CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	ARTAI: an evaluation platform to assess societal risk of recommender algorithms
Authors	Ruan, Qin;Xu, Jin;Dong, Ruihai;Younus, Arjumand;Mai, Tai Tan;O'Sullivan, Barry;Leavy, Susan
Publication date	2024
Original Citation	Ruan, Q., Xu, J., Dong, R., Younus, A., Mai, T.T., O'Sullivan, B. and Leavy, S. (2024) 'Artai: an evaluation platform to assess societal risk of recommender algorithms'. arXiv. https://doi.org/10.48550/ARXIV.2409.12396
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://arxiv.org/pdf/2409.12396 - 10.48550/ARXIV.2409.12396
Rights	© 2024, the Authors. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-17 12:47:21
Item downloaded from	https://hdl.handle.net/10468/17311

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

arXiv:2409.12396v1 [cs.CY] 19 Sep 2024

ARTAI: An Evaluation Platform to Assess Societal Risk of Recommender Algorithms

Qin Ruan*
Jin Xu*
qin.ruan@ucdconnect.ie
jin.xu@ucd.ie
School of Computer Science and Insight SFI Research Centre for Data Analytics
University College Dublin
Dublin, Ireland

Ruihai Dong
ruihai.dong@ucd.ie
School of Computer Science and Insight SFI Research Centre for Data Analytics
University College Dublin
Dublin, Ireland

Arjumand Younus
arjumand.younus@ucd.ie
School of Information and Communication
University College Dublin
Dublin, Ireland

Tai Tan Mai
tai.tanmai@dcu.ie
School of Computing
Dublin City University
Dublin, Ireland

Barry O'Sullivan
b.osullivan@cs.ucc.ie
Insight SFI Research Centre for Data Analytics and School of Computer Science & IT
University College Cork
Cork, Ireland

Susan Leavy
susan.leavy@ucd.ie
Insight SFI Research Centre for Data Analytics
University College Dublin
Dublin, Ireland

Abstract

Societal risk emanating from how recommender algorithms disseminate content online is now well documented. Emergent regulation aims to mitigate this risk through ethical audits and enabling new research on the social impact of algorithms. However, there is currently a need for tools and methods that enable such evaluation. This paper presents ARTAI, an evaluation environment that enables large-scale assessments of recommender algorithms to identify harmful patterns in how content is distributed online and enables the implementation of new regulatory requirements for increased transparency in recommender systems.

Keywords

Recommender Algorithms, Societal Risk Evaluation, Simulation System

ACM Reference Format:

Qin Ruan, Jin Xu, Ruihai Dong, Arjumand Younus, Tai Tan Mai, Barry O'Sullivan, and Susan Leavy. 2024. ARTAI: An Evaluation Platform to Assess Societal Risk of Recommender Algorithms. In *Proceedings of FAccTRec 2024: The 7th Workshop on Responsible Recommendation. In 18th ACM Conference on Recommender Systems (RecSys '24).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Societal risks emanating from how content is disseminated by recommender algorithms in online platforms to increase engagement are now well documented [3, 4, 9]. The Digital Services Act (DSA) [1] currently being implemented across the European Union (EU) address these issues and mandates improved algorithmic transparency through third-party auditing. It also recognises the crucial role of research, granting access rights for independent researchers to evaluate the social impacts of digital platforms [**?**]. Many countries are also introducing online safety legislation to ensure the protection of children using online platforms and mandating increased levels of transparency (e.g.[6, 12]). To enable researchers and auditors to conduct risk assessments of recommender algorithms there is a need for the development of evaluation tools [10]. ARTAI (Assessing Risk for Trustworthy AI) is an evaluation platform to address this requirement to assess the social impact of recommender algorithms.

