LOU RIM

LOUVAIN RESEARCH INSTITUTE IN MANAGEMENT AND ORGANIZATIONS

WORKING PAPER

2026/03

TRACE: A scalable and extensible framework for auditing YouTube's recommendation algorithm

Colin Timmers,
Corentin Vande Kerckhove,
Louvain Research Institute in Management and Organizations

Quentin Dessain,
Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM)

Louvain Research Institute in Management and Organizations
Working Paper Series
Editor: Prof. Karine Charry
(president-lourim@uclouvain.be)

TRACE: A scalable and extensible framework for auditing YouTube's recommendation algorithm

Colin Timmers, Louvain Research Institute in Management and Organizations
Corentin Vande Kerckhove, Louvain Research Institute in Management and Organizations
Quentin Dessain, Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM)

Summary

Auditing large-scale recommender systems like YouTube remains a methodological challenge due to the trade-off between behavioral realism, scalability, and reproducibility. We present TRACE, an open-source framework that integrates Large Language Models to simulate context-aware, persona-driven user journeys. TRACE combines containerized browser automation, database-backed traceability, and asynchronous data enrichment to enable reproducible large-scale audits of YouTube's recommendation ecosystem. By decoupling experimental contexts from personas and supporting multiple behavioral modes, it allows researchers to model diverse user identities and explore how recommendation dynamics evolve over time. The framework's modular architecture, reproducible design, and web-based control interface facilitate transparent, comparative studies of algorithmic personalization and potential filter bubble formation. TRACE establishes a scalable foundation for empirically grounded, extensible, and ethically sound auditing of recommender systems.

Keywords : YouTube; Recommendation algorithm; Algorithmic auditing; Persona-driven simulation; Large Language Models;

JEL Classification: L66.

This work was supported by an FSR Seedfund grant from UCLouvain

Corresponding author :
Colin Timmers
Louvain Louvain Research Institute in Management and Organizations / Campus Mons
Université catholique de Louvain
M1.01.01
Chaussée de Binche 151
7000 Mons, BELGIUM

TRACE: A Scalable and Extensible Framework for Auditing YouTube's Recommendation Algorithm

Quentin Dessain
UCLouvain
Institute of Information and Communication Technologies, Electronics and Applied Mathematics
Louvain-la-Neuve, Belgium
quentin.dessain@uclouvain.be

Colin Timmers
UCLouvain
Louvain Research Institute of Organizations and Management
Louvain-la-Neuve, Belgium
colin.timmers@uclouvain.be

Corentin Vande Kerckhove
UCLouvain
Louvain Research Institute of Organizations and Management
Louvain-la-Neuve, Belgium
corentin.vandekerckhove@uclouvain.be

Abstract

Auditing large-scale recommender systems like YouTube remains a methodological challenge due to the trade-off between behavioral realism, scalability, and reproducibility. We present TRACE, an open-source framework that integrates Large Language Models to simulate context-aware, persona-driven user journeys. TRACE combines containerized browser automation, database-backed traceability, and asynchronous data enrichment to enable reproducible large-scale audits of YouTube's recommendation ecosystem. By decoupling experimental contexts from personas and supporting multiple behavioral modes, it allows researchers to model diverse user identities and explore how recommendation dynamics evolve over time. The framework's modular architecture, reproducible design, and web-based control interface facilitate transparent, comparative studies of algorithmic personalization and potential filter bubble formation. TRACE establishes a scalable foundation for empirically grounded, extensible, and ethically sound auditing of recommender systems.

CCS Concepts

• **Information systems → Recommender systems**; **Evaluation of retrieval results**.

Keywords

YouTube; Recommendation algorithm; Algorithmic auditing; Persona-driven simulation; Large Language Models; Filter bubbles

ACM Reference Format:
Quentin Dessain, Colin Timmers, and Corentin Vande Kerckhove. 2026. TRACE: A Scalable and Extensible Framework for Auditing YouTube's Recommendation Algorithm. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining (WSDM '26), February 22–26, 2026, Boise, ID, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3773966.3779410

1 Introduction

YouTube's recommendation system plays a crucial role in shaping how billions of users discover and engage with information, thereby influencing global political discourse and societal norms. Concerns have been raised through controlled audits that its algorithmic feedback loops may inadvertently create "filter bubbles", potentially reinforcing users' existing beliefs, spreading misinformation, or even leading them down paths toward ideologically extreme content [6, 11]. However, empirical studies often show that such content constitutes only a minor part of what users typically watch, highlighting a significant disconnect between controlled audits and actual user behavior [1].

