Assessing the impact of AI-driven recommenders on Human-AI ecosystems

HHAI2025

The tutorial

June 10, 2025
09:00 - 13:00
Aula VI

Who we are

Valentina

Giuliano

Margherita

Virginia

Luca

Mauro

Gizem

SOBIGDATA RESEARCH INFRASTRUCTURE

Consiglio Nazionale delle Ricerche

SCUOLA NORMALE SUPERIORE

KDD Lab

This tutorial

1. Introduction
2. Social Media Ecosystem
3. Online Retail Ecosystem
4. Urban Mapping Ecosystem
5. Generative-AI Ecosystem
6. Open Challenges

Material

- D. Pedreschi et al.
  Human-AI coevolution
  Artificial Intelligence (2025): 104244.

- L. Pappalardo et al.
  A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions
  arXiv:2407.01630 (2024).

Very Large Online Platforms (VLOPs)

[**DSA**, article 33] VLOPs are online platforms with more than **45M average active users** per month in the EU

The **Digital Services Act** (DSA) mandates that:

> "VLOPs need to tackle the risks they pose to Europeans and society when it comes to illegal content and **their impact** on fundamental rights, public security, and wellbeing."

Designated VLOPs

https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses#ecl-inpage-Infinite

updated to February 6th, 2025

facebook
TikTok
amazon
Snapchat
SHEIN
AliExpress
Booking.com
Pornhub
X
Wikipedia
The Free Encyclopedia
Google Maps
Google Play
Instagram
Pinterest
XVIDEOS
LinkedIn
YouTube
TEMU
Google Shopping
zalando

[Article 34, Risk assessment - Digital Services Act]

“Providers of VLOPs [...] **shall diligently identify, analyse, and assess any systemic risks** in the [European] Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services

Recommenders behind VLOPs

Algorithms that **suggest** items or content on VLOPs based on users' preferences or specific requests

- The use *machine learning* to capture users' preferences
- They mediate, *through VLOPs*, most of our actions by exerting instant influence over many specific choices
- Studying the role of recommenders constitutes a **vantage point** to analyse human-AI coevolution

Some examples

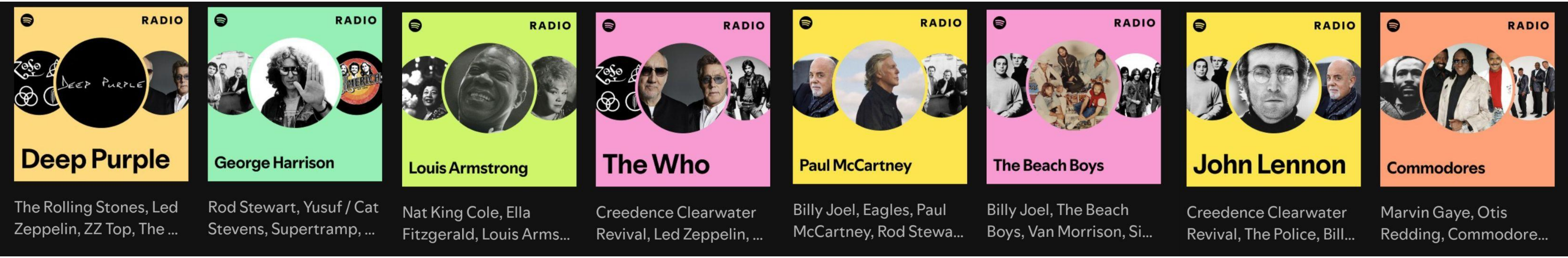
- Personalised suggestions on **social media** guide our content consumption and social connections

- **Online retail** recommenders propose products (e.g., items, songs, movies) for consumption

- **Navigation services** suggest routes to reach our destinations

- **Generative AI** creates content in response to users' wishes.

EXAMPLE OF OUTPUT



1 2 3 4 5 6 7 8

Types of Recommenders

Recommenders are of three main categories:

1. **User-based** collaborative filtering
2. **Item-based** collaborative filtering
3. **Content-based** filtering

Collaborative filtering

4. and combinations of 1, 2, 3

User-based CF

It recommends items to a user based on the **preferences of similar users**

Two steps:

1) select similar users
2) select items from them

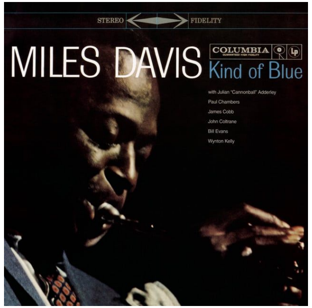
Example:

