Model-based Algorithmic Auditing of Social Media AI Algorithms

IVAN SRBA, Kempelen Institute of Intelligent Technologies, Slovakia
BRANISLAV PECHER, Kempelen Institute of Intelligent Technologies, Slovakia
JAKUB SIMKO, Kempelen Institute of Intelligent Technologies, Slovakia
ROBERT MORO, Kempelen Institute of Intelligent Technologies, Slovakia
MARIA BIELIKOVA, Kempelen Institute of Intelligent Technologies, Slovakia

This position paper introduces a novel paradigm for oversight of social media AI algorithms, called the *model-based algorithmic auditing*. In general, the algorithmic auditing is a process of automated dynamic black-box assessment of real-world software system behavior. In a so-called sockpuppeting audit, impostor bots stimulate the platform with a simulated user behavior and observe the responses of the audited system (e.g., a recommender system). Algorithmic auditing is able to disclose interesting traits of AI systems (e.g., biases), which may be otherwise opaque. However, technical and methodological difficulties make audits costly, hard to reproduce, and hard to transfer cross-platform and cross-domain. To overcome this, the *model-based algorithmic auditing* introduces a platform-agnostic social media model which provides a simplified and aggregated representation of users, content, and interactions between them. The model supports or even automates challenging steps of the audit, like assisting human experts in creation of abstract audit scenarios, or predicting next user interactions. The reduction of manual effort makes the auditing more representative, cross-platform, and longitudinal, ultimately enabling more efficient oversight of social media algorithms by regulators, auditors and other stakeholders.

Keywords: Algorithmic auditing, Social media, Model-based auditing, Algorithmic fairness, Recommender systems

Reference Format:
Ivan Srba, Branislav Pecher, Jakub Simko, Robert Moro, and Maria Bielikova. 2025. Model-based Algorithmic Auditing of Social Media AI Algorithms. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 7 pages.

1 Introduction

A number of recent research studies [10, 14, 15], EU and NGOs' reports [6] as well as real-world cases [4], demonstrate significant negative societal and psychological effects of social media AI algorithms (recommender systems, search engines or advertisement systems). Some roots of this problem can be traced to users' as well as algorithms' biases (e.g., preference of viral, sensational, or polarizing content), and feedback loops of algorithms [16]. User generated content, user behavior, and AI algorithms together compound a socio-technical dynamic ecosystem, in which echo chamber effect [1] causes that such biases (even if negligible at first) are continuously accumulating [6].

Authors' Contact Information: Ivan Srba, Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia, ivan.srba@kinit.sk; Branislav Pecher, Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia, branislav.pecher@kinit.sk; Jakub Simko, Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia, jakub.simko@kinit.sk; Robert Moro, Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia, robert.moro@kinit.sk; Maria Bielikova, Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia, maria.bielikova@kinit.sk.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

In Europe, oversight and regulation over these phenomena has already been legislated, yet auditors, regulators, as well as researchers lack proper tooling to assess the platforms' AI algorithmic systems. Our analysis of *DSA audit reports*[1] for five VLOPs (TikTok, YouTube, Instagram, Facebook, LinkedIn) and one VLOSE (Google Search), all published in Q4 2024, revealed that some DSA obligations related to recommender system transparency (Articles 27 & 38 of the DSA), the protection of minors (Article 28), and advertising practices (Articles 26 & 39) - were not always assessed positively or could not be fully audited. The latter was often due to the absence of appropriate tools, insufficient documentation, or auditors' inability to understand platform's systems.

We believe that the key tool, which is currently missing, are independent *algorithmic audits*. They can be defined as "assessments of the algorithm's negative impact on the rights and interests of stakeholders, with a corresponding identification of situations and/or features of the algorithm that give rise to these negative impacts" [3]. Since social media AI algorithms are black-boxes (we cannot analyze or influence their inner workings), audits must explore their properties behaviorally: user interactions with an AI algorithm are simulated (e.g., content visits), and observed responses (e.g., recommended items) are examined for the presence of the audited phenomenon. Typically, bots [5, 17] or human agents [11] are employed to simulate such user interactions in so-called sockpuppeting audits and crowdsourcing (collaborative) audits respectively [12]. If deployed with sufficient coverage, algorithmic audits can help to obtain a deeper understanding of negative effects of social media AI algorithms and to improve their algorithmic fairness as well as independently validate DSA audit reports by VLOPs and VLOSEs.