This discrepancy often arises from the methodological limitations of existing auditing techniques [2]. Many auditing frameworks depend on rigidly scripted agents, which, while offering reproducibility, fail to capture the nuances of human decision-making and curiosity. Consequently, there is a pressing need for a tool that can bridge the gap between behavioral realism, experimental control, and large-scale reproducibility.

To address this challenge, we introduce TRACE (Transparent Recommendation Auditing and Collection Engine). TRACE is an open-source framework designed for scalable and extensible audits of YouTube's recommendation algorithm. It leverages Large Language Models (LLMs) to simulate realistic, persona-driven user journeys, providing a more nuanced approach to understanding algorithmic behavior.

2 Related work

The practice of auditing YouTube's recommendation algorithm spans a range of methodologies, each with its own strengths and weaknesses. System-level analyses, which often use the YouTube API or third-party data, are effective in mapping the structural connections between topics and channels [10, 12]. While these methods can uncover inherent biases in the platform's content network, they do not account for personalization, as they lack behavioral inputs.

To capture these individualized dynamics, many researchers employ automated agents, often called "sock puppets," to simulate user interactions like watching or clicking on videos. Some bots are programmed to follow predetermined viewing paths to analyze reorientation effects [13], while others mimic specific interests or explore recommendation trees through various strategies such as autoplay or random walks [4, 5]. More sophisticated designs incorporate human-like behaviors, including varied timing and scrolling, to enhance realism [8]. Recent advancements have even introduced reinforcement learning to create adaptive auditing bots that can dynamically adjust their engagement based on content [9].

WSDM '26, Boise, ID, USA

ACM ISBN 979-8-4007-2292-9/2026/02
https://doi.org/10.1145/3773966.3779410

However, these approaches present a trade-off. Highly scripted bots offer excellent control and reproducibility but often lack the adaptability needed to elicit genuine algorithmic responses, making their findings less representative of real-world user experiences. While studies involving real users provide the most authentic insights into YouTube's recommendation dynamics, they are expensive, difficult to scale, and pose significant ethical challenges, particularly when it involves exposing participants to potentially harmful content [1, 7].

Recent work has proposed LLM-based user simulation frameworks for recommender systems, aiming to generate proxies for user behavior. Agent4Rec and RecUserSim, for instance, emphasize behavioral realism by equipping agents with personas, memory, and decision modules to approximate real user behavior [3, 14].

This review of existing methods reveals a clear gap in the field: the need for a framework that combines the behavioral realism of human users with the scalability, control, and ethical safety of automated agents. TRACE is designed to fill this void by using LLMs to simulate psychologically plausible user behavior at scale, offering a transparent and reproducible alternative to current auditing techniques.

3 Framework design and methods

TRACE is engineered to simulate human-like user journeys at scale, enabling researchers to conduct rigorous and reproducible audits of YouTube's recommendation algorithm. Its core innovation lies in the use of LLM-driven "personas" to navigate the platform, creating realistic browsing patterns that can trigger genuine algorithmic personalization. Each audit conducted with TRACE is structured into a two-phase process to ensure both experimental control and behavioral realism.

3.1 A two-phase simulation methodology

3.1.1 First phase: contextual initialization. To ensure that comparisons between different experimental runs are meaningful, each audit begins with a context-setting phase. This phase establishes a standardized algorithmic baseline for each agent. This is achieved by having the agent watch a predefined list of "warm-up" videos, which are stored as a reusable "context." During this stage, the agent sequentially visits each video in the context list, simulating a watch event and collecting the recommendations that are generated. By decoupling the initial algorithmic exposure (the context) from the agent's behavioral profile (the persona), TRACE allows for controlled, cross-persona comparisons, making it possible to attribute differences in recommendation pathways directly to the persona's behavior.

3.1.2 Second phase: persona-driven exploration. Here, the agent's navigation is guided by a specific persona defined as a detailed, natural-language description of a user's interests, biases, and viewing habits. This persona serves as a continuous input for an LLM, which makes decisions on behalf of the agent; while individual decisions are stochastic by the LLM's nature, TRACE relies on large numbers of parallel agents and aggregation across runs to average out random variation and recover stable behavioral signals. The agent operates in an iterative loop:

(1) **Data collection:** It scrapes all visible recommendations on the current video page.
(2) **Decision-making:** To simulate a user with heterogeneous interests, the agent probabilistically selects a persona from a weighted pool before each interaction.
(3) **Interaction:** The agent navigates to the selected video and simulates a viewing event.
(4) **Data logging:** Every recommendation, its ranking, the chosen video, and the LLM's justification for the choice are meticulously logged in a central database.

