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

**Artificial Intelligence and Society**

Faculdade de Ciências & Faculdade de Engenharia, Universidade do Porto

José Guilherme Lourenço Correia Marques dos Santos
Francisco José Gomes da Silva
Mateus Maria Gomes Eça de Queiroz Cabral
Victor Daniel Tomás Rodriguez

December 1, 2025

**Abstract.** Recommender systems quietly shape what we watch, read, and listen to. On platforms like YouTube, they decide which video auto-plays next and which topics follow us across sessions. This blog-style article explores how these systems, particularly YouTube's recommendation algorithm, can amplify political bias and even nudge users toward more extreme content. Combining a matrix-factorization-based simulation with recent empirical audits and counterfactual studies, the article argues that recommendation algorithms are not neutral tools but design artefacts that reflect business incentives, data biases, and societal power structures. It concludes by discussing how regulation, technical design choices, and informed user behaviour can work together to steer these systems toward more democratic outcomes.

# 1 Why Recommender Systems Matter Today

## 1.1 A Polarized World Curated by Algorithms

Artificial intelligence recommender systems have fundamentally transformed how individuals discover and consume information. YouTube, the most popular video platform globally, is a prime example. With around 81% of the U.S. population using the platform and about 70% of watch time driven by algorithmic recommendations, YouTube's recommender has become one of the most influential information gatekeepers in contemporary society [3][4].

This power operates in a context of deep political polarization. The gap between left and right on key issues has widened, hostility between partisan groups has increased, and support for political violence is not negligible [3][4]. Social media and recommendation algorithms are regularly accused of locking users into "filter bubbles" and leading them down "rabbit holes" of increasingly ideological or conspiratorial content [3]. YouTube, in particular, has been labelled "the great radicalizer", a platform whose algorithmic logic allegedly pushes users from mainstream content toward increasingly extreme material [3][4].

## 1.2 The Core Question: Are Algorithms Driving Radicalization?

At the heart of the debate lies a deceptively simple question: do recommendation algorithms systematically direct users toward extreme, conspiratorial, or otherwise problematic content? [4] Answering this requires unpacking what we call the **loop effect**, a feedback cycle with four interacting components:

1) **Selective exposure**: users tend to consume information that matches their pre-existing views [3].

2) **Homophily**: social networks, online and offline, are biased toward like-minded connections [3].

3) **Filter bubbles**: personalization algorithms learn from this biased behaviour and recommend more of the same [3].

4) **Feedback**: as users follow these recommendations, their future options and sometimes their beliefs become more homogeneous [3].

Empirically, the picture is complex. Large-scale audits report that YouTube's algorithm favours ideologically congenial content and that this effect is especially strong for right-leaning users [3][4]. More recent counterfactual studies, however, suggest that after major algorithmic changes in 2019, user preferences may play a larger role than algorithmic bias, with recommendations sometimes moderating rather than radicalizing viewing behaviour [6]. Reality lies somewhere in between: algorithm and user co-produce the information environment.

### 1.3 Key Stakeholders in This Ecosystem

Several actors shape and are shaped by these algorithms:

**YouTube / Google.** The platform optimizes recommendations primarily for engagement and advertising revenue. Former employees have described how problematic content was promoted because it held user attention, and how attempts to reduce such promotion sometimes clashed with business interests [3].

**Users.** Most YouTube users consume relatively mainstream content. Yet a small, highly engaged subset repeatedly encounters and sometimes seeks out ideologically extreme or conspiratorial material [4][6]. This group is disproportionally important for both political impact and algorithmic training data.

**Creators.** From major news organizations to fringe influencers, content creators respond to the platform's incentives. Provocative, emotionally charged, or divisive content often performs better, reinforcing the supply of such material.

**Regulators.** The European Union's Digital Services Act (DSA), in force since 2024, requires large platforms to explain in plain language how their recommender systems work and to offer at least one non-personalized recommendation option [74]. This is one of the first serious attempts to regulate the logic of recommendation itself.