A typical audit consists of several steps. Based on an initial *audit question*, *audit scenarios* are specified, comprising user profiles (personal information, user's interaction history) and user actions that simulate user behavior (e.g., interactions with the pre-selected or presented/recommended content). Next, during *audit execution and evaluation*, bots/human agents execute the created audit scenarios, platform responses are recorded, and further investigated for the presence of *audited phenomenon* (e.g., tendencies for harmful content recommendation).

Unfortunately, although technically possible, current *first-generation algorithmic audits* require lot of effort and expertise [13], which renders them impractical for regulatory purposes. Specifically, they excessively rely on manual inputs: from specification of audit scenarios to audit execution and evaluation. Audit scenarios are created ad-hoc on intuition of researchers, incomplete (covering only a small subset of relevant user/content/interaction space) and inauthentic (heavily prescribed or too random). At the same time, only few attempts have been made to automate the labeling of platform responses (e.g., for the presence of harmful content [9, 15, 17]). Consequently, current audits are oversimplified and artificial, and do not reflect the complexity of real-world social media environment. They are also limited in scope and replicability, platform dependent, and executed for short-time intervals only.

We aim to fundamentally change the way we audit social media AI algorithms. Our main idea is a novel paradigm of *model-based algorithmic auditing*. With this paradigm, our vision is to have large-scale next-generation audits that will use an appropriate level of automation to overcome the limitations of the current audits. Such audits should be: 1) representative – they will comprehensively and systematically cover social media environment and authentically replicate user behavior; and 2) cross-platform, longitudinal and multilingual – they will be able to perform highly demanded comparison/benchmark of AI algorithms across multiple platforms and temporal dimension (continuously over time - to capture trends and changes in an AI) and spatial dimension (over multiple languages - to identify whether AI algorithms do not struggle with some specific languages, especially minor ones).

[1] Under Article 37 of the Digital Services Act (DSA), Very Large Online Platforms (VLOPs) and Very Large Search Engines (VLOSEs) are required to undergo independent annual audits conducted by qualified external auditors to assess their compliance with DSA obligations.

The main obstacles why such next-generation audits have not been conducted yet is that in order to execute them, it is necessary to tackle with dynamic and complex nature of social media and AI algorithms, and overcome absence of (labeled) data. Fortunately, the recent progress in machine learning and natural language processing, like learning with limited labeled data or large language models, provides us with the necessary grounds to such research.

2 Paradigm of Model-based Algorithmic Auditing

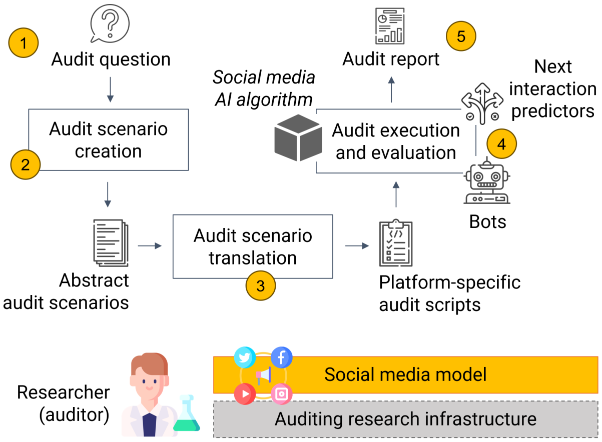
1
Audit question
2
Audit scenario creation
Abstract audit scenarios
Audit scenario translation
3
Platform-specific audit scripts
Social media AI algorithm
Audit execution and evaluation
Audit report
5
Next interaction predictors
4
Bots
Researcher (auditor)
Social media model
Auditing research infrastructure

Fig. 1. Process of model-based algorithmic auditing

In order to address the drawbacks of first-generation algorithmic audits, we propose the *model-based algorithmic auditing* – a novel paradigm in which a *social media model* is built and continuously maintained to support and automate the challenging steps of auditing process (see Figure 1).