This loop repeats for a configurable number of steps. TRACE does not aim to reproduce individual-level human optimality, but rather to generate behaviorally plausible and internally consistent interaction patterns sufficient to compare algorithmic responses under controlled conditions.

3.2 Modeling behavioral diversity

TRACE's flexibility is enhanced by its support for multiple experimental modes, which dictate how personas are applied during the simulation. This allows researchers to model a wide range of user behaviors:

(1) **Single persona mode:** The agent consistently adheres to a single persona, representing a user with stable interests.
(2) **Mixed persona mode:** The agent randomly selects from a weighted list of personas at each step, simulating a user with multiple, diverse interests.
(3) **Sequential persona mode:** The agent switches between different personas after a set number of steps, modeling an evolution in a user's preferences over time.
(4) **Random choice mode:** The agent ignores all personas and selects videos at random, providing a neutral baseline to isolate the effects of algorithmic personalization.

3.3 System architecture

TRACE is built on a modular, container-based architecture designed for scalability and reproducibility. As shown in Figure 1, the framework consists of four decoupled components orchestrated via Docker Compose.

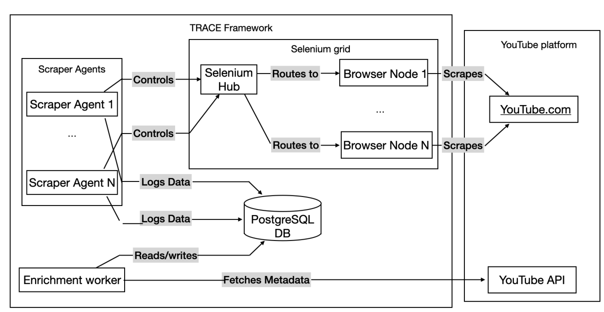

Figure 1: Overview of the TRACE architecture. Persona-driven agents interact with YouTube through a Selenium Grid, logging data to a central PostgreSQL database enriched asynchronously through the YouTube Data API.

The core components are:

- A central PostgreSQL database that serves as the system's backbone, ensuring complete data traceability and reproducibility.
- A scalable Selenium Grid for parallel browser automation, allowing multiple agents to run concurrently.
- Independent Scraper Agents, which are persona-driven bots that simulate human-like navigation.
- An asynchronous Enrichment Worker that fetches additional metadata and transcripts for collected videos using the YouTube Data API, preventing bottlenecks during the scraping process.

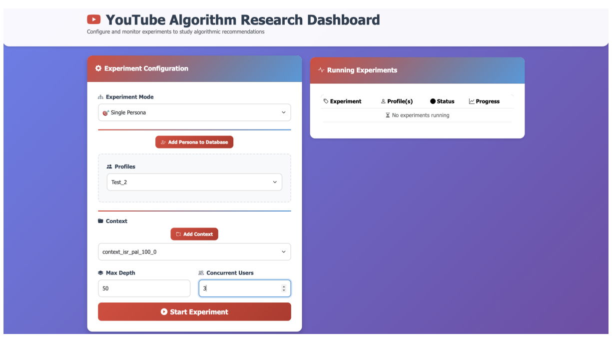

Figure 2: The web interface for managing TRACE experiments, which allows for the configuration of profiles, contexts, as well as real-time monitoring.

This modular design ensures that experiments are reproducible and can be scaled to run large-scale comparative audits with statistical robustness, while also allowing the underlying LLM to be easily swapped to study model-specific biases and variability. To enhance usability, TRACE can be operated either from the command line or through a web-based graphical user interface (GUI), as depicted in Figure 2. This interface allows researchers to easily configure experiments, define personas, and monitor progress in real time.

4 Research capabilities and applications

TRACE provides a scalable, flexible, and reproducible framework for auditing YouTube's recommendation algorithm. Its architecture enables the concurrent operation of many independent agents, allowing for large-scale, comparative, and statistically robust experiments. Personas can be easily created, combined, or adapted, making it possible to investigate evolving research questions about user behavior and algorithmic personalization without modifying the core system. The containerized design guarantees reproducibility, while structured logging ensures full transparency and auditability.

