Evaluating recommender systems with and for children: towards a multi-perspective framework

Emilia Gómez*1*, Vicky Charisi*1* and Stephane Chaudron*2*

1Joint Research Centre, European Commission. Edificio Expo, C. Inca Garcilaso, 3, 41092 Seville, Spain.

2Joint Research Centre, European Commission. Via Enrico Fermi, 2749, 21027 Ispra (VA), Italy.

Abstract

Children are common users of recommender systems (RSs) when watching videos on streaming services, accessing information on the web or playing games, being tablets or phones their favourite devices. Some concerns have been raised by parents and educators on the risks that these systems pose to children and the need to develop products and services that empower children by design and support children's rights. The RSs literature shows that children scenarios are difficult for evaluation, which makes it a clear example of the need to integrate perspectives from multiple stakeholders. Motivated by the need for practical methodologies for children-centric trustworthy artificial intelligence, this paper provides a comprehensive view of the different perspectives involved in the evaluation of RSs for children. We first carry out a literature review, with a focus on the RSs literature, on children-related research, which integrates knowledge from disciplines such as engineering, cognitive science and human-computer interaction. From this review, we identify the main opportunities, challenges and risks related to children-centred RSs and their evaluation. Finally, we propose a multi-perspective framework for the evaluation of RSs for children.

Keywords

recommender systems, information retrieval, children, evaluation, impact assessment, trustworthy artificial intelligence

1. Introduction

A recommender system (RS) is a type of information retrieval (IR) system whose goal is to suggest items from a large collection that meets the preference of a user [1]. RSs are used in a variety of domains, with well-known applications such as video services (e.g. YouTube), product recommenders in online shopping, content recommenders in social media and web content recommenders in a different topic such as restaurants, wines, dating, news, language teachers or financial services. Children are common users of recommender systems. Watching videos is one of the most common digital activities of children reported in the literature [2], where tablets seem to be their favourite devices in studies carried out in Europe and USA [3, 4].

Despite the opportunities for new personalized learning and play experiences that RSs provide to children, parents and educators have raised certain concerns regarding their use in

digital environments by children. One of most promising direction for the mitigation of those risks is the development of services that empower children by design and support children's rights [3, 5]. However, the RSs research literature is limited in child-centric studies compared to adult evaluations so that design decisions on datasets, algorithms and interaction designs are mostly driven by adult needs. Existing literature has shown that children scenarios are difficult for the evaluation of RSs [6, 7] and that they require a multi-stakeholder evaluation as defined in [8].

Motivated by the need for practical methodologies for child-centred artificial intelligence (AI) as defined by UNICEF [9], the goal of this paper is to provide a comprehensive view on the different perspectives involved in the evaluation of RSs for children. We first carry out a literature review on research related to children and RSs, with a comprehensive review of KidRec proceedings (*International and Interdisciplinary Perspectives on Children and Recommender and Information Retrieval Systems* workshop series), key contributions from ACM Recommender Systems Conference - RecSys, and insights from other communities such as cognitive science and human-computer interaction. Then, we identify the main potential opportunities, the emerging risks and the challenges in the evaluation of RSs. As a follow up, we propose a multi-perspective framework for the comprehensive evaluation of RSs with children and for children's well-being.

2. Recommender systems components and evaluation

Recommender systems are implemented through different components which contribute to their outcome and impact, as illustrated in Figure 1. Datasets are crucial component for their development and evaluation, as they set up the application part of the context and scope in terms of information sources used for the recommendation. Machine learning algorithms learn from these data to propose recommendations, which are presented to users by means of a graphical user interface (GUI) and adapted to a particular hardware device, such as computer, mobile phone or tablet. By means of several user-interaction components, the system is able to capture user behaviour with the system, and perform the relevant adaptations to the analyzed data, algorithm and user interface.

State-of-the-art recommendation algorithms are hybrid and combine different approaches such as collaborative filtering techniques (e.g. recommending to a user the items that a similar user liked in the past), content-based algorithms (e.g. recommending to a user similar items to the ones she/he likes), demographic systems (e.g. targeting specific languages or countries) or knowledge-base approaches (e.g. case-based reasoning systems) [1]. As an example, music recommendation systems implement hybrid approaches using collaborative filtering (e.g. play counts, information from peer users), music content description (e.g. features extracted from music audio recordings such as melody, tempo or volume), music context descriptors (e.g. information about the artist or lyrics found on the web), and user properties and behaviour (e.g. demographics, mood). They now integrate state-of-the-art data-driven machine learning techniques [10].