The process of *model-based algorithmic auditing* follows these steps: (1) A researcher (an auditor) formulates the *audit question*, for example: "how does the prevalence of persuasive antivax false claims in YouTube's search engine differs for various age groups". (2) The researcher interactively defines a set of *abstract audit scenarios*. In the process, she queries the *social media model* to determine user profiles (e.g., to achieve a representative age/gender/location distribution corresponding to the audited platforms) and to construct user actions (e.g., to select phrases a user should search for, such as "vaccination causes autism"; or to determine how many videos should be watched/skipped/up-voted to faithfully mimic the behavior of users on the audited platforms). (3) Next, such abstract audit scenarios are translated to *platform-specific audit scripts* (e.g., a user profile is converted to a real YouTube user account, topics are matched to the actual YouTube videos, user actions are mapped to their corresponding platform implementations). (4) During audit execution, the bots follow the prescribed audit scripts. User interactions can be either pre-determined by the script itself or automatically and dynamically predicted with utilization of *next interaction predictors*. The reactions of the platforms' AI algorithms (e.g., recommended videos) are observed and recorded. (5) The recorded reactions are automatically annotated for the presence of the audited phenomenon (e.g., the presence of disinformation claims in the recommended videos) and quantified in the resulting *audit report*.

The *social media model*, the key underlying element of the proposed paradigm, provides a unique (previously non-existing) representation of: 1) user population on social media platforms (e.g., distribution of users' age, gender and location), 2) (harmful) content (e.g., distribution of topics or fact-checked false claims), and 3) user interactions (e.g., distribution of occurrence of various user actions, either for a whole population or for a specific user stereotypes). The social media model is semi-automatically derived from outcomes of research studies (e.g., social science studies examining user behaviour on social media platforms), public opinion polls, from harmful content combating activities (e.g., fact-checking), as well as from the real-world social media platforms' data.

The social media model is platform-agnostic – it jointly models users, content, and interactions using the same representation for multiple social media platforms.

The *abstract audit scenarios*, in contrast with the previous first-generation audits, also use a platform- and content-agnostic abstraction. Instead of specific user accounts, abstract scenarios define the representative user population to be simulated during the audit (e.g., the audit will employ 100 bots, 56% of them are female, 44% are male). Instead of user actions with the predetermined content (e.g., a bot should like/follow these manually preselected videos or authors), abstract scenarios determine platform-agnostic interactions (e.g., up-vote) and more latent content representations (e.g., keywords to be used to identify currently available/popular content). As such abstract scenarios are not directly executable, audit scenario translation is executed before each audit run and convert scenarios to the executable *platform-specific scripts*. This makes it possible to execute the "same audit" on multiple platforms, repetitively over longer time periods, and in multiple languages; as well as to make the audit outcomes directly comparable.

An easily extendable set of *next interaction predictors* allows to simulate more organic user behavior. Instead of determining user actions (e.g., a decision whether to skip or like a recommended video) randomly or by simple heuristics (e.g., if a video has any of these hashtags, watch it until the end), we introduce a possibility for more advanced decision mechanisms. Next interaction predictors are able to fully consider characteristics of a user as well as currently displayed content to determine the next action. To this end, the content may be automatically annotated (e.g., whether it is a harmful content or not), and appropriate interaction type can be determined either by a rule-based system (that reflects the intention of the audit question) or even by querying the social media model.

The whole paradigm of model-based algorithmic auditing is following a *human-centered approach*. Ubiquitous human oversight is achieved by adopting a human-in-the-loop approach, from creation of abstract audit scenarios to data annotation and audit evaluation. Stakeholders (e.g., regulators) are also involved in formulating audit questions and interpreting the results provided in the audit reports.