**Researchers.** Teams at UC Davis, Yale, EPFL, the University of Pennsylvania and elsewhere have developed both sock-puppet audits and counterfactual bot experiments to isolate algorithmic influence from user choice [4][6]. Their findings underpin much of the discussion that follows.

## 2 How Modern Recommender Systems Work

### 2.1 From Click Logs to Latent Factors

At their core, recommender systems predict which items a user is likely to appreciate based on historical interaction patterns. On YouTube, the items are videos; on Netflix, they are movies or series; on Spotify, songs and playlists. A central technical paradigm for this prediction task is **matrix factorization**, which starts from a large user–item interaction matrix $R$, where each entry $r_{ui}$ might represent a rating, a view, or another engagement signal.

Matrix factorization assumes that both users and items can be embedded into a lower-dimensional latent space. Formally, we approximate:

$$R \approx UV^T,$$

where $U$ is a matrix of user latent factors and $V$ is a matrix of item latent factors [7]. The dot product between a user's and an item's latent vectors approximates the strength of their interaction.

On platforms like YouTube this idea is instantiated across several *surfaces*: homepage recommendations (what users see when they open the site), up-next recommendations (what auto-plays after the current video) and search-based suggestions. Each surface can use slightly different models or weighting strategies [3][4].

## 2.2 Design Choices That Create Feedback Loops

YouTube's recommender reflects a set of optimisation choices:

- **Objective function:** engagement metrics such as watch time, click-through rate and session length are primary targets [3].
- **Personalization:** models are trained on each user's history and the behaviour of "similar" users [3].
- **Heterogeneous surfaces:** homepage and up-next recommendations are tuned differently, with homepage often being more strongly personalized [3][4].

Combined, these choices create the loop effect. If a user repeatedly watches content from a particular political leaning, the model updates that user's latent vector towards that region of the space and increases the score of similarly positioned videos. Over time, the user's local neighbourhood in latent space can become ideologically homogeneous, even if the global catalogue is diverse.

## 2.3 Our Simulation: Matrix Factorization in a Political Toy World

To make these dynamics concrete, our project implements a recommender simulation based on the MovieLens 100K dataset. While MovieLens is about movies, not politics, it provides a clean, well-understood environment to prototype algorithms.

**Dataset and Political Mapping.** The original dataset contains 100,000 interactions over 31 attributes. For experimentation, we select a subset of 50 users and 100 movies (including a focal user with ID 196). We then map 19 movie genres into three political categories:

- **Neutral** (10 genres): Comedy, Animation, Children's, Musical, Drama, Romance, Sci-Fi, Fantasy, Documentary, Unknown.
- **Mildly political** (5 genres): Action, Adventure, Thriller, Crime, Mystery.
- **Extreme** (3 genres): War, Film-Noir, Horror.

This mapping is obviously stylized but serves as a proxy to observe how an algorithm shifts a user's recommendations along a neutral–extreme axis.

**Simulation Parameters.** We train models with:

- 50 users and 100 movies,
- latent dimension $k = 10$,
- 5 simulation rounds for observing drift,
- learning rate 0.30.

### 2.4 SVD: Decomposing the Ratings Matrix

Singular Value Decomposition (SVD) factorizes the rating matrix $R$ into three matrices:

$$R = U\Sigma V^T,$$

where $U$ and $V$ are orthogonal matrices and $\Sigma$ is diagonal with non-negative singular values. For recommendation, we keep only the top $k$ components:

$$\hat{R} \approx U_k \Sigma_k V_k^T.$$

In our configuration:

- $U_k \in \mathbb{R}^{50\times 10}$: latent factors for 50 users.
- $V_k^T \in \mathbb{R}^{10\times 99}$: latent factors for 99 items.

A bias-aware version predicts ratings as [7]:

$$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T p_u,$$

where $\mu$ is the global mean, $b_i$ and $b_u$ are item and user bias terms, and $p_u, q_i$ are $k$-dimensional latent vectors.

In our experiment:

- the top 10 singular values explain 75.37% of total variance,
- MSE $= 1.0076$,
- RMSE $= 1.0038$.