Current approaches to risk evaluation involve analysis of code, user surveys, data scraping, sock-puppet audits and crowd-sourcing. Code audits are increasingly infeasible given the use of deep learning and the dynamic nature of such systems in live environments. Approaches involving collecting data from users or setting up sample profiles in sock-puppet audits can have limitations in terms of scaling to very large samples [2]. To overcome these limitations and provide scalable and comprehensive risk assessments this project developed a simulation environment. These environments allow researchers to systematically measure user engagement and assess the societal impact of recommender algorithms in a controlled setting. They can also address issues pertaining to user data in research through the generation of synthetic user data. Simulation frameworks such as Google's RecSim [7], Alibaba's Virtual-Taobao [13] and Microsoft's MindSim [8] have been developed primarily for use by developers of recommendation algorithms to test and refine their systems. This project builds on this work developing a prototype simulation environment for auditors, researchers and industry to assess risk in recommender algorithms.

*Both authors contributed equally to this research.

RecSys '24, October 14–18, 2024, Bari, Italy
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

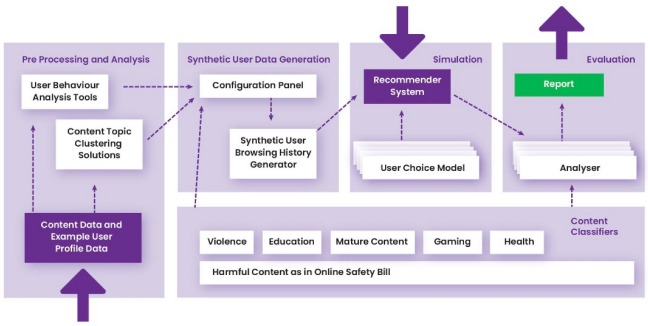


Figure 1: ARTAI Components

2 Evaluation Platform Design

ARTAI is designed as a simulation platform for recommendation algorithms to impart transparency and enable large-scale evaluation of trends in what content is suggested to users online. The simulation environment is integrated within a user interface designed for platform regulators, third-party auditors and industry. It allows for sample profiles of users and different models of online interaction to be generated and interaction with the recommendation algorithm simulated. The output of the algorithms, the recommended content, is categorised and trends are compiled in an interpretable report which visualises what kind of content algorithms are recommended. This approach allows for the identification of both instances of harmful content and also potentially harmful content distribution trends. The overall aim is that societal risks, particularly risks to vulnerable groups may be identified before the negative consequences become evident in society, thus reflecting a core aim of the DSA which is to "foresee", rather than witness societal risk.

The platform comprises five main components: *Pre-Processing and Analysis*, *Content Classifiers*, *Synthetic User Data Generation*, *Simulation* and *Risk Evaluation* (see Figure 1).

Pre-Processing and Analysis: Given a large-scale dataset such as videos or text, along with profile and online behaviour data of users, this component provides a suite of tools to pre-process the data and uncover patterns and trends that can be used to generate data synthetically. For instance, user behaviour analysis tools track and analyse interactions across various dimensions such as engagement frequency and content preferences, to uncover behavioural trends. Content topic clustering solutions categorise and group content based on topics to reveal underlying structures within the data. These insights can then inform the generation of large-scale synthetic user data for simulations. In our work, we utilised a recently released large-scale dataset of short videos MicroLens [11]. This dataset includes extensive user-item interaction data along with rich visual and textual information such as the original video content, comments and titles. These details enable a more fine-grained classification of the items and facilitate the learning of user preference distributions for data generation.

Content Classifiers: This component includes a number of trained classifiers designed to identify various types of videos or text. The primary goal is to discern the kind of content being recommended, rather than merely identifying harmful content. The content classification tool developed by ARTAI incorporates advanced natural language processing (NLP) technologies to ensure that classification outcomes are both explainable and meaningful. This approach enhances the transparency and utility of the tool, enabling users to understand and trust the classification process. It also allows researchers studying the social impact of recommender algorithms to define and investigate new topics.

Synthetic User Data Generation: This component involves example user groups that can be specified and synthetic data generated. Groups of users can be set up with different interest distributions. For instance, we may be interested in users who have the same interests but marginal differences in their browsing histories in relation to a particular topic. Setting up this interest distribution will allow for the simulation of recommender system outputs to capture the consequences of even marginal changes in online behaviour for user groups.

