Political Alignment in Recommendations: A Public Good Dilemma

Philipp Chapkovski Dietmar Jannach Silvia Milano Caterina Giannetti
Cecilia Vergari Nicola Meccheri Marco Catola

University of Duisburg-Essen

University of Klagenfurt

TUM Ethical Digital Initiative

University of Pisa

25th–27th March 2026
Kreuzlingen (CH)

Political alignment in recommender systems

- **Recommender systems** shape what users see and do online.
- Political alignment asks what content mix or viewpoint balance should count as **appropriate**.
- It is therefore not only a question of **accuracy**, but also of **whose political values** the system reflects.
- Political alignment is a **normative choice**, not just a technical one.

Political alignment as a public good

- Steering a recommender toward a preferred political balance requires **costly user input** (ratings, feedback, value-relevant information).
- Once the system shifts, the benefits are partly **shared across users**.
- This creates incentives to **free-ride** on others' effort.
- Political alignment can therefore be **underprovided**, even when many users want more of it.

Contribution & Research Questions

Objective and contribution

- **Objective**
  Understand when users are willing to make **costly contributions** to improve a **shared recommender** under **value alignment or misalignment**.

Objective and contribution

- **Objective**
  Understand when users are willing to make **costly contributions** to improve a **shared recommender** under **value alignment or misalignment**.

- **Contribution**
  - Reframe political alignment as a **public-good dilemma**.
  - Use a **light economic model** to derive hypotheses.
  - **Test them experimentally** in an interdisciplinary design.

Research questions

- Does **perceived political alignment** affect users' willingness to make **costly contributions** to a shared recommender system?
- Is free-riding stronger when users are paired with **aligned partners** or with **misaligned partners**?
- Is this effect stronger in **political** than in **non-political** domains?

Theoretical framework

Model intuition

- The closer the system is to a user's **ideal political level**, the higher the user's **click probability**.
- Improving alignment requires **costly contributions**.
- Since these contributions improve a **common system**, their benefits are partly **shared**.

Click probability Platform profit

A simple model

- Two user groups choose contributions w_1 and w_2.
- A platform then selects the recommender's **political level** E.
- Higher contributions induce the platform to choose a higher E.
- Each group benefits when the system is closer to its **ideal political level**, but contributing is **costly**.
- Political alignment is therefore a **collective-action problem**.

Equilibrium intuition

$$\frac{x_2 - w_2}{x_1 - w_1} = \frac{\varepsilon_{p_1}}{\varepsilon_{p_2}}$$

- In equilibrium, contribution incentives must **balance across user types**.
- Users who are more sensitive to **political misalignment** must be willing to give up more surplus, $x_i - w_i$, to shift the common system.
- Because both users influence the **same platform choice**, each internalizes only part of the benefit from contributing.

Main behavioral implication

- Because both benefit from the same improvement, contributions are **strategic substitutes**.
- If a user expects a **similar / aligned partner** to contribute more, the best response is to contribute less.
- This yields the core prediction: **homogeneous matches** may display more free-riding than **heterogeneous matches**.

Experimental Design

Experimental design: overview

- *Three phases*: Fig
  1. **Pre-treatment survey + preference elicitation** (**ground truth**)
  2. **Repeated contribution game**(10 rounds; stranger matching)
  3. **Prediction & payment stage** (transparent mapping contributions $\rightarrow$ accuracy $\rightarrow$ payoff)
- **Treatments**: (*similar vs opposed*) $\times$ (*political vs non-political domain*).

Phase 1: Pre-treatment survey & ground truth

- **To define matching labels (similar vs opposed)**, survey elicits:
  - **political positions** (e.g. immigration, abortion, gun control),
  - **non-political preferences** (e.g. dogs vs cats, print vs eBooks).
- **Ground truth**: each participant ranks 5 **movies**
  - selection from a common pool (e.g. IMDb)
  - possibility to drag from most to least preferred. Figure
- **At the end of the experiment**: Rankings used to determine **recommendation accuracy** and payments.

Phase 2: Rating rounds

- **10 rounds** with **stranger matching**.
- In each round, participants see:
  - one movie from a **balanced pool of 10 pre-tested titles**,
  - the **cost** of rating, Movie neutral Movie political + rating
  - whether the partner is **Similar** or **Opposed**. Message
- **No identities, no chat, no partner history.**
- The **learning rule is fixed**; only the **context** changes across treatments.