3 Implementation and Preliminary Results

The implementation of the *model-based algorithmic auditing*, requires a research of novel (AI-based) methods for: 1) representation and extraction of social media model, 2) audit scenario creation, 3) next user interaction prediction, and 4) (real-time) content annotation. Moreover, an underlying technical (software and hardware) infrastructure is required to streamline the auditing process (e.g., to automatically translate scenarios and execute bot behavior; or to automatically annotate the social media platform content). At present, no such methods and infrastructure exist, and thus, we are developing them as a part of the *AI-Auditology platform*[2].

We aim for flexibility and extendability of the platform to support a wide range of future auditing assignments. To this end, the platform allows to easily reuse existing or add new above-mentioned methods supporting or automating individual steps of the auditing process. To perform the sockpuppeting audits, the technical infrastructure implements bots interacting with the audited platforms. In this direction, we build upon our developed proof-of-concept solution that was used in our awarded work [17]. Since platforms regularly change their interfaces, which can break functioning of bots (e.g., the platform-specific implementation for clicking on a like/upvote button may no longer work), we implement appropriate monitoring measures to spot and fix these changes in a timely manner.

[2] https://kinit.sk/project/ai-auditology

We also aim to support high interpretability and transparency of the audit process. All quantitative results (e.g., differences between control and experimental groups of bots) are verified for statistical significance by statistical tests and calculation of effect sizes, and accompanied with corresponding confidence intervals. In this way, we aim to avoid misinterpretation of the audit results as well as transparently report the accuracy of the underlying techniques. Even if the partial methods may introduce a small inevitable level of noise (e.g., inaccuracies in content annotation), the large-scale nature of the audits (e.g., comprehensive sets of audit scenarios that will be run multiple times) will suppress a possible influence of such random noise.

As we develop the platform, we demonstrate parts of its functionality in a pilot TikTok audit. TikTok is the only platform that received a "negative" appraisal regarding compliance with recommender system transparency requirements (Article 27 and 38 of the DSA). In a preliminary study [7], we reproduced several representative first-generation audits of the personalization factors on TikTok [2, 8, 18], with a goal to better understand the challenges posed by the platform (e.g., anti-bot measures). We observed poor reproducibility of the existing auditing studies, due to constant content changes (e.g., average video diversity between repeated runs of the audits is as high as 70%), and platform changes (e.g., user interfaces change frequently and differs across countries). Furthermore, we observe significant changes in the behavior of the platform, e.g., increased importance of implicit actions (watching videos), but also stronger focus on preventing automated behavior and banning harmful content. Finally, we observe that audit outcomes heavily depend on evaluation metric selection. All these findings reinforce our motivation for a more authentic, structured, multiplatform, and longitudinal audits that can more faithfully discern whether the changes in findings are only a result of content change or due to enacted policies and algorithm changes.

4 Conclusion

The contribution of this position paper is twofold. First, we describe the current generation of social media algorithmic audits and their drawbacks preventing their wider adoption. Second, we introduce a novel concept of model-based algorithmic auditing, the current state of its realization, and our future plans.

The proposed model-based algorithmic auditing and next-generation audits fall into a so-far under-researched area of algorithmic auditing. With this novel concept, we aim to contribute to a new and impactful research direction, which addresses challenges hampering further progress in auditing social media AI algorithms. To achieve this, we are researching novel methods and models supporting auditing process and implement the necessary research infrastructure. Our next steps will focus on auditing the selected obligations most commonly assessed negatively in the DSA audit reports. In addition, we aim to test the proposed solution by means of a longitudinal audit on multiple social media platforms, which will demonstrate the capabilities of the proposed solution.

Acknowledgments

This work was partially supported by the project AI-Auditology funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V03-00020.

In addition, we would like to thank Adrian Bindas, Marek Havrila, Filip Hossner, Jan Jakubcik, Simon Liska, Matej Mosnar, Peter Sakalik, Adam Skurla, Andrej Suty, Matus Tibensky, Matus Tuna for their participation in research and development activities within AI-Auditology project, as well as Matus Mesarcik, Sara Solarova, Juraj Podrouzek and Katarina Marcincinova for their contributions to ethical and legal aspects of the project.