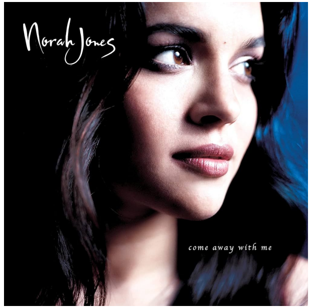
If A and B both like Action Movies:

- A watches *Mad Max*
- B is likely to receive *Mad Max* as recommendation

Rita

Rose





Marc

Item-based CF

It recommends items by finding those similar to what a user interacted with, based on the **preferences of many users**

Two steps:

1) select co-interacted items
2) suggest an item

Example:

If many people who watched *Inception* also watched *Interstellar*:

- the system recommends *Interstellar* to a user who has watched *Inception*

Rita

rock, 1969

rock, 1965

rock, 1966

rock, 1994

rock, 1969

pop/R&B, 2024

pop, 2024

rock, 1973

jazz, 1989

rock, 1988

Content-based Filtering

It recommends items to a user by comparing **item features** with the user's past preferences

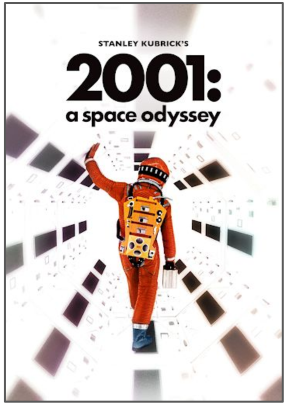
Two steps:

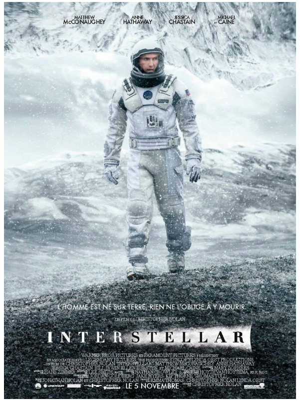
1) extract features for items
2) compute similarities
3) suggest an item

Example:

If a user watches many sci-fi movies, the system recommends other sci-fi movies, even if no other users have watched them

Chen


STANLEY KUBRICK'S
2001:
a space odyssey


INTERSTELLAR
LE 5 NOVEMBRE


PAR LE RÉALISATEUR DE PREMIER CONTACT ET PRISONERS
RYAN GOSLING HARRISON FORD
BLADE RUNNER 2049
LE 4 OCTOBRE


ARRIVAL

2


WRITTEN AND DIRECTED BY
ALEX GARLAND THE WRITER
28 DAYS LATER
EX_MACHINA
THERE IS NOTHING MORE HUMAN THAN THE WILL TO SURVIVE
"A TENSE, THOUGHT-PROVOKING THRILLER"
15
15
DVD


Ben Affleck Jennifer Aniston Drew Barrymore Jennifer Connelly Kevin Connolly Bradley Cooper Ginnifer Goodwin Scarlett Johansson Justin Long
he's just not that into you

1


MATT DAMON
THE MARTIAN

4


5 BAFTAS BEST FILM
6 ACADEMY AWARDS BEST DIRECTOR
RYAN GOSLING EMMA STONE
LA LA LAND
FROM DIRECTOR DAMIEN CHAZELLE
"A SUN-DRENCHED MASTERPIECE"
HERE'S TO THE FOOLS WHO DREAM
12
12

3


Being the adventures of a young man whose principal interests are rape, ultra-violence and Beethoven.
STANLEY KUBRICK'S
CLOCKWORK ORANGE

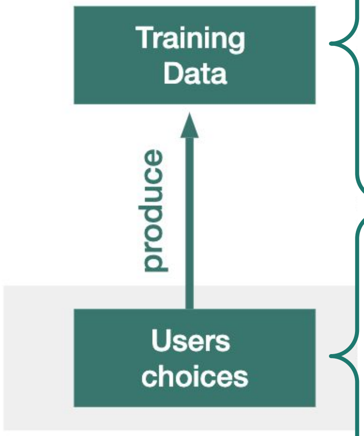
THE FEEDBACK LOOP

Interactions between *users* and *recommenders* always generate a **feedback loop**

- Users' choices determine data on which recommenders are trained;
- The trained recommenders exert influence on users' choices
- Which affect the next round of training
- and so on….