### 2.5 PMF: A Probabilistic View of Preferences

Probabilistic Matrix Factorization (PMF) interprets ratings as noisy observations generated from latent factors. For observed pairs $(u, i) \in \mathcal{K}$, it assumes: [6]

$$p(R \mid P, Q, \sigma^2) = \prod_{(u,i)\in\mathcal{K}} \mathcal{N}(r_{ui} \mid p_u^T q_i, \sigma^2),$$

with Gaussian priors:

$$p(P \mid \sigma_P^2) = \prod_u \mathcal{N}(p_u \mid 0, \sigma_P^2 I), \qquad p(Q \mid \sigma_Q^2) = \prod_i \mathcal{N}(q_i \mid 0, \sigma_Q^2 I).$$

Learning maximizes the log-posterior over $P$ and $Q$. In our setting PMF achieves:

- MSE $= 0.4024$,
- RMSE $= 0.6343$,

substantially improving over plain SVD and capturing uncertainty more effectively.

### 2.6 ALS: Alternating Least Squares at Scale

Alternating Least Squares (ALS) optimizes a regularized squared-error objective:

$$\min_{P,Q} \sum_{(u,i)\in\mathcal{K}} (r_{ui} - p_u^T q_i)^2 + \lambda \left( \|p_u\|^2 + \|q_i\|^2 \right).$$

By fixing $Q$ and solving for $P$ and then fixing $P$ and solving for $Q$, ALS decomposes the problem into a sequence of convex sub-problems. This makes it attractive for industrial-scale systems with sparse, implicit feedback, which is close to YouTube's actual setting.

### 2.7 An Interactive "Radicalization Index"

To make these ideas tangible, we built a Streamlit-based interactive demo:

1) It visualizes users and items in a reduced 3D latent space, showing how similar items cluster [2].
2) The user can "click" on neutral, mildly political, or extreme items and see how the model updates their profile.
3) A *radicalization index* tracks the share of extreme items in the current top-10 recommendations.

4) The interface allows switching between SVD, PMF and ALS to compare how each algorithm responds to the same interactions.

5) Users can attempt to “recover” by deliberately watching neutral content and observing whether recommendations rebalance.

Even in this simplified world, repeated selection of “extreme” items gradually increases the radicalization index. Conversely, neutral interactions can pull it back down, but typically more slowly. This mirrors patterns observed in real YouTube audits [2][4].

### 2.8 Connecting the Toy World to Real YouTube Data

Large-scale empirical work on YouTube complements these simulations:

**UC Davis sock-puppet audit.** Researchers trained 100,000 automated browser “sock puppets” on different ideological diets and recorded 9,930,110 watched videos from 120,073 channels [4]. Key results include:

- very-right sock puppets received congenial top-1 homepage recommendations 65.5% of the time, compared to 58.0% for very-left puppets;
- along up-next trails, very-right users experienced a 37% increase in congenial recommendations with depth, while very-left users did not;
- only about 2.5% of recommendations pointed to “problematic” channels (e.g. Alt-right, Conspiracy, QAnon), but 36.1% of users encountered at least one such channel, rising to 40% for very-right profiles.

**Counterfactual bots.** Hosseinmardi et al. trained programmatic bots to follow either real user histories or simple rule-based recommendation-following paths [6]. Post-2019 results suggest that when bots simply follow YouTube’s sidebar or homepage recommendations, they often end up consuming *less* partisan content than the corresponding real users. In other words, the algorithm shows some moderating tendencies relative to users’ own preferences.

Together, these studies and our simulation support a nuanced view: recommender systems can amplify bias and occasionally lead users toward problematic content, but user intent and broader social forces remain central.

## 3 What We Learn from Outcomes and Implications

### 3.1 The Upside: Navigation, Personalization, Scale

It is important to acknowledge that recommender systems deliver genuine value.

**Navigation.** They help users navigate enormous catalogues. Without some form of algorithmic triage, platforms like YouTube would be nearly unusable.

**Personalization.** Good recommendations surface content that users might never have found manually. For many, this is experienced as serendipity rather than manipulation.