RSs evaluation practices intend to assess the effectiveness of the system and includes the definition of the different aspects of the evaluation:
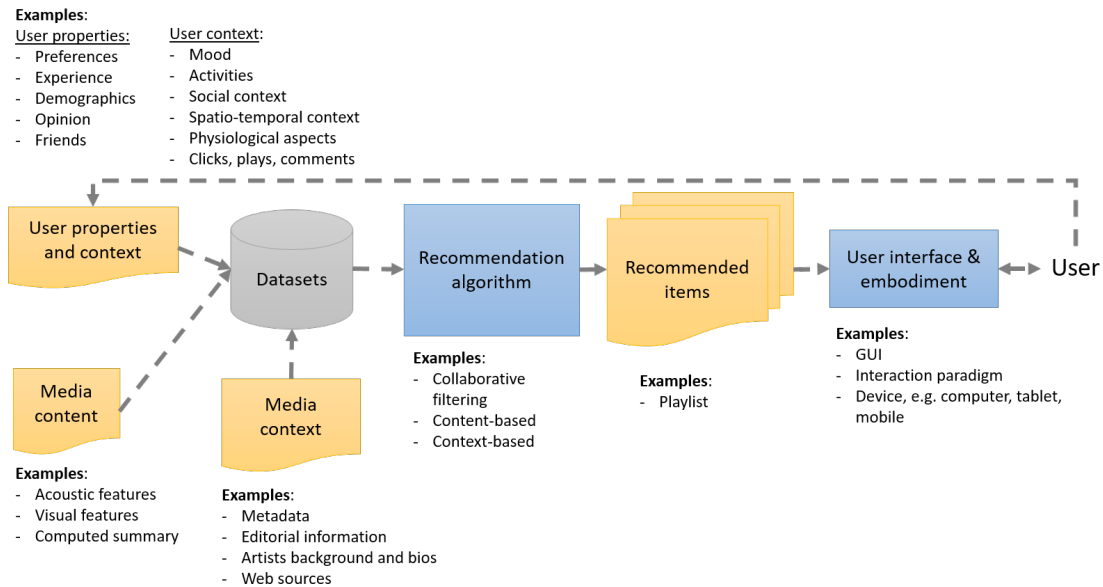
Figure 1: Components of a RS, adapted and complemented from [10]. Yellow blocks refer to system inputs and outputs, grey ones to data and blue ones to data-driven and human-computer interaction approaches. Examples are tailored to audio-visual media recommendation.

- **Content** to be recommended in the evaluation exercise, including the selection or creation of specific content datasets.
- **User population**, defined as the target population for the recommendations in terms of background, experience, age, culture, etc.
- **Methodology** for the evaluation exercise and protocol, e.g. user study, online evaluation (where recommendation results are shown to real users of the system and we observe their behavior to measure their satisfaction with the system, e.g. if they select or play the recommended items), and offline evaluations (where we use existing datasets built from historical data, we then discard some of this data and we try to predict it using the recommendation algorithm) as explained in [11].
- **Criteria or metrics** for system evaluation, with a focus on accuracy, which is usually represented by standard metrics such as mean squared error, root mean squared error, IR metrics such as precision and recall. Other aspects going beyond it include diversity, novelty, coverage, robustness, serendipity, trust, privacy or reproducibility.

3. From general to children-centred recommender systems

General recommender systems, even if not adapted to or designed especially for children, are widely used by children. For instance, according to Statista[1], as of March 2020, a survey on

---

[1]https://www.statista.com/statistics/1150571/share-us-parents-young-child-watch-youtube-videos/

parenting in USA showed that 89% of parents with children aged 5-11 years old and 57% of parents with children aged 0 to 2 years old reported that their children had watched YouTube videos. The usage varies for different countries, media modalities and platforms. For instance, according to the same source, 23% of Brazilian parents with children between 10 and 12 years old stated that their kids used Spotify, and 9% percent of Brazilian parents with children between the ages of 7 and 9, as well as those with toddlers, reported that these children used the digital music, podcast, and video streaming service[2].

This extended usage is confirmed by some research studies. According to Izci et al. [4], *it is recognized that children from all ages use YouTube and researchers found that children as young as 6 months are exposed to videos on the YouTube platform.* Radeski et al. also found YouTube and YouTubeKids to be one of the most commonly used applications in a study with 346 English-speaking parents and guardians of children aged 3 to 5 [2]. Chaudron et al. carried out a cross-national study covering 21 European countries on young children (0 to 8) and digital technologies [3]. Their analysis, grounded on data from 234 family interviews, showed that children usually have their first interaction with digital technologies at a very early age, through their parent's devices, which are not tailored for them in the first place (below 2). In a similar line, parents have identified their own perspectives on the use of RSs by children that relate to the transformation of their own parental role, by the use of online applications with RS as a "digital babysitter" [3, 4].

But as mentioned by Cunningham and Zhang [12], children are not miniature adults, they have different needs, capabilities and expectations of computer products. Some studies have addressed specific children's needs, challenges and risks of this kind of technology [4], and industry has also adapted their products to children (e.g. YouTube Kids[3] or Spotify Kids[4]).

