Report on NORMalize: The First Workshop on the Normative Design and Evaluation of Recommender Systems

Sanne Vrijenhoek[1], Lien Michiels[2], Johannes Kruse[3], Alain Starke[1,4], Jordi Viader Guerrero[5] and Nava Tintarev[6]

[1]*University of Amsterdam, Amsterdam, the Netherlands*
[2]*University of Antwerp, Antwerp, Belgium*
[3]*Ekstra Bladet, Copenhagen, Denmark*
[4]*University of Bergen, Bergen, Norway*
[5]*TU Delft, Delft, the Netherlands*
[6]*Maastricht University, Maastricht, the Netherlands*

Abstract

Recommender systems are among the most widely used applications of artificial intelligence. Because of their widespread use, it is important that practitioners and researchers think about the impact they may have on users, society, and other stakeholders. To that effect, the NORMalize workshop seeks to introduce *normative thinking*, to consider the norms and values that underpin recommender systems in the recommender systems community. The objective of NORMalize is to bring together a growing community of researchers and practitioners across disciplines who want to think about the norms and values that should be considered in the design and evaluation of recommender systems, and further educate them on how to reflect on, prioritise, and operationalise such norms and values. This document is a report on the first workshop, co-located with ACM RecSys '23 in Singapore.

Keywords

normative thinking, normative design, recommender systems, norms, values, value-sensitive design

1. Introduction

Users and developers of recommender systems are becoming increasingly aware of the possible societal impact of their systems [1]. As 'beyond-accuracy' metrics are becoming more common in recommender research, much attention has been given to methods related to notions of fairness, such as statistical parity or equality of opportunity in the design or evaluation of recommender systems [2, 3]. However, many values could be considered in the development

NORMalize 2023: The First Workshop on the Normative Design and Evaluation of Recommender Systems, September 19, 2023, co-located with the ACM Conference on Recommender Systems 2023 (RecSys 2023), Singapore

s.vrijenhoek@uva.nl (S. Vrijenhoek); lien.michiels@uantwerpen.be (L. Michiels); johannes.kruse@eb.dk (J. Kruse); a.d.starke@uva.nl (A. Starke); j.viaderguerrero@tudelft.nl (J. V. Guerrero); n.tintarev@maastrichtuniversity.nl (N. Tintarev)
0000-0002-1031-4746 (S. Vrijenhoek); 0000-0003-0152-2460 (L. Michiels); 0009-0007-5830-0611 (J. Kruse); 0000-0002-9873-8016 (A. Starke); 0009-0003-8556-7670 (J. V. Guerrero); 0000-0003-1663-1627 (N. Tintarev)

CEUR Workshop Proceedings (CEUR-WS.org)

and goal of a recommender systems, of which fairness towards the end-users of the system is but one example [4].

Identifying and balancing these values requires so-called *normative thinking* and decision-making [5, 6, 7]. Normative thinking requires us to reflect on how or what the system *should be*, rather than focusing on what the current state of the system (output) *is*. Besides identifying relevant values, this includes determining how these values would be expressed in what is recommended by a system, how different values may be conflicting, and justifying how certain values in such cases should be prioritised over others [8].

2. Interactive Session

In the on-site morning session, participants were first introduced to the principles and practices of normative thinking. After this short lecture, participants were split into breakout groups. In these groups, they discussed a specific use case of a recommender system. There were three groups: X (formerly known as Twitter), BBC News, and Spotify. First, participants were asked to identify when, where, and how the system would used and what it recommended. For example, the Spotify recommender system(s) need to recommend songs, albums and artists, but also podcasts and playlists, and so on. Then, they identified relevant stakeholders and the norms and values that mattered to them. Again using the Spotify example, the participants identified many different stakeholders: from advertisers, to creators, end-users to investors and more. They then catalogued these stakeholders' values, for example, discoverability matters greatly to creators and indie labels, whereas profit matters most to investors. Next, they considered the relationships between values and their possible (negative) consequences. Using the example of X, we might ask *if we value freedom of speech, could that lead to hate speech and misinformation?* Subsequently, each group was allocated a total of one hundred points to be divided among various values. Each group member was given the responsibility to represent one or more stakeholders of the recommender system and to champion their respective values.