Beyond these general features, TRACE expands the analytical reach of algorithmic audits. It reduces the sensitivity of experiments to the initial context-setting phase by maintaining behavioral control throughout the entire navigation process. In conventional audits based on random selection, control over agent behavior typically ends once the initial exposure context has been defined. TRACE, by contrast, preserves behavioral coherence across multiple viewing and recommendation cycles. This continuity enables researchers to analyze how recommendation pathways evolve dynamically as simulated users interact with content, rather than restricting their analysis to the initial exposure moment.

This approach enables the study of radical pathways, where users consistently select content aligned with a specific ideology, providing insight into how reinforcement dynamics unfold. It also supports the exploration of mixed or evolving personas, revealing how YouTube's algorithm reacts to complex, shifting user profiles and how its memory adapts when a user transitions from one viewpoint to another.

The framework also supports comparative analyses that quantify the respective influence of user behavior and algorithmic personalization. Persona-driven experiments can be directly compared with random-choice baselines conducted under identical conditions, allowing researchers to isolate algorithmic effects from behavioral ones.

Finally, the framework facilitates systematic comparisons between opposing ideological personas under identical experimental conditions. Unlike random or single-context approaches, TRACE ensures that both sides of an ideological spectrum are explored through equally realistic, behaviorally consistent simulations. This feature yields balanced insights into how recommendation systems respond to differences in user identity and engagement patterns, contributing to more transparent and accountable auditing of algorithmic personalization.

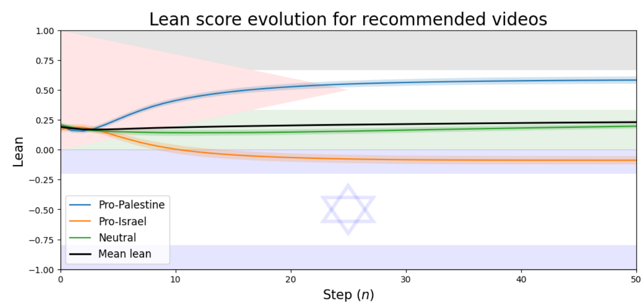

Figure 3: Evolution of rank-weighted lean scores for Pro-Palestine, Pro-Israel, and Neutral personas over fifty steps. All agents shared the same initialization context before diverging in single-persona mode.

Proof of Concept

To illustrate TRACE's capabilities, we conducted a controlled proof-of-concept experiment designed to assess the system's capacity to generate, control, and analyze ideologically contrasted recommendation trajectories. The objective was not to draw substantive conclusions about political bias or to validate user behavior, but rather to demonstrate how TRACE supports transparent, reproducible, and rank-aware audits of algorithmic personalization.

Given its sensitive nature, the Israeli-Palestinian conflict was selected as a challenging test case to evaluate the framework's robustness. The seed videos for initialization were chosen based on pre-registered, stance-balanced criteria and were applied uniformly across all agents to establish a shared baseline. All sessions were executed in a controlled, containerized environment with consistent region and language settings. To ensure reproducibility and cost-efficiency, agents utilized Mistral Small 3.2 Instruct (<$0.01/agent) and residential proxies ($0.10/agent) to mitigate bot detection. The framework demonstrated scalability on commodity hardware (Ubuntu desktop, Intel i7-8700k, 32GB RAM), supporting 10 concurrent instances (2GB RAM each).

Three distinct personas were defined in natural language: a Pro-Israel, a Pro-Palestine, and a Neutral user. For each of these personas, 100 independent scraper agents were deployed. Each was executed in single-persona mode, ensuring stable behavioral profiles throughout the simulation. To guarantee comparability, all agents underwent the same initialization phase. During this phase, each persona viewed an identical set of introductory videos (five per ideological stance, for a total of fifteen) to establish a shared contextual baseline within YouTube's recommendation ecosystem. This controlled exposure ensured that any divergence in subsequent recommendations could be attributed to differences in persona-driven behavior rather than contextual bias.

Following initialization, each persona engaged in a fifty-step navigation sequence, selecting videos according to its defined behavioral and ideological preferences (94%, 93%, and 65% stance-coherent choices for the pro-Israel, pro-Palestine, and neutral personas, respectively). At each step, TRACE collected all visible recommendations and their corresponding ranks, alongside the video ultimately chosen by the agent and the language model justification for that choice. This meticulous logging enabled the computation of a rank-aware lean metric, defined as the rank-weighted expected stance of the recommendation list (with Pro-Israel = −1, Neutral/Unrelated = 0, and Pro-Palestine = +1), thereby quantifying the ideological orientation of the recommendations while reflecting the relative prominence of higher-ranked content within the user interface.