In the ACM Recommender Systems (Recsys) conference[5], the most well-known international forum for RS research, there are only a few (four) papers having the keyword "child" in the title or abstract [13, 14, 15, 16]. Pera and Ng [13] propose and evaluate a book recommender for K-12 users and a readability analysis tool to determine the grade level of books. Milton et al. [15] carry out an empirical study to identify the traits affecting children's preferences in books. They found out preference differences between children from different ages in terms of preferred colours, emotions, length, writing style and topics, and signal the small availability of recorded interaction among young unders and recommender systems. Based on this work the authors present in [16] StoryTime, a web-based book recommender specifically co-designed with children, based on images which elicit their preferences. Fails et al. [14] established in 2017 the *International and Interdisciplinary Perspectives on Children and Recommender and Information Retrieval Systems* workshop series - KidRec[6] - as a research forum on children-specific recommender systems, now in its fifth edition (2017-2021). KidRec proceedings contain 21 research works on different topics related to the design of children-centred RSs, dealing with the specific challenges, applications, evaluation practices and ethical concerns.

---

[2]https://www.statista.com/statistics/1193642/children-using-spotify-brazil/
[3]https://www.youtube.com/kids/
[4]https://www.spotify.com/us/kids/
[5]https://recsys.acm.org/
[6]https://kidrec.github.io/

4. Potential opportunities and applications

The literature identifies several domains where recommender systems can bring value and support children's autonomy in several tasks by facilitating the access to different information sources and modalities. The specific application of RSs for children, as presented in KidRec proceedings, include information search [17], video recommendation [18, 4] (e.g. YouTube or dedicated apps), music recommendation [19], learning [20, 21, 22, 23], second language learning [24, 25], smart toys [26], story and book recommendation [13, 27, 16] and social media (e.g. MessengerKids[7]).

The above-mentioned RS-based platforms for children have the potential, under certain circumstances, to bring unique opportunities for learning, play and entertainment. First, these platforms have the capacity to accommodate and render accessible to children large sets of material that otherwise would not be accessible to them. This has a particular impact in school-based activities, especially for children from less-advantaged socio-economic backgrounds, due to the fact that, otherwise, they would not have access to a teacher to manually curate the information for them. In addition, RSs for game-based learning for children can facilitate self-guided cognitive training, especially when the system has an orientation towards transparency with explainable recommendations [28]. Moreover, these systems can support children's diversification by allowing each individual child to have control in their own learning or play and entertainment trajectories by selecting among a large set of recommended material to be engaged with. In this way, children even from very young age are empowered to develop their agency, especially in online environments, while avoiding information overload [29]. Another particularly beneficial feature of RSs platforms is the possibility for children to send each other messages, thus expanding the database to peer-to-peer recommendations [30]. RSs that are used in educational setting can support cognitive self-regulated learning skills in children, considering individual differences on abilities, preferences, and needs [28]. At the same time, RS-based applications for children are being often developed not only to scaffold the child but to monitor and report the child's progress and predict future performance [31] which might prove beneficial for the teaching process. These features can be used by parents and educators in order to further support child's development and well-being.

The above-mentioned examples of RSs have the potential to benefit children only under certain circumstances. For instance, in the case of a reading recommendation system that collects data on a child's engagement with books, and generates graph data and predictions, it can easily turn into a monitoring and surveillance tool [29] which would probably violate children's rights for privacy. The identification and the mitigation of the potential risks of the use of RSs by children will help us understand the possible necessary future actions for the development of RSs that support children's well-being. In the following sections we elaborate on the relevant literature on the emerging risks and the challenges we face for their evaluation in order to propose certain future directions.

---

[7]https://www.technologyreview.com/2018/02/07/145469/facebooks-app-for-kids-should-freak-parents-out/

5. Emerging risks of recommender systems

While the use of RSs by children brings certain opportunities for children's learning, play and entertainment, recent research literature, policy reports and press articles have identified several risks that children may encounter when using recommender systems:

- **Personal data collection**: data related to the behaviour and interaction of users with RSs is crucial for their development. However, as a vulnerable population it is important to protect children's data and privacy [9] by considering which information is appropriate to gather. For instance, the Children's Online Privacy Protection Act (COPPA) [32] states that people at age 13 can participate in social-media platforms, which need to make sure this age limit is correctly defined and enforced. The General Data Protection Regulation (GDPR) [33] also states that children should merit specific protection with regard to their personal data. As a consequence, there is an additional effort for researchers to establish data ownership, responsibility and protection procedures when they design RSs for children [3].

- **Over-exposure**: several studies mention the risk for children of being exposed to the same or similar media repeatedly, due to the fact that media recommender systems are driven by the notion of similarity [4]. This includes the so-called **information bubbles**, understood as the risk for children to encounter a certain type of content based on its previous choices, reinforcing those and giving less opportunity and room for discovering something different.

- Being exposed to **undesirable content**, given that entertaining videos are not always adequate for children and may for instance contain sexual content, include physical violence or refer to unhealthy food or habits [3, 4].

- Online **advertising** is also considered as a risk, as platforms may treat children as "young consumers", linked to the concept of the "commodification of childhood" [4].