**Scalability.** Matrix factorization and related methods allow platforms to operate at web scale, computing recommendations for millions of users and items in near real-time [7].

### 3.2 The Downside: Filter Bubbles and Radicalization Pathways

Yet these very mechanisms also have serious downsides.

**Filter bubbles for politically engaged users.** While most users see a relatively mixed information diet, politically active partisans are more likely to inhabit echo chambers, both in their social networks and their recommender-driven feeds [3][4]. This small group is often more vocal, more politically engaged and more influential than average citizens.

**Algorithmic radicalization pathways.** Empirical evidence shows that some users move from centrist to more extreme channels over time [4][6]. Our simulation shows the mechanism: when the model updates user factors after every extreme click, the latent profile shifts in that direction, and so do future recommendations.

**Right-leaning asymmetry.** Multiple audits find stronger amplification of congenial and problematic content for right-leaning users than for left-leaning ones [3][4]. This asymmetry is not yet fully understood, but may relate to:

- different content supply patterns across the ideological spectrum,
- the higher engagement provoked by certain right-wing narratives,
- incomplete or uneven enforcement of platform policies.

**Normalisation through exposure.** Even if 97.5% of recommendations do not point to problematic channels, the fact that over a third of users eventually encounter such channels somewhere in their recommendation trails matters. Repeated incidental exposure can gradually shift what counts as “normal” or “acceptable”. [4]

### 3.3 Opacity and Accountability Gaps

From the outside, YouTube’s recommender is effectively a black box. Researchers can log what is recommended and what gets watched, but not directly see the internal model structure or objective functions [5].

This opacity has several consequences:

- platforms can claim that certain policy changes (e.g. reducing “borderline” content) had large positive effects without independent verification [65];
- users cannot meaningfully understand or contest why specific items are recommended;
- regulators struggle to assess systemic risks or enforce standards;
- causal responsibility for harms becomes hard to attribute.

### 3.4 A Supply-and-Demand Perspective

A helpful way to cut through the “algorithm vs. users” dichotomy is to adopt a supply-and-demand lens [4][5]:

- **Supply:** extremist and conspiratorial creators produce content that is engaging to certain audiences, sometimes financially rewarded through monetization.
- **Demand:** users with pre-existing grievances, prejudices or curiosity seek out or dwell on such content.
- **Matching:** recommender systems optimised for engagement become increasingly efficient at connecting this supply and demand.

In this view, the algorithm amplifies and structures existing dynamics rather than creating them ex nihilo. Effective interventions therefore need to address supply (e.g. moderation, demonetization), demand (media literacy, social support) and the matching mechanism (recommender design and constraints).

## 4 Why This Matters for Society and Governance

### 4.1 Media Narratives and Platform Responses

High-profile journalism has played a major role in framing YouTube as a radicalizing force. Stories like “The Making of a YouTube Radical” in the *New York Times* and essays such as Tufekci’s “YouTube, The Great Radicalizer” cast recommender systems as almost autonomous agents driving individuals toward extremism [3][67].

In response to mounting criticism, YouTube announced a “Four Rs” policy framework:

- **Remove** content that clearly violates policies.
- **Raise** up authoritative sources in recommendations.
- **Reward** trusted, policy-compliant creators.

- **Reduce** the spread of “borderline” content videos that do not quite violate rules but are considered harmful [65].

YouTube claims these measures significantly reduced views of borderline content, but independent verification is limited due to black-box constraints.

### 4.2 Regulation: The Digital Services Act and Beyond

The EU’s Digital Services Act directly targets recommender systems:

- platforms must describe the main parameters of their recommender systems in “plain and intelligible” language;
- very large online platforms must provide at least one non-profile-based recommendation option (e.g. chronological ordering);
- regulators receive stronger auditing and data access powers [66][74].

The EU’s AI Act, while not ultimately categorising recommenders as high-risk systems, further signals regulatory concern about opaque, large-scale algorithmic systems [77].