Phase 2: The contribution decision

- Each participant starts with an endowment of **€2.50**.
- In each round, the choice is binary:
  - **Rate**: pay **€0.25** to submit a rating
  - **Skip**: keep the money and do not rate
- A rating is **privately costly**, but improves the recommender for **both** users.
- This creates a **shared-learning dilemma**: one can benefit from the partner's effort without contributing.

Phase 2: Movie stimuli

- To avoid familiarity and strong priors, we use **fictional movies** rather than existing titles.
- A **separate subject pool** is used to pre-test and select **10 movies** from an initial pool of **16**.
- The final set is balanced across content:
  - **5 political** movies (2 left, 2 right, 1 neutral)
  - **5 non-political** movies
- We also balance the stimuli across **genres**.
- Each movie is presented with a **poster** and a short **synopsis**; all treatments use the same validated pool.

Phase 3: Individual payoff

$$\Pi_i = e - c\, r_i + b\, K_i$$

- $e = €2.50$: fixed **endowment**.
- Each rating lowers payoff by $c = €0.25$.
- The final bonus depends on $K_i \in \{0, 1, 2, 3, 4, 5\}$, the number of **correctly guessed positions**.
- Each correct position pays $b = €2$.

This slide defines the payoff components; the next slide shows how **total ratings** improve K_i.

Phase 3: From ratings to accuracy

$$\lambda = 1 + a(r_i + r_j), \qquad a = 0.10$$

- λ is the recommender's **learning / accuracy parameter**.
- It depends on **total pair ratings**, $r_i + r_j$, not only on my own effort.
- Each additional rating raises expected accuracy by about **0.10 positions**.
- No ratings $\Rightarrow \lambda = 1$: about **1 correct position** on average.
- Full mutual rating ($r_i = r_j = 10$) $\Rightarrow \lambda = 3$: about **3 correct positions** on average.

More total ratings $\Rightarrow$ better predictions $\Rightarrow$ higher expected bonus for **both** participants.

Phase 3: Incentive structure

$$a = 0.10, \qquad b = €2, \qquad c = €0.25$$

- One rating costs the contributor **€0.25**.
- It increases expected bonus by approximately

  $$ba = 2 \times 0.10 = €0.20$$

  for **each** participant.
- So the **private return** is below the private cost:

  $$€0.20 < €0.25$$
- But the **joint return** exceeds the cost:

  $$2ba = €0.40 > €0.25$$

Hence the structure is **individually costly but jointly beneficial**: a public good /

Treatment and Hypotheses

Treatments

- **22 design**: domain of disagreement $\times$ match type
- **Domain**: Political vs. Non-political
- **Match type**: Homogeneous (Similar) vs. Heterogeneous (Opposed)
- Assignment based on survey responses

	Homogeneous (Similar)	**Heterogeneous (Opposed)**
Non-political	Neutral–Homogeneous	Neutral–Heterogeneous
Political	Political–Homogeneous	Political–Heterogeneous

Hypotheses

- **H1 — Alignment fosters free-riding**
  Contributions are **lower** in **homogeneous / similar** matches than in **heterogeneous / opposed** matches.
- **H2 — Polarization amplifies the effect**
  The gap between **heterogeneous** and **homogeneous** matches is **larger in the political domain**.
- **H3 — Broader content effect**
  Contributions differ between **political** and **non-political** movie content.

Preliminary Results

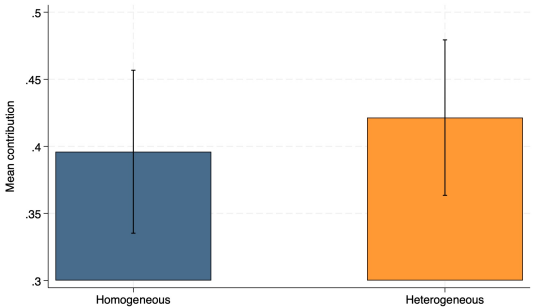
Pilot on Prolific: descriptive overview

- **Completed sample:** **53 participants** completed all 10 rounds.
- **Average contribution:** mean movies rated $\approx$ **4.09**, median $\approx$ **4.00**

Treatment	N	Mean rated	SD
Heterogeneous / Neutral	15	4.80	3.45
Homogeneous / Neutral	12	3.50	2.81
Heterogeneous / Political	13	3.54	1.61
Homogeneous / Political	13	4.38	1.89