- **Addictions or dependency** on screen has been also identified as a relevant risk[8]. Lukoff et al. [34] carried out a survey with 120 YouTube adult users and a set of co-design experiences to analyze how some internal mechanisms implemented in the app can support user agency, as low sense of agency can relate to negative life impacts such as loss of social opportunities, sleep or productivity. The authors found out that, on the one hand, some mechanisms such as autoplay or automatic recommendations, decrease the user sense of agency. On the other hand, some other functionalities such as search, or playlist creation can support it. Research studies addressing children provide conclusions in a similar line. Hiniker et al. [35] carried out a behavioural study with 24 3-5 y.o. children, and they found that some design features can support children's autonomy and self-regulation, such as those providing opportunity for planning and making choices, the ones reminding children of their intentions and those asking questions to the child. However, others such as post-play, can undermine it.

- Social media platforms or apps such as Messenger Kids allow children to post and message friends through a federation mechanism monitored by parents. Some voices have signaled

---

[8]What Screen Addictions and Drug Addictions Have in Common https://www.pbs.org/wgbh/nova/article/screen-time-addiction/

the risk of these applications to be used to **familiarize children with commercial products** that will be used when they become teenagers[9].

- **Difficulty for parents to monitor children's behaviour**, as recommender systems are consumed by children mostly on personal devices such as tablets or phones [3].
- Propagation of existing **gender stereotypes** present in search and recommendation systems [36].

Although some of these risks also appear in adult population, children need special protection, given their vulnerability and potential impact in their cognitve and socio-emotional development [9]. In addition, the particular tendency of children to use trial-and-error methods to learn how to use a tool increases several risks such as the deviation from non-suitable content, the accidental disclosure of personal information, and the unintended contact with people [3].

6. Challenges of RSs evaluation with children

In the previous sections we elaborated on the potential advantages and the emerging risks of the use of RSs by children. For the design and development of RSs that promote children's rights and benefit their well-being while taking advantage of the unique opportunities of the use of those systems, we need to development scientifically rigorous and responsible techniques for their evaluation which, as we will discuss, is still a challenging endeavor.

Some studies have identified and analysed children behaviour in adult-centred platforms. One example in the music domain is the work by Shedl and Bauer [19] in the Last-FM platform. Among all users, the authors found a small presence of children 6-17 vs adults, and a small but significant presence of young children (e.g. 6-10), including 5,953 users (12.9% of users in the platform). For those children using Last-FM , Schedl and Bauer found that recommendation algorithms based on collaborative filtering seem to work better for children than for adults. The authors also found significant differences in the musical genres preference between young and adult listeners. For instance, young listeners were found to have a high preference for rock music and low preference for blues. In addition, the youngest age group (6-12) was found to appreciate electronic music the most in comparison to the other age groups, and rock, folk, punk, alternative, and metal were the least liked genres by this youngest group, compared to the older groups [19]. The need to define children-specific musical genres is also visible in some commercial products, e.g. Spotify Kids, with genres such as movies music, bedtimes tunes, party jams and stories.

We also find studies specifically focused on children, such as the work by Cunningham and Zhang [12], who propose a participatory design activity for children with mmusic recommender systems, and is the only paper of ISMIR (*International Society for Music Information Retrieval Conference*[10]) with the "Child" keyword in the title. The authors develop *Kids Music Box*, a music recommendation system created with 6-10 y.o. children in mind. In this work, the authors organize the different challenges for children to use music recommendation platforms in terms of their cognitive and physical development and their preferred functionalities. In terms of

---

[9]Child health advocates call for Facebook to shutter Messenger Kids app http://social.techcrunch.com/2018/01/30/child-health-advocates-call-for-facebook-to-shutter-messenger-kids-app/

[10]http://www.ismir.net

cognitive development, the authors mention that children should not be forced to use software designed with complex interaction and interfaces, requiring good spelling, and reading skills beyond their current abilities. In addition, they mention the need for children to get constant visual or acoustic feedback, which is not always provided by textual interfaces. Children may have difficulty with abstract concepts, so the selected icons should represent familiar, real-world objects. Finally, they signal the fact that children use trial-and-error methods to learn how to use a tool, which is not always the case for adults. In terms of physical development, the authors suggest that children may have difficulty controlling the mouse, targeting small areas on the screen or typing on the keyboard. We then need to design simple physical interactions for this user population. Finally, as regards the specific needs or preferred functionalities for music RSs, they mention the rating of songs, the synchronization of visuals with the music, the incorporation of games while listening to music and the option to have parental setting or control. These findings are inline with the need to integrate children-specific design recommendations, which are adapted to their cognitive and physical needs and abilities.