Simulation: The use of simulation is crucial for understanding how different recommendation strategies influence user behaviour and content dissemination over time. By creating a controlled environment, we can systematically analyse the impact of various factors and conditions on the outcomes of recommender systems. Multiple user choice models are represented in the platform [5] and this is used to simulate the interaction between users and recommendation algorithms. For example, in a position-based user choice model, users are most likely to select items recommended at the top of a personalised recommendation list. How actively a user interacts with recommender systems over time will also be included as part of the modelling.

Risk Evaluation: The platform will develop reports based on collated patterns and trends of recommended content for synthetic user data, simulated interactions and the kind of personalised recommendations suggested to users over time. This component will highlight how certain user profiles and online behaviours give rise to different content recommendations. The content classifiers allow for the evaluation, not only of what conditions may lead to the recommendation of particular harmful content types, but also of the proportion of content categories that are recommended overall and how this changes over time.

3 Conclusion

This paper presented ARTAI, a simulation environment for the evaluation of recommender algorithms for societal risk. The platform aims to address the lack of oversight into what kind of content is being recommended to different user groups, which is a particular issue concerning vulnerable groups such as children. ARTAI is designed primarily for use by auditors and vetted researchers to enable the implementation of the European Union DSA. While simulation environments have been used in industry (e.g. [7, 8, 13]), they have been created with developers in mind and require advanced technical skill to implement and use. The ARTAI system is designed from the outset with the full range of stakeholders involved resulting in an accessible user interface. The goal of this project is to equip auditors with tools to conduct ethics evaluations and increase the scale at which researchers across disciplines can evaluate the social impact of online platforms.

Acknowledgments

This publication has emanated from research conducted with the financial support of the EU Commission Recovery and Resilience Facility under the Science Foundation Ireland OurTech Challenge Grant Number 22/NCF/OT/11077 and with the financial support of Science Foundation Ireland 12/RC/2289_P2 at Insight the SFI Research Centre for Data Analytics at University College Dublin.

References

[1] 2020. Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) 2000/31/EC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A825%3AFIN

[2] Ada Lovelace Institute. 2021. *Technical Methods for Regulatory Inspection of Algorithmic Systems.* Technical Report. Ada Lovelace Institute, London, UK. https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/

[3] Catherine Baker, Debbie Ging, and Maja Brandt Andreasen. 2024. *Recommending Toxicity: The role of algorithmic recommender functions on YouTube Shorts and TikTok in promoting male supremacist influencers.* Technical Report. Anti-Bullying Centre, Dublin City University.

[4] Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. 2022. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning*. PMLR, 2686–2708.

[5] Naieme Hazrati and Francesco Ricci. 2022. Simulating users' interactions with recommender systems. In *Adjunct proceedings of the 30th acm conference on user modeling, adaptation and personalization*. 95–98.

[6] Houses of the Oireachtas. 2022. Online safety and media regulation Act 2022. https://www.oireachtas.ie/en/bills/bill/2022/6. Accessed: 2024-8-7.

[7] Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. 2019. Recsim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847* (2019).

[8] Xufang Luo, Zheng Liu, Shitao Xiao, Xing Xie, and Dongsheng Li. 2022. Mindsim: user simulator for news recommenders. In *Proceedings of the ACM Web Conference 2022*. 2067–2077.

[9] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *Ai & Society* 35 (2020), 957–967.

[10] Jakob Mökander, Maria Axente, Federico Casolari, and Luciano Floridi. 2022. Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. *Minds and Machines* 32, 2 (2022), 241–268.

[11] Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. 2023. A Content-Driven Micro-Video Recommendation Dataset at Scale. *arXiv preprint arXiv:2309.15379* (2023).

[12] Parliament of the United Kingdom. 2023. Online Safety Act 2023. https://www.legislation.gov.uk/ukpga/2023/50/pdfs/ukpga_20230050_en.pdf. Accessed: 2024-9-18.

[13] Jing-Cheng Shi, Yang Yu, Qing Da, Shi-Yong Chen, and An-Xiang Zeng. 2019. Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4902–4909.