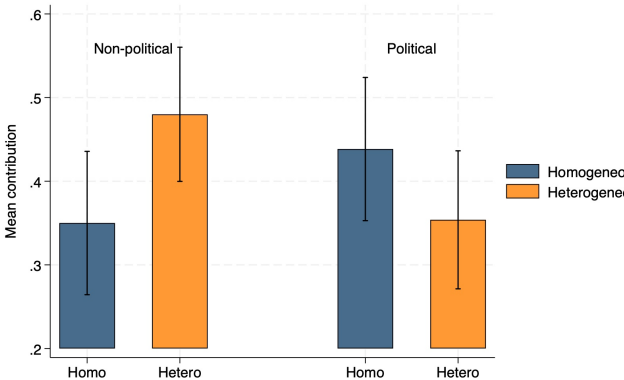
Results: Hypothesis 1

- **H1:** Contribution was slightly lower in homogeneous than heterogeneous matches (0.396 vs. 0.421), but the difference was **not statistically significant**.
- **Power:** Detecting a **meaningful increase** from 39% to 50% requires about $N = 640$ participants for 80% power, or roughly 160 per cell.



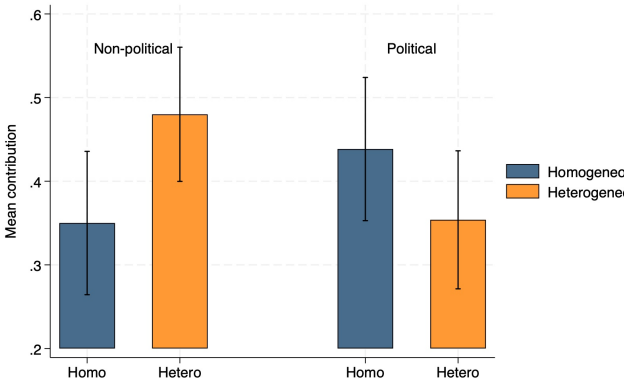
Results: Hypothesis 2

- **H2:** The heterogeneous effect was **positive in the non-political treatment** ($+0.13$, $p = 0.031$) but **negative in the political treatment** (-0.085, $p = 0.165$). $DiD = -0.215$, $p = 0.012$.
- **Power:** Simulation suggests that detecting an interaction of this size with 80% power requires roughly 160 participants per cell (about $N = 640$).



Results: Hypothesis 2

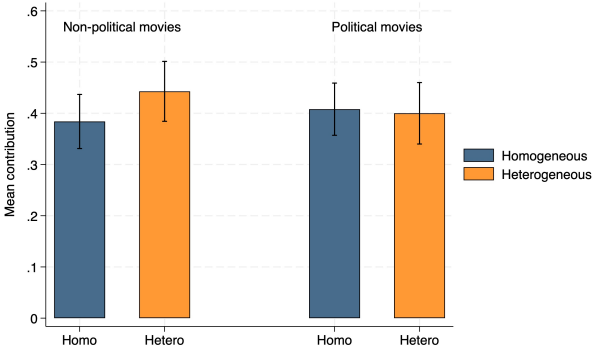
- **H2:** The heterogeneous effect was **positive in the non-political treatment** ($+0.13$, $p = 0.031$) but **negative in the political treatment** (-0.085, $p = 0.165$). $DiD = -0.215$, $p = 0.012$.
- **Power:** Simulation suggests that detecting an interaction of this size with 80% power requires roughly 160 participants per cell (about $N = 640$).



Interesting pilot pattern: the domain gap is significant, but opposite to the preregistered prediction and based on a small sample.

Results: Hypothesis 3

- **H3:** The political–non-political movie gap was small in both match types: $+0.024$ in homogeneous matches and -0.043 in heterogeneous matches, yielding a small and imprecise interaction ($DiD \approx -0.067$).
- **Power:** Simulated power with 150/200 participants per cell remains small, suggesting that the design is underpowered to detect an H3 interaction of this size.



Interpreting the pilot: mechanism or priming?

$$\frac{x_2 - w_2}{x_1 - w_1} = \frac{\varepsilon_{p_1}}{\varepsilon_{p_2}}$$

Theoretical mechanism: if I expect my partner to contribute more, I can contribute less.

Similar matches may therefore induce **more free-riding** if they are perceived as more informative or more likely to contribute.

Why this is informative: monetary incentives are held constant across treatments, so treatment differences reflect user-side beliefs and interpretation.

But the pilot does not yet isolate a purely strategic channel.

Interpreting the pilot: instruction channel

Important design note: the manipulation is not purely strategic.

Participants are also **primed by the instructions**: they are told they may act *alone vs together*, and that ratings from users with **similar tastes** are especially informative.

This means the treatment may affect:

- beliefs about partner contribution,
- beliefs about informativeness of one's rating,
- broader framing responses (identity, trust, demand effects).

Open issue: can we separate strategic response from framing and priming induced by the instructions?