In fact, Human-Computer Interaction research has widely addressed the evaluation of interfaces with children. Soni et al. review existing design recommendations for children's touchscreen interfaces based on cognitive, physical and socio-emotional developmental appropriateness [37]. In their work, the authors define, from a review of the state of the art, the *Touchscreen Interaction Design Recommendation for Children* - TDRC framework, incorporating 57 different design recommendations found in the literature, organized by interface dimensions. This framework was used to empirically analyze how these recommendations were considered in 50 popular apps, finding out that only 63% of those apps followed design recommendations to fulfill children's cognitive (51%), physical (67%) and socio-emotional (72%) needs. This study illustrates the existing divergence between research findings and practical children-centred touchscreens applications.

In the recommender systems literature, Ekstrand [6] summarizes the challenges of evaluating RSs with children, confirmed by other authors, and including the following issues:

- **Data availability**: the lack of data (i.e. the so called "cold-start problem") is one of the limitations of children-centric studies, emphasized with the above mentioned rights of data protection, also signaled in [7]; All these aspects limit the availability of benchmarking datasets including children users, which are crucial for algorithm evaluation and development and to ensure the reproducibility of studies [6].
- **Limited survey abilities** when dealing with children. Surveys provide a common strategy and practical way for large-scale evaluation of RSs. However, some studies have signaled the limitations of this methodology for children [38, 3, 6]. For instance, click logs from children interactions with a system are likely to be noisier than those from general users, and children are unlikely to be able to provide robust ratings particularly when attempting to accommodate different factors such as educational value or information accuracy. Other methodologies such as user studies, usability exercises and participatory design processed are then required for children, which are costly to be carried out on a large scale.
- **Multi-stakeholder evaluation**: as mentioned in [6], RS evaluation has been traditionally centred on metrics and protocols that measure how the different system components

impact "user" behaviour (e.g. accuracy, satisfaction, play counts) or platform/business outcomes (e.g. sales, user retention). However, in child-centred recommendation, we need to consider different stakeholders as related to the target "user" or "consumer" as indicated by Bauer and Jannach [8]: the child has particular interests and information needs; the caretaker might decide on the information and content that are suitable for the child; in educational scenarios, the teacher uses a RS to support certain learning outcomes; Other stakeholders mentioned in [8] include the RS provider (e.g. platform), supplier (e.g. product manufacturer) and society in general. For instance, the RS provider want to ease the discovery of specific content according to their business model. These different views need to be formulated and integrated into the design of the evaluation protocol. As mentioned by Bauer and Jannach in [8], multi-stakeholder evaluation implies the optimization of multiple objectives in parallel, and needs to be considered from the dataset and algorithm itself to the evaluation methodology. The authors also reflect on the concept of fairness and other ethical questions arising from the consideration of different stakeholders, e.g. provider vs consumer, which is another research gap [8].

7. Towards a multi-perspective evaluation framework

After analyzing existing literature, we observe that the evaluation of recommender systems for children is very challenging, and different studies have approached varied evaluation aspects in specific contexts. Landoni et al. [39] proposed an evaluation framework allowing the comparative analysis of diverse IR *strategies by a given *user group, task* and *context*. Building upon this work and the previous review, we propose a multi-perspective framework covering the four dimensions represented in Figure 2: component, stakeholder, methodology and temporal scale, as illustrated in Figure 2 and further detailed in the following subsections. These dimensions should be driven from the intended **context, purpose and expected value** of the RS [8].
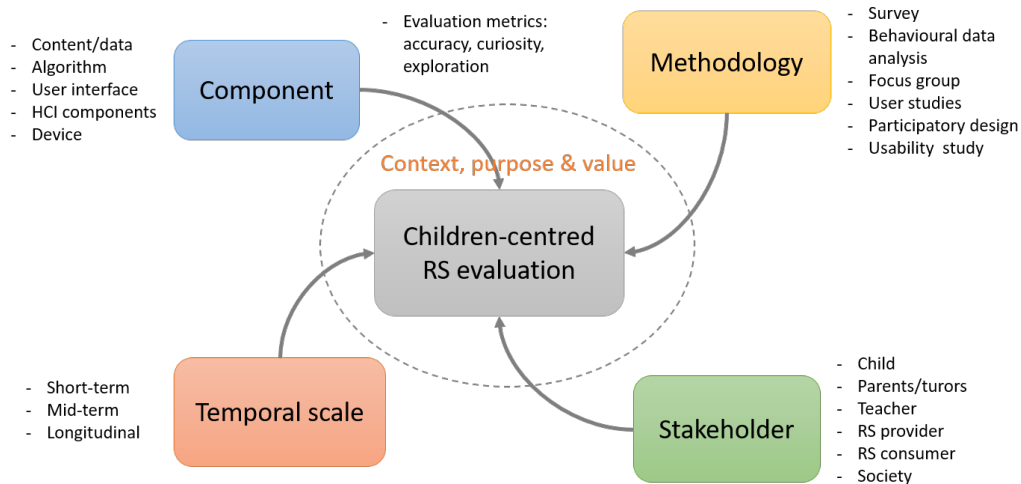
Figure 2: Perspectives to be considered in the evaluation of RS for children.

