: 14th Conference on Natural Language Processing (KONVENS 2018)*.