Conclusions

- **H1:** The pattern is directionally consistent with the theory, but **imprecise**.
- **H2:** The domain difference is **interesting**, but opposite to the preregistered prediction.
- **H3:** The broader political–non-political content effect is **small and imprecise**.
- **Power:** A substantially larger sample is needed, especially for **interaction effects**.
- **Interpretation:** the pilot is informative, but the current treatment may mix strategic response with beliefs and framing.

Thank you

Thank you

Phase1: Rank 5 movies

Part 1: Your Private Movie Ranking

Open full instructions

Search and build your top 5 movies. Drag to reorder your preference list (rank 1 = strongest preference).

Add your top 5 movies before continuing.
You've added 0 movies. 5 more to add.

finding nemo

Search | Clear

Found 3 movies.

Preferred 5

Drag rows to reorder your ranking.

No movies selected yet.

Reset list

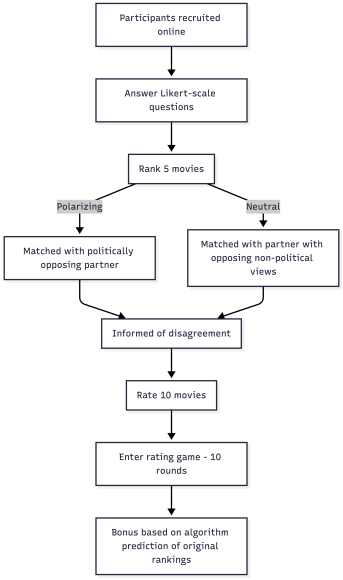
Search Results

Finding Nemo — Add
2003 • Director: Andrew Stanton
Cast: Albert Brooks, Ellen DeGeneres, Alexander Gould

Children's Aquarium: Finding the Real Nemo & Dory — Add
2015 • Director: Johannes Kernmeyer

Close

Game flow



Back

Matching Message

Important Matching Information

Remember: this movie will also be ranked by another participant. This participant **disagrees** with you at least on one of the following statements:

- Women should have the right to choose to have an abortion.
- Immigrants today are a burden on our country because they take our jobs, housing and healthcare.
- Climate change is an immediate crisis that demands radical changes in our lifestyles, even if it means sacrificing personal conveniences.

Back

Phase 2: Rate 10 movies (e.g. neutral)

Current budget: €2.50
Spent so far: €0.00

In the Shadows

MOVIE DESCRIPTION

A violent criminal strikes again, leaving detectives with one night to stop the next attack. With leads collapsing and witnesses too scared to talk, the hunt turns personal. As rain floods the streets and the clock runs out, one investigator follows a clue that could end it or get them killed.

How much would you like to see this movie realized? *

- Skip (no cost)
- Rate this movie (pay EUR 0.25)

Complete

Back

Phase 2: Rate 10 movies (e.g. political)

Current budget: €2.25
Spent so far: €0.25

Silicon Circus

MOVIE DESCRIPTION

Silicon Circus loves "global talent," but many workers are immigrants, and some are undocumented. They joke between meetings until panic hits. When an audit arrives, one employee risks everything to protect the people who built the company.

How much would you like to see this movie realized? *

- Skip (no cost)
- Rate this movie (pay EUR 0.25)

If you choose to rate, what is your rating? *

- 1 - Strongly dislike
- 2 - Dislike
- 3 - Neutral
- 4 - Like
- 5 - Strongly like

Complete

Back

User's engagement/Click probability

$$d_i = \left| E - E_i^{\text{ideal}} \right|, \qquad p_i(E) = f(E, d_i)$$

$$\frac{\partial p_i}{\partial E} \begin{cases} > 0 & \text{if } E < E_i^{\text{ideal}}, \\ < 0 & \text{if } E > E_i^{\text{ideal}}, \end{cases} \qquad \frac{\partial p_i}{\partial d_i} < 0$$

- Engagement is highest when $E = E_i^{\text{ideal}}$.
- Engagement decreases with distance from the user's ideal.

Back

Platform profit

$$\Pi_a(E; w_1, w_2) = \alpha\, p_1(E)\, w_1 + (1-\alpha)\, p_2(E)\, w_2 - C(E)$$

- The platform benefits from user engagement and contributions.
- Higher political alignment can increase engagement.
- But stronger political constraints are costly: $C(E)$.

Back

Phase 3: Payoff details

$$\Pi_i = e - c\, r_i + b\, K_i$$

- e: fixed endowment
- r_i: number of ratings submitted by participant i
- c: cost per rating
- K_i: correctly guessed positions in the final ranking
- b: reward per correctly guessed position

Back