As a complement to these four perspectives, we consider the aspects of **reproducibility and transparency** as key requirements for meaningful evaluations, as open protocols and community-built toolkits are the only way towards incremental, comparative and comprehensive evaluations of RSs.

7.1. Component

We have seen that RSs are complex systems as represented in Figure 1. Different components contribute to the system output, such as the device (e.g. computer, tablet, phone), user interface, interaction mechanisms, functionalities, recommendation algorithm, data collected from children or content information used for training. These components may also need varied evaluation strategies, and there is a need to understand the impact of each of the components into the final evaluation outcomes. Most evaluation approaches reviewed here focus on full system evaluation or stay in the particular device, set of functionalities, the user interface or the interaction paradigm. Up to our knowledge there are no comprehensive evaluations on how specific steps of the recommendation process affect and should target children, e.g. content description methods, item similarity metrics, emotion recognition models, or collaborative filtering strategies. This indicates a clear risk of bias and malfunction of state of the art RS for children. Full-system evaluation, combined with the understanding of the role of the different components, should be the target goal.

7.2. Stakeholder

We see the need to consider the evaluation exercise from the perspective of the different stakeholders involved, which might have different needs and expectations from the RS. In addition, we need to adapt the evaluation methodology to the particular user (e.g. as mentioned before, surveys might be more adequate for adults). Moreover, the interaction between those stakeholders and the processes that develop among them should also be understood. We reflect now on the main stakeholders involved in the design of child-specific RSs:

- <u>Children</u> have specific information needs and preferences, empowering their participation in the design process. From this perspective, evaluation practices should reflect on the child's individual history and current behaviour, the current context and culture. Importantly, children are not an homogeneous group: age, gender and family and social background affect their choices and preferences. From this perspective, evaluation practices should reflect on the child's individual history and current behaviour, the current context and culture.
- <u>Parents</u> or guardians should be able to incorporate their preferences in terms of protection and values to be transmitted by RSs. We should note that, although some studies such as [3, 2] are based on interviews with parents/guardians, according to Radesky et al., *parent-reported duration of mobile device use in young children has low accuracy, and the use of objective measures is needed in future research.* This reveals the need to contrast the evaluation results obtained with different methodologies and stakeholders.
- <u>Educators</u>: educational goals and expected learning outcomes are of particular relevance when using RSs in educational contexts. Often, parents are involved in educational

activities with their children, especially in informal settings, and educators have a great role in the protection of children as well as in the creation of opportunities of children's participation. Notably, tensions can appear between decisions supporting children's online protection and participation.

- Companies need to consider the effect of the RS on their business and business model. Being the main developers and integrator of the RS, it is important to understand their needs and limitations as related to system evaluation.
- Policy makers: evaluation practices and results can provide the needed scientific evidence to design policies that can minimize risks, ensure children protection, empower their participation, and support shaping the current educational systems to prepare children in the best possible way.

7.3. Temporal scale

A third important dimension is the temporal scope of the evaluation exercise, that should also fit its purpose:

- Short-term: if we design a one-shot co-design exercise, user/usability study or survey, we will research on the immediate effect of recommendations.
- Mid-term: in this case, we would follow children in their interaction with a RS to study the potential impact after several sessions or exercises.
- Long-term: longitudinal studies are also needed, with the goal of understanding in the impact RSs may have on children in the long-term, e.g. for their future development as teenagers or adults.

7.4. Methodology

Finally, we have mentioned the need to combine different methodologies for a comprehensive RSs evaluation:.

- Criteria and metrics: evaluation goals and criteria are linked to the selected metrics, either algorithm-centred accuracy metrics or application-specific holistic ones [40]. Metrics are also linked to the needs of different stakeholders and the target component and temporal scope. We consider, for instance, that the overall challenge of designing a child-specific RS is to understand how the RS tackle child's cognitive models and developmental aspects. This includes the consideration of different aspects such as: (1) How the RS supports the child's need for agency acquisition; (2) How to implement design decisions that better fit children's attention span; (3) Which is the role of interactivity in children's connections between online and offline scenarios; and (4) Which is the correct balanced regarding children's rapid development and their predisposition for repetition (especially in early childhood). Although traditionally the goal of the evaluation of a RS is linked to its accuracy, we also need to evaluate whether the tool scaffolds child's well-being and development by prioritizing children's innate characteristics such as curiosity and exploration.

- Set up and protocol: here, we would need to detail and select relevant evaluation methods including quantitative (survey, behavioural data analysis) and qualitative (participatory design exercise, usability study, interviews, focus group, ethnographic studies) approaches.

8. Conclusions

In this paper we have summarized existing literature on the evaluation, opportunities, risks and challenges of children using recommender systems. An analysis of the literature on children-specific RSs has revealed the main challenges researchers address to evaluate recommender systems with children's audiences, and the importance of children-centred design to minimize the risks that recommender systems pose, without sacrificing the opportunities such systems can bring to children. Our review shows that evaluation practices typically focus on RS accuracy; however we need to include other points for evaluation such as whether the tool scaffolds the child's well-being and development by prioritizing children's innate characteristics such as curiosity, exploration and creativity.