### 4.3 Challenging the Myth of Algorithmic Neutrality

The idea that algorithms are neutral, objective tools is increasingly untenable. Design decisions embed human values at multiple levels:

- **Objective choice:** engagement optimisation vs. accuracy, diversity or civic value.
- **Preference modelling:** users are treated as isolated individuals rather than socially embedded agents whose preferences evolve.
- **Content ranking:** default favouring of popular or trending items can systematically marginalise minority viewpoints.
- **Training data:** historical patterns of inequality and discrimination are reproduced in model behaviour [8].

These are not mere implementation details; they are normative choices with political and ethical consequences.

### 4.4 Agency, Determinism, and Time

Counterfactual bot studies remind us that users retain significant agency [6]. Many radicalization trajectories involve users actively searching for, subscribing to, or repeatedly selecting extreme content. The algorithm is complicit but not omnipotent.

At the same time, the effects of recommenders change over time. Pre-2019 YouTube likely behaved differently from the post-2019 version. This means that:

- empirical results must always be interpreted in temporal context;
- platform changes can and do alter outcomes, which is evidence that different designs are possible;
- long-term research and monitoring are essential to track the impact of both platform self-regulation and formal regulation.

### 4.5 Fairness and Asymmetry

The persistent asymmetry whereby right-leaning users receive stronger amplification of congenial and problematic content raises fairness concerns [3][4]. At minimum, it suggests that the recommender does not treat all political orientations equivalently.

Possible contributors include:

- higher density of right-wing alternative media on YouTube,
- different engagement patterns across audiences,
- uneven moderation or enforcement practices.

Regardless of exact causes, such asymmetries challenge claims of neutrality and highlight the need for fairness-aware recommender research [8].

## 5 Takeaways: Steering Recommender Systems Toward Better Futures

Putting all the pieces together, recommender systems on platforms like YouTube are best understood as socio-technical infrastructures: they are neither neutral tools nor all-powerful puppet masters. Instead, they are powerful amplifiers shaped by business models, design decisions, data flows and user behaviour.

Key takeaways include:

1) **Algorithmic bias is real but not the whole story.** Recommenders can and do amplify ideological bias and occasionally promote problematic content, yet user preferences and broader social dynamics remain crucial.
2) **Design choices matter.** Objective functions, model architectures, and ranking heuristics embed values. Post-2019 algorithm changes on YouTube demonstrate that different trade-offs can be made.

3) **Transparency and governance are essential.** Regulatory frameworks like the DSA, combined with independent auditing, are needed to make recommender systems accountable rather than opaque.

4) **Multiple levers must be pulled.** Effective responses must combine supply-side moderation, demand-side education, design changes (e.g. diversity-promoting objectives) and user empowerment tools.

5) **Education helps.** Simulations like our matrix-factorization-based "radicalization index" interface allow students, policymakers and the wider public to see how seemingly abstract algorithms can produce concrete social patterns.

Recommender systems will remain central to digital life. The question is not whether we use them, but whose values they encode, whose interests they serve, and how we collectively govern them.

# References

[1] Haroon, M., Wojcieszak, M., Chhabra, A., Liu, X., Mohapatra, P., & Shafiq, Z. (2023). Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, 120(50).

[2] CS8 Project Proposal (2025). *Exploring Bias, Radicalization, and Human Choice in AI Recommender Systems.*

[3] Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z., & Wojcieszak, M. (2022). YouTube, The Great Radicalizer? Auditing and mitigating ideological biases in YouTube recommendations. arXiv:2203.10666.

[4] Whittaker, J. (2022). Recommendation Algorithms and Extremist Content: A Review of Empirical Evidence. GIFCT Transparency Working Group.

[5] Hosseinmardi, H., Ghasemian, A., Rivera-Lanas, M., Ribeiro, M. H., West, R., & Watts, D. J. (2023). Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proceedings of the National Academy of Sciences.*

[6] *Matrix Factorization Techniques for Recommender Systems.* (2021). Foundational techniques for collaborative filtering.

[7] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8), 30–37.

[8] Raza, S., et al. (2024). A Comprehensive Review of Recommender Systems. arXiv.