References

[1] Alessandro Bessi. 2016. Personality traits and echo chambers on facebook. *Computers in Human Behavior* 65 (2016), 319–324. https://doi.org/10.1016/j.chb.2016.08.016

[2] Maximilian Boeker and Aleksandra Urman. 2022. An Empirical Investigation of Personalization Factors on TikTok. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2298–2309. https://doi.org/10.1145/3485447.3512102

[3] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8, 1 (2021), 2053951720983865. https://doi.org/10.1177/2053951720983865 arXiv:https://doi.org/10.1177/2053951720983865

[4] Jeff Horwitz. 2021. *The Facebook Files; A Wall Street Journal investigation*. Retrieved March 10, 2025 from https://www.wsj.com/articles/the-facebook-files-11631713039

[5] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 48 (May 2020), 27 pages. https://doi.org/10.1145/3392854

[6] Alistair Knott, Kate Hannah, Dino Pedreschi, Andrew Trotman, Ricardo Baeza-Yates, Rituparna Roy, David Eyers, Virginia Morini, and Valentina Pansanella. 2021. *Responsible AI for Social Media Governance A proposed collaborative method for studying the effects of social media recommender systems on users*. Retrieved March 10, 2025 from https://gpai.ai/projects/responsible-ai/social-media-governance/responsible-ai-for-social-media-governance.pdf

[7] Matej Mosnar, Adam Skurla, Branislav Pecher, Matus Tibensky, Jan Jakubcik, Adrian Bindas, Peter Sakalik, and Ivan Srba. 2025. Revisiting Algorithmic Audits of TikTok: Poor Reproducibility and Short-term Validity of Findings. *arXiv preprint arXiv:2504.18140* (2025).

[8] Sepehr Mousavi, Krishna P Gummadi, and Savvas Zannettou. 2024. Auditing Algorithmic Explanations of Social Media Feeds: A Case Study of TikTok Video Explanations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1110–1122.

[9] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2020. "It is just a flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. arXiv:2010.11638 [cs.CY]

[10] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2021. "How over is it?" Understanding the Incel Community on YouTube. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 412 (Oct. 2021), 25 pages. https://doi.org/10.1145/3479556

[11] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 131–141. https://doi.org/10.1145/3351095.3372879

[12] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014), 4349–4357.

[13] Jakub Simko, Matus Tomlein, Branislav Pecher, Robert Moro, Ivan Srba, Elena Stefancova, Andrea Hrckova, Michal Kompan, Juraj Podrouzek, and Maria Bielikova. 2021. Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 411–414. https://doi.org/10.1145/3450614.3463353

[14] Larissa Spinelli and Mark Crovella. 2020. How YouTube Leads Privacy-Seeking Users Away from Reliable Information. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 244–251. https://doi.org/10.1145/3386392.3399566

[15] Ivan Srba, Robert Moro, Matus Tomlein, Branislav Pecher, Jakub Simko, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Adrian Gavornik, and Maria Bielikova. 2023. Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. *ACM Trans. Recomm. Syst.* 1, 1, Article 6 (Jan. 2023), 33 pages. https://doi.org/10.1145/3568392

[16] Coalition to Fight Digital Deception. 2021. *Trained for Deception: How Artificial Intelligence Fuels Online Disinformation: A report from the Coalition to Fight Digital Deceptionn*. Retrieved March 10, 2025 from https://www.newamerica.org/oti/policy-papers/trained-for-deception-how-artificial-intelligence-fuels-online-disinformation/

[17] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. 2021. An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior

Changes. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3460231.3474241

[18] Karan Vombatkere, Sepehr Mousavi, Savvas Zannettou, Franziska Roesner, and Krishna P. Gummadi. 2024. TikTok and the Art of Personalization: Investigating Exploration and Exploitation on Social Media Feeds. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 3789–3797. https://doi.org/10.1145/3589334.3645600