In addition, while most research focus on partial aspects of the evaluation such as the effect of design decisions and interfaces in particular contexts, we propose a comprehensive multi-perspective framework to develop reproducible and incremental evaluation practices allowing the scientific understanding on the impact, potential bias and needed adaptations of RSs for children, to make sure these systems support their current and future welfare.

As mentioned in [8], we think that only by evaluating RSs from these different perspectives the research community will understand the effect that their designs may have on individual stakeholders (e.g. children, parents, business) and the wider society.

References

[1] F. Ricci, L. Rokach, B. Shapira, P. Kantor, Recommender Systems Handbook, Springer, 2011. doi:10.1007/978-0-387-85820-3.

[2] J. S. Radesky, H. M. Weeks, R. Ball, A. Schaller, S. Yeo, J. Durnez, M. Tamayo-Rios, M. Epstein, H. Kirkorian, S. Coyne, R. Barr, Young children's use of smartphones and tablets, Pediatrics 146 (2020). URL: https://pediatrics.aappublications.org/content/146/1/e20193518. doi:10.1542/peds.2019-3518.

[3] S. Chaudron, R. D. Gioia, M. Gemo, Young Children (0-8) and Digital Technology - A qualitative study across Europe, Publication Office of the European Union, 2018. doi:10.2760/294383.

[4] B. Izci, I. Jones, T. Ozdemir, L. Alktebi, E. Bakır, Youtube and young children: Research, concerns, and new directions., Lisbon School of Education, 2019, pp. 81–92.

[5] EU strategy on the rights of the child COM/2021/142 final, Technical Report, European Commission, 2021. URL: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0142.

[6] M. Ekstrand, Challenges in evaluating recommendations for children, in: KidRec 2017, 2017.

[7] A. Milton, #thehorror: Evaluating information retrieval systems for kid, in: KidRec 2017, 2020.

[8] D. Jannach, C. Bauer, Escaping the mcnamara fallacy: Towards more impactful recommender systems research, AI Magazine 41 (2020) 79–95. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/5312. doi:10.1609/aimag.v41i4.5312.

[9] V. Dignum, M. Penagos, K. Pigmans, S. Vosloo, Policy Guidance on AI for Children, Technical Report, UNICEF, 2020. URL: https://www.unicef.org/globalinsight/reports/policy-guidance-ai-children.

[10] M. Schedl, E. Gómez, J. Urbano, Music Information Retrieval: Recent Developments and Applications, Now Foundations and Trends, 2014. doi:10.1561/9781601988072.

[11] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, B. Gipp, A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation, in: Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, RecSys '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 7–14. URL: https://doi.org/10.1145/2532508.2532511. doi:10.1145/2532508.2532511.

[12] S. J. Cunningham, E. Zhang, Development of a music organizer for children, in: J. P. Bello, E. Chew, D. Turnbull (Eds.), ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008, 2008, pp. 185–190. URL: http://ismir2008.ismir.net/papers/ISMIR2008_123.pdf.

[13] M. S. Pera, Y.-K. Ng, What to read next? making personalized book recommendations for k-12 users, in: Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 113–120. URL: https://doi.org/10.1145/2507157.2507181. doi:10.1145/2507157.2507181.

[14] J. A. Fails, M. S. Pera, F. Garzotto, M. Gelsomini, Kidrec: Children & recommender systems: Workshop co-located with acm conference on recommender systems (recsys 2017), in: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 376–377. URL: https://doi.org/10.1145/3109859.3109956. doi:10.1145/3109859.3109956.

[15] A. Milton, M. Green, A. Keener, J. Ames, M. D. Ekstrand, M. S. Pera, Storytime: Eliciting preferences from children for book recommendations, in: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 544–545. URL: https://doi.org/10.1145/3298689.3347048. doi:10.1145/3298689.3347048.

[16] A. Milton, L. Batista, G. Allen, S. Gao, Y.-K. D. Ng, M. S. Pera, "don't judge a book by its cover": Exploring book traits children favor, in: Fourteenth ACM Conference on Recommender Systems, RecSys '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 669–674. URL: https://doi.org/10.1145/3383313.3418490. doi:10.1145/3383313.3418490.

[17] M. Landoni, E. Murgia, T. Huibers, M. Pera, My name is sonny, how may i help you searching for information?, in: IDC '19, Association for Computing Machinery (ACM), United States, 2019. 18th ACM International Conference on Interaction Design and Children, IDC 2019, IDC 2019 ; Conference date: 12-06-2019 Through 15-06-2019.

[18] Y. Deldjoo, C. Frà, M. Valla, M. A. Tuncel, F. Garzotto, P. Cremonesi, A. Paladini, D. Anghi-

leri, Enhancing children's experience with recommendation systems, in: KidRec 2017, 2017.