Each group was given a starting kit to work with: Marked envelopes with instructions for each step of the process, sticky notes, sharpies, and pens. Groups were free to come up with their own creative process. Each group approached the task differently: Whereas the Spotify group immediately created a mind map on the wall, the Twitter group only made notes on paper. The group work concluded with a discussion of what a recommender system that prioritizes values and stakeholders in such a way would look like. Finally, each group presented the outcomes of their discussion to all workshop participants and organizers. Interestingly, we found that outcomes differed greatly as well: Every use case had different amounts of stakeholders and values, and as a result, took a different amount of time to complete every step.

After the session concluded, participants were asked to complete a short survey to gauge their satisfaction with the process, as well as provide oral feedback. Generally speaking, the interactive session was well received by the participants. Those who completed the survey unanimously found that instructions were clear, and indicated that they had learning something during the session. The most well-liked parts of the interactive session were those that required discussion: Assigning values to stakeholders, discussing consequences and prioritizing values as a group. We also asked participants how we could further improve the interactive session.

One participant mentioned that by using real companies as use cases, people were forced to reason as if they were a part of these companies, which limited their creativity somewhat. They suggested to use fictional company descriptions in the future instead.

3. Keynote

The subsequent keynote addressed the norms and governance of recommender systems on digital platforms like YouTube and TikTok, especially in relation to user-generated content. It addressed concerns about the platforms' algorithmic systems contributing to user harm and challenged the notion of platforms as mere content conduits. There were three main points: First, the need to question the established definitions of recommendation-related harms and to encourage diverse frameworks for evaluating these systems. Second, the importance of considering the long-term effects of information landscape commercialization and the potential of algorithmic recommendation for elevating historically excluded voices. Lastly, the keynote called for greater appreciation of the nature of the 'items' being recommended, which opens up possibilities for more sophisticated discussions on normative frameworks for curation.

4. Submitted Work

The accepted work (13 registered abstracts, 9 accepted) can be thematically clustered into papers dealing with "Power Structures", "News Recommendation" and "Practical Applications". Each paper received three reviews by members of the program committee, at least one of which was from a technical- and one from a social science/humanities background.

4.1. Power Structures

Recommender systems often exist in a complex environment, with multiple stakeholders competing for optimization of their own objectives. In "Towards a Pragmatic Approach for studying Normative Recommender Systems: exploring Power Dynamics in Digital Platform Markets", Binst et al. argue that decision power exists primarily on the side of the system providers. They illustrate this with the key bottlenecks "lock-in and monopolization" and "engagement-centric logic", and make concrete suggestions for regulatory principles that may alleviate them.

"Designing and Implementing Socially Beneficial Recommender Systems: An Interdisciplinary Approach" by Mallia presents a theoretical argument on how we can move from engagement-centric recommender systems to recommender systems that have a positive impact on society. Central to this discussion is the definition of what a 'positive social outcome' actually is, as this is dependent on a multitude of socio-cultural factors. Furthermore, their actual implementation requires interdisciplinary methodologies and collaboration.

The design of recommender systems is often rooted in a utilitarian or consequentialist world view. "Digital Humanism and Norms in Recommender Systems" by Prem et al. details how Digital Humanism can serve as a useful lens to approach complex issues surrounding the design of recommender systems, and as such promote values such as human rights, democracy, inclusion and diversity. For example, from the Digital Humanism perspective users should be empowered to understand the system, and as a consequence will make better choices regarding their interactions with it.

4.2. News Recommendation

Publicly available datasets are crucial for tackling challenges faced by news recommender systems, especially in terms of news diversity. Fortunately, Lucas et al. introduced the News Portal Recommendations (NPR) dataset in their work, ”NPR: A News Portal Recommendations Dataset”. Distinct from the Microsoft News Dataset (MIND) [9], the NPR dataset focuses on frequent user interactions with hard news. Furthermore, to assess diversity metrics, Lucas et al. enriched the dataset with the metadata needed to employ the RADio framework [4] on the NPR dataset.

Building on the theme of enhancing news recommendations, another noteworthy study delves deeper into a specific challenge. In ‘Improving and Evaluating the Detection of Fragmentation in News Recommendations with the Clustering of News Story Chains’, Polimeno et al. focus on quantifying fragmentation in news recommendations. Specifically, they examine how to accurately measure the fragmentation of information streams in news recommendations. To do this, they employ Natural Language Processing (NLP) to identify distinct news events, stories, or timelines. Their work features a thorough investigation of different approaches, such as hierarchical clustering coupled with SentenceBERT text representation, along with the analysis of simulated scenarios. These results could provide valuable insights for stakeholders concerning the measurement and interpretation of fragmentation.