Figure 3 illustrates the evolution of the mean lean score over time. While the figure suggests an initial Pro-Palestinian bias and a gradual polarization toward each persona's initial stance, the focus here is methodological: TRACE makes it possible to detect, quantify, and reproduce such ideological dynamics across multiple independent runs, with complete control over contextual and behavioral variables. Each trajectory results from an autonomous, persona-driven browsing process executed under identical experimental settings, providing a transparent and auditable foundation for analyzing algorithmic sensitivity to user identity and initial recommendation bias.

5 Conclusion

TRACE offers a scalable and extensible framework that redefines how researchers can audit YouTube's recommendation algorithm. Integrating Large Language Models into persona-driven simulations bridges the methodological gap between behavioral realism, control, and reproducibility. Its modular, containerized architecture and centralized database ensure full traceability and large-scale parallelization, while the intuitive web interface allows researchers to configure, launch, and monitor experiments in real time. Through its proof-of-concept, TRACE demonstrates how ideological trajectories can be transparently generated, logged, and analyzed under identical conditions, providing valuable insight into the interaction between user behavior and algorithmic personalization.

Funding

This work was supported by an FSR Seedfund grant from UCLouvain.

Data availability statement

The TRACE software, including all installation instructions, configuration files, and LLM prompts used in the experiments, is available here: https://github.com/Hyedryn/TRACE_public.

References

[1] Megan A. Brown, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2022. Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users. doi:10.2139/ssrn.4114905

[2] Sarmad Chandio, Muhammad Daniyal Pirwani Dar, and Rishab Nithyanand. 2024. How Audit Methods Impact Our Understanding of YouTube's Recommendation Systems. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 241–253. doi:10.1609/icwsm.v18i1.31311

[3] Luyu Chen, Quanyu Dai, Zeyu Zhang, Xueyang Feng, Mingyu Zhang, Pengcheng Tang, Xu Chen, Yue Zhu, and Zhenhua Dong. 2025. RecUserSim: A Realistic and Diverse User Simulator for Evaluating Conversational Recommender Systems. In *Companion Proceedings of the ACM on Web Conference 2025 (WWW '25)*. Association for Computing Machinery, New York, NY, USA, 133–142. doi:10.1145/3701716.3715258

[4] Hussam Habib and Rishab Nithyanand. 2025. YouTube Recommendations Reinforce Negative Emotions: Auditing Algorithmic Bias with Emotionally-Agentic Sock Puppets. doi:10.48550/arXiv.2501.15048 arXiv:2501.15048 [cs].

[5] Shengchun Huang and Tian Yang. 2024. Auditing Entertainment Traps on YouTube: How Do Recommendation Algorithms Pull Users Away from News. *Political Communication* 41, 6 (Nov. 2024), 903–920. doi:10.1080/10584609.2024.2343769

[6] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (2020), 1–27.

[7] Prerna Juneja, Md Momen Bhuiyan, and Tanushree Mitra. 2023. Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3544548.3580846

[8] Mohammed Lahsaini, Mohamed Lechiakh, and Alexandre Maurer. 2024. Auditing health-related recommendations in social media: A Case Study of Abortion on YouTube. doi:10.48550/arXiv.2404.07896 arXiv:2404.07896 [cs].

[9] Hieu Le, Salma Elmalaki, Zubair Shafiq, and Athina Markopoulou. 2025. AutoLike: Auditing Social Media Recommendations through User Interactions. doi:10.48550/arXiv.2502.08933 arXiv:2502.08933 [cs].

[10] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 522–533. doi:10.1609/icwsm.v14i1.7320

[11] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 131–141.

[12] Lu Tang, Kayo Fujimoto, Muhammad Tuan Amith, Rachel Cunningham, Rebecca A. Costantini, Felicia York, Grace Xiong, Julie A. Boom, and Cui Tao. 2021. "Down the Rabbit Hole" of Vaccine Misinformation on YouTube: Network Exposure Study. *Journal of Medical Internet Research* 23, 1 (Jan. 2021), e23262. doi:10.2196/23262

[13] 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*. 1–11.

[14] An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On Generative Agents in Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1807–1817. doi:10.1145/3626772.3657844