[19] M. Schedl, C. Bauer, Online music listening culture of kids and adolescents: Listening analysis and music recommendation tailored to the young, in: KidRec 2017, 2017.

[20] M. S. Pera, Y.-K. Ng, With a little help from my friends: Generating personalized book recommendations using data extracted from a social website, in: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, volume 1, 2011, pp. 96–99. doi:`10.1109/WI-IAT.2011.9`.

[21] M. Landoni, E. Murgia, F. Gramuglio, G. Manfredi, Teaching an alien: Children recommending what and how to learn, in: KidRec 2018, 2018.

[22] T. Horiuchi, M. Rothschild, R. Barrera, S. Gururajan, Designing a personally meaningful abcmouse.com: Challenges and questions in an edtech recommendation system, in: KidRec 2018, 2017.

[23] A. Milton, E. Murgia, M. Landoni, T. Huibers, M. Pera, Here, there, and everywhere: Building a scaffolding for children's learning through recommendations, in: O. Shalom, D. Jannach, I. Guy (Eds.), ImpactRS 2019: Impact of Recommender Systems 2019, CEUR workshop proceedings, CEUR, 2019. 1st Workshop on the Impact of Recommender Systems, ImpactRS 2019, ImpactRS ; Conference date: 19-09-2019 Through 19-09-2019.

[24] W. Ma, M. Zhang, C. Zhang, Y. Chen, Q. Xie, W. Sun, Y. Liu, S. Ma, A game-based data collecting framework for the recommendation of kids' second language learning, in: KidRec 2017, 2017.

[25] H. Xie, M. Wang, D. Zou, F. L. Wang, A personalized task recommendation system for vocabulary learning based on readability and diversity, in: International conference on blended learning, Springer, 2019, pp. 82–92.

[26] F. Delprino, O. F. Bravo, M. Mariani, C. Piva, N. Izzo, M. Matera, R. Tassi, Playing outdoor, recommending new content: Stimulating kids' learning through the abbot smart object, in: KidRec 2017, 2017.

[27] M. S. Pera, K. Wright, M. Ekstrand, Recommending texts to children with an expert in the loop, in: KidRec 2018, 2018.

[28] K. Tsiakas, E. Barakova, J. V. Khan, P. Markopoulos, Brainhood: towards an explainable recommendation system for self-regulated cognitive training in children, in: Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2020, pp. 1–6.

[29] N. Kucirkova, 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 (2019) 80–95.

[30] I. Picton, The impact of ebooks on the reading motivation and reading skills of children and young people: A rapid literature review., National Literacy Trust (2014).

[31] M. Ueno, Y. Miyazawa, Irt-based adaptive hints to scaffold learning in programming, IEEE Transactions on Learning Technologies 11 (2017) 415–428.

[32] Children's Online Privacy Protection Rule ("COPPA"), Technical Report, Federal Trade Commission, United States, 2021. URL: https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.

[33] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing

of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Technical Report, European Parliament and Council, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.

[34] K. Lukoff, U. Lyngs, H. Zade, J. V. Liao, J. Choi, K. Fan, S. A. Munson, A. Hiniker, How the Design of YouTube Influences User Sense of Agency, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445467.

[35] A. Hiniker, S. S. Heung, S. R. Hong, J. A. Kientz, Coco's Videos: An Empirical Investigation of Video-Player Design Features and Children's Media Use, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–13. URL: https://doi.org/10.1145/3173574.3173828.

[36] A. Raj, A. Milton, M. D. Ekstrand, Pink for princesses, blue for superheroes: The need to examine gender stereotypes in kid's products in search and recommendations, CoRR abs/2105.09296 (2021). URL: https://arxiv.org/abs/2105.09296. arXiv:2105.09296.

[37] N. Soni, A. Aloba, K. S. Morga, P. J. Wisniewski, L. Anthony, A framework of touchscreen interaction design recommendations for children (tidrc): Characterizing the gap between research evidence and design practice, in: Proceedings of the 18th ACM International Conference on Interaction Design and Children, IDC '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 419–431. URL: https://doi.org/10.1145/3311927.3323149. doi:10.1145/3311927.3323149.

[38] N. Borgers, E. de Leeuw, J. Hox, Children as respondents in survey research: Cognitive development and response quality 1, Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 66 (2000) 60–75. URL: https://doi.org/10.1177/075910630006600106. doi:10.1177/075910630006600106. arXiv:https://doi.org/10.1177/075910630006600106.

[39] M. Landoni, D. Matteri, E. Murgia, T. Huibers, M. Pera, Sonny, cerca! evaluating the impact of using a vocal assistant to search at school, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, Springer, Netherlands, 2019, pp. 101–113. doi:10.1007/978-3-030-28577-7_6, 10th International Conference of the CLEF Association, CLEF 2019.

[40] O. Anuyah, M. Green, A. Milton, S. Pera, The need for a comprehensive strategy to evaluate search engine performance in the classroom, in: KidRec 2019, 2019.