Going beyond data and fragmentation, there is also an emerging emphasis on enhancing user experience in news recommendations. Kiddle et al. formulate a novel user-centric approach for promoting serendipity in news recommender systems. This approach leverages user familiarity with the algorithmic language of recent social media, particularly TikTok, to nurture news discovery. They introduce the concept of ‘navigable surprise’, which they define as the experience of encountering novel, diverse, relevant, and unexpected information under conditions of immediate (i.e., real-time) and bounded (i.e., item-oriented) agency. To realize ‘navigable surprise’, they propose a combination of short-term interest modeling with consumption-based (implicit) user signaling. As such, they highlight the centrality of short-term interest modeling to serendipity in recommender design.

4.3. Practical Applications

In “Value-Based Nudging in News Recommender Systems – Results From an Experimental User Study”, Modre et al. explore the potential of nudges for changing people’s news reading behavior. They evaluate two types of nudges: feedback-based and social norms-based, both grounded in theory from psychology and related social sciences, and find social norms-based nudges achieve the best results. Their study provides a great example of how interdisciplinary work that bridges the social and computer sciences can help develop more effective, socially responsible recommender systems.

“Refining Deliberative Standards for Online Political Communication: Introducing a Summative Approach to Designing Deliberative Recommender Systems” by Stolwijk et al. formulate design guidelines, rooted in political theory, for recommender systems that wish to foster deliberative democracy. By proposing a set of concrete metrics and objectives that can be used to design and evaluate deliberative recommender systems, they contribute to the operationalization of normative goals that were previously overlooked.

In "Classification of Normative Recommender Systems", Heitz proposes a classification of recommender systems into four types related to how and when normative goals are introduced to the recommender system; for example, in the preprocessing stage or as a postprocessing step. He argues that different types are not directly comparable and will lead to different results. As such, his classification contributes to a more 'mature' debate on normative goals in recommender systems.

Acknowledgments

We would like to thank the participants and authors of accepted contributions for their valuable inputs to the workshop, our program committee for their thoughtful reviews, as well as the RecSys'23 organisers for their support in the organisation of NORMalize. Finally, we would like to thank our employers and funding bodies. Sanne Vrijenhoek's contribution to this research is supported by the AI, Media and Democracy Lab. Lien Michiels' contribution to this research was supported by the Research Foundation Flanders (FWO) under grant number S006323N and the Flanders AI research program. Johannes Kruse's contribution to this research is supported by the Innovation Foundation Denmark under grant number 1044-00058B and Platform Intelligence in News under project number 0175-00014B. Alain Starke's contribution was in part supported by the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the Centre for Research-based Innovation scheme, project number 309339. Jordi Viader Guerrero's contribution is supported by the TU Delft AI Labs Initiative and the AI DeMoS Lab. Nava Tintarev's contribution is supported by the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), RTL, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023. *All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.*

References

[1] M. D. Ekstrand, M. Tian, M. R. I. Kazi, H. Mehrpouyan, D. Kluver, Exploring author gender in book rating and recommendation, in: Proceedings of the 12th ACM conference on recommender systems, 2018, pp. 242–250.

[2] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, F. Diaz, Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems, in: Proceedings of the 27th acm international conference on information and knowledge management, 2018, pp. 2243–2251.

[3] E. Purificato, L. Boratto, E. W. De Luca, Do graph neural networks build fair user models? assessing disparate impact and mistreatment in behavioural user profiling, in: Proceedings

of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 4399–4403.

[4] S. Vrijenhoek, G. Bénédict, M. Gutierrez Granada, D. Odijk, M. De Rijke, Radio – rank-aware divergence metrics to measure normative diversity in news recommendations, in: Proceedings of the 16th ACM Conference on Recommender Systems, RecSys '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 208–219. URL: https://doi.org/10.1145/3523227.3546780. doi:10.1145/3523227.3546780.

[5] S. Buckler, Normative theory, Theory and methods in political science 3 (2010) 156–180.

[6] J. J. Thomson, Normativity, 2010.

[7] T. A. Christiani, Normative and empirical research methods: Their usefulness and relevance in the study of law as an object, Procedia-Social and Behavioral Sciences 219 (2016) 201–207.

[8] B. C. Stahl, Morality, ethics, and reflection: a categorization of normative is research, Journal of the association for information systems 13 (2012) 1.

[9] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, M. Zhou, MIND: A large-scale dataset for news recommendation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3597–3606. URL: https://aclanthology.org/2020.acl-main.331. doi:10.18653/v1/2020.acl-main.331.