IAS Group 8 Project Proposal

Exploring Bias, Radicalization, and Human Choice in AI Recommender Systems

1. Project Description

This project investigates how AI recommender systems, such as those used by YouTube, can influence users toward extremist or polarizing content.

The study focuses on the interaction between algorithmic recommendations and human choice, examining how biases, personalization, and feedback loops affect the trajectory of content consumption.

By combining literature research, algorithmic simulations, and interactive demonstrations, the project aims to explore the societal implications of AI-driven content recommendation.

The primary objective is to illustrate that **recommender systems are not inherently neutral**.

By simulating user interactions and analyzing algorithmic outputs, the project will expose how human biases and system design together shape information consumption, potentially leading to radicalization or ideological reinforcement.

This approach is grounded in empirical research that identifies both algorithmic and human factors contributing to exposure to extreme content:

Research Findings to Inform the Project

1. **Amplification of Extremist Content:** Studies show that platforms like YouTube can amplify extremist content through their recommendation algorithms. Users may be nudged down a “rabbit hole” of increasingly extreme content, particularly among right-leaning audiences ([1] UC Davis, 2021).
2. **Role of User Intent:** Individual choices play a key role in content exposure. Users with higher levels of predisposition toward certain beliefs are more likely to engage with extreme content, highlighting the interplay between algorithmic recommendations and human behavior ([4] GNET Research, 2022).
3. **Counterfactual Studies:** Research using counterfactual bots demonstrates that algorithmic recommendations, on average, may moderate rather than exacerbate partisan consumption, indicating that the system’s effect is nuanced ([2] arXiv, 2023).
4. **Debiasing Interventions:** Interventions aimed at reducing bias in recommendation algorithms have shown promise, promoting more diverse and ideologically neutral content, though challenges remain for some user groups ([3] arXiv, 2022).

These findings emphasize that **recommendation systems can influence societal behaviors but that individual user choices and algorithm design both significantly shape outcomes.**

2. Team Members

A: Francisco José Gomes da Silva

B: José Guilherme Lourenço Correia Marques dos Santos

C: Victor Daniel Tomás Rodriguez

Team Leader

A

Roles of Each Team Member

A: Overall coordination, methodology design, implementation of SVD-based simulations.

B and C: Development of interactive Streamlit demos, user interface design, implementation of click interactions. Data preparation, literature review, documentation, and bibliography compilation.

3. Resources

Data: Synthetic dataset of videos or Movielens 100K dataset for SVD experiments; videos categorized into neutral, mildly political, and extreme for controlled simulations.

Code: Python, Streamlit, NumPy, Pandas, SciPy, Matplotlib.

Examples: YouTube content for discussion (no copyrighted content included in the demo).

Open-source Tools: Python libraries, Streamlit, GitHub repositories such as Failed Machine Learning and Awful AI.

Other Tools: Jupyter Notebook, VS Code, PDF generation tools.

Rationale for Dataset Choice

Although using a YouTube dataset might seem ideal, full user interaction data is not publicly accessible due to privacy restrictions.

Moreover, such datasets are extremely large and require extensive preprocessing.

Using a controlled synthetic dataset or Movielens allows the team to simulate different user personalities and interactions, making it easier to demonstrate the effects of algorithmic bias and feedback loops in an educational and interpretable setting.

4. Bibliography (Selected References)

[1] M. Haroon, M. Wojcieszak, A. Chhabra, X. Liu, P. Mohapatra, and Z. Shafiq, "Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations," *Proc. Natl. Acad. Sci. U.S.A.*, edited by D. Massey, Princeton University, Princeton, NJ; received Aug. 1, 2022; accepted Sep. 21, 2023.

[2] H. Hosseinmardi, A. Ghasemian, M. Rivera-Lanas, M. H. Ribeiro, R. West, and D. J. Watts, "Causally estimating the effect of YouTube's recommender system using counterfactual bots," *arXiv preprint arXiv:2308.10398v2 [cs.SI]*, Dec. 2023.

[3] M. Haroon, A. Chhabra, X. Liu, P. Mohapatra, Z. Shafiq, and M. Wojcieszak, “YouTube, The Great Radicalizer? Auditing and mitigating ideological biases in YouTube recommendations,” *arXiv preprint arXiv:2203.10666v2 [cs.CY]*, Mar. 2022.

[4] J. Whittaker, *Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence*, Global Internet Forum to Counter Terrorism (GIFCT) Transparency Working Group, 2022. Available: https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-TR-Empirical-1.1.pdf

5. Proposed Methodology

1. Develop a **synthetic dataset of videos or movies** classified into categories (neutral, mildly political, extreme) to simulate user interactions.
2. Implement **SVD-based recommender simulations** to track latent space drift and propagation of user preferences over time.
3. Design an **interactive Streamlit demo**, allowing participants to make click selections and observe their effect on recommendations and category distributions.
4. Conduct a **literature review** of real-world cases where recommender systems have influenced radicalization or polarization.
5. Synthesize findings to create educational materials highlighting the societal influence of algorithmic recommendation systems.

6. Potential Practical Activities

- **Interactive Simulation:** Participants adjust user biases and make click choices, observing algorithmic outcomes in real time.
- **Workshop/Seminar:** Students explore different user personas and feedback loops, analyzing the effects on content diversity. Discussion of case studies on algorithmic radicalization, mitigation strategies, and ethical considerations.
- **Hands-On Exercise:** Compare simulated algorithmic clicks versus human-chosen interactions to highlight the role of personal agency.

7. Expected Outcomes

This project will provide an educational framework demonstrating how AI recommender systems can amplify biases and influence content consumption. It aims to foster understanding of algorithmic transparency, critical thinking about AI in society, and discussions on ethical AI design and governance.