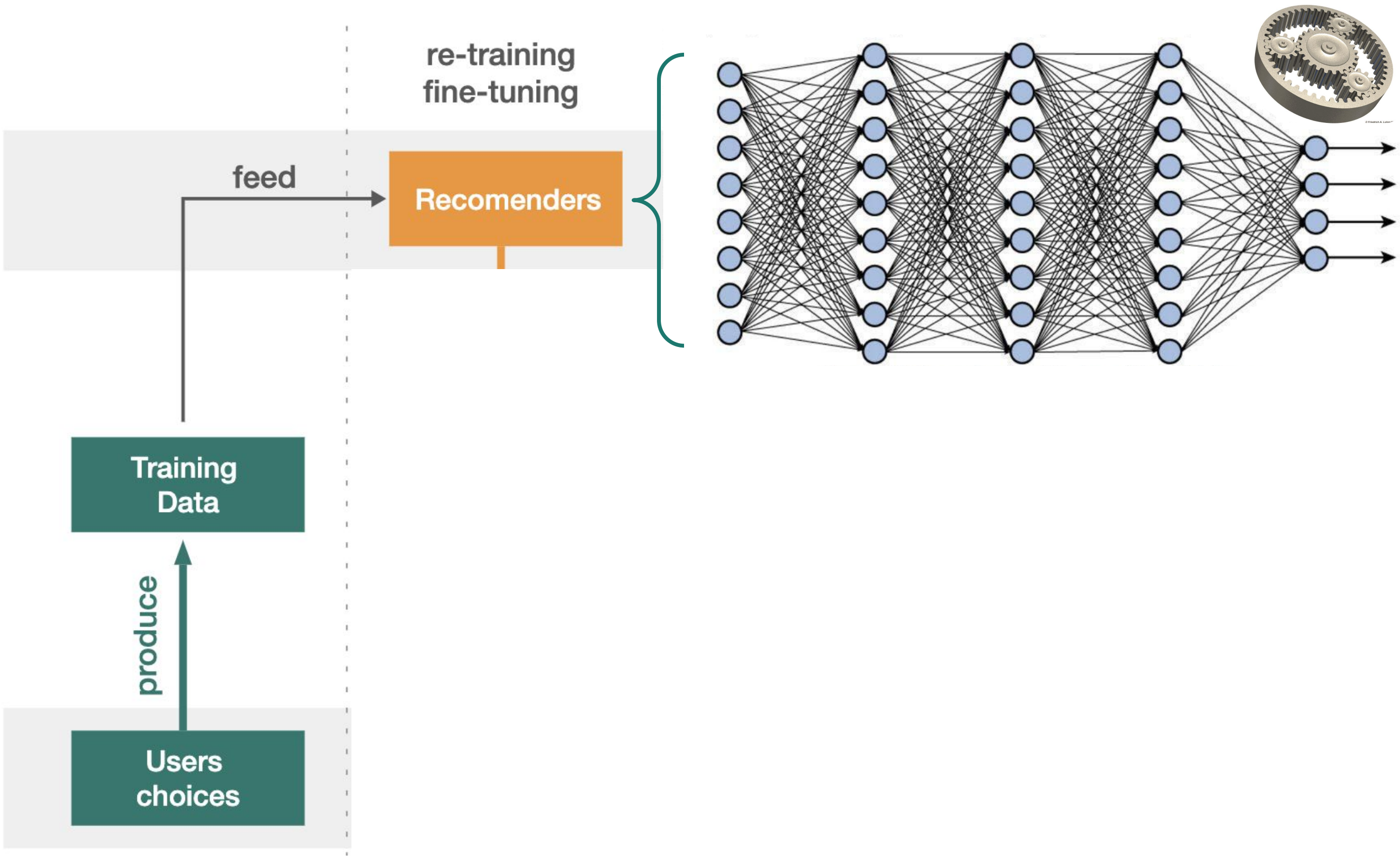
Users choices

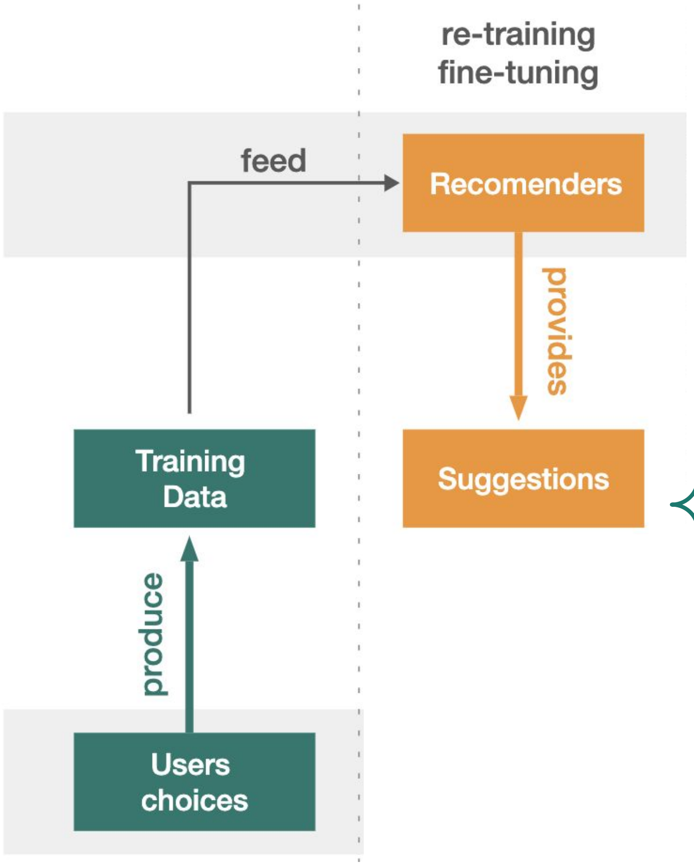
movies selected on Netflix, songs selected on Deezer
products visited or bought on Amazon or Taobao
friends followed (and interactions) on X or Instagram
requests made on DeepSeek or chatGPT
routes requested/followed on TomTom



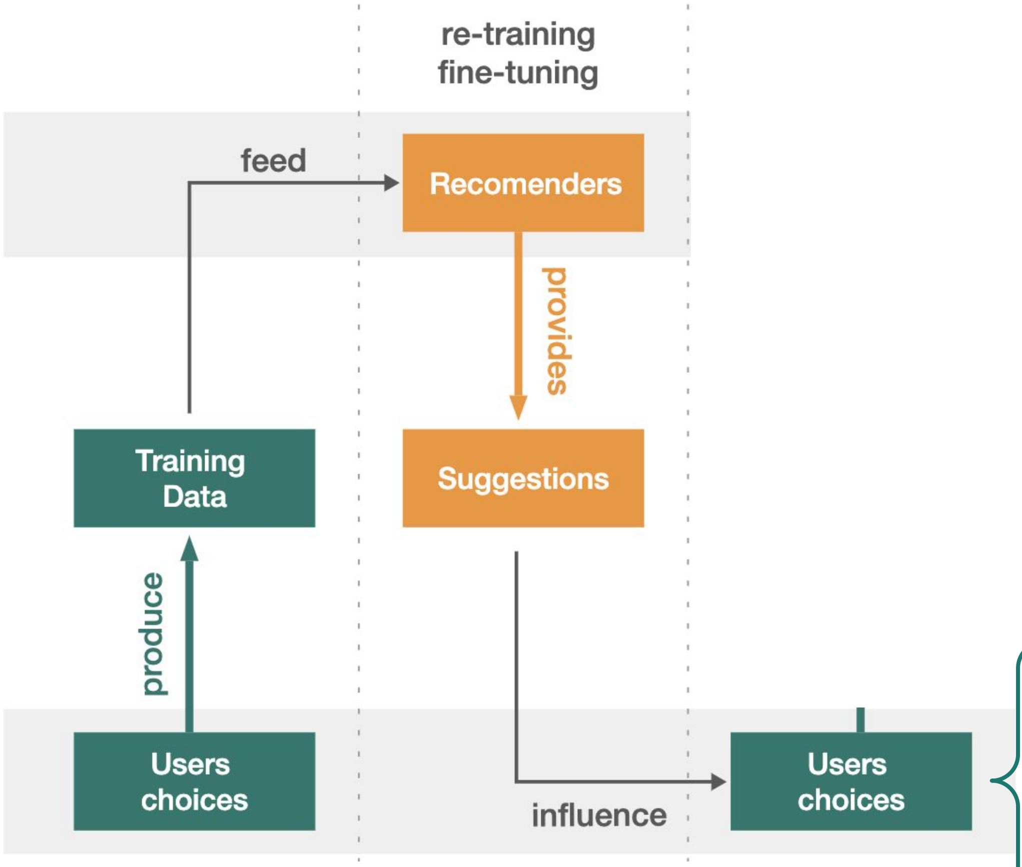
Breaking Bad 👍, The Godfather 👍, He's Just Not That into You 👎
Fifty Shades of Grey 👎, Shantaram 👍, Zen in the Art of Archery 👍
@serenawilliams **+1**, @KingJames ❤️
["What's the capital of Latvia?", "It is Riga"]
Jardin de Luxembourg → Panthéon → Notre-Dame de Paris

movies selected on Netflix, songs selected on Deezer
products visited or bought on Amazon or Taobao
friends followed (and interactions) on X or Instagram
requests made on DeepSeek or chatGPT
routes requested/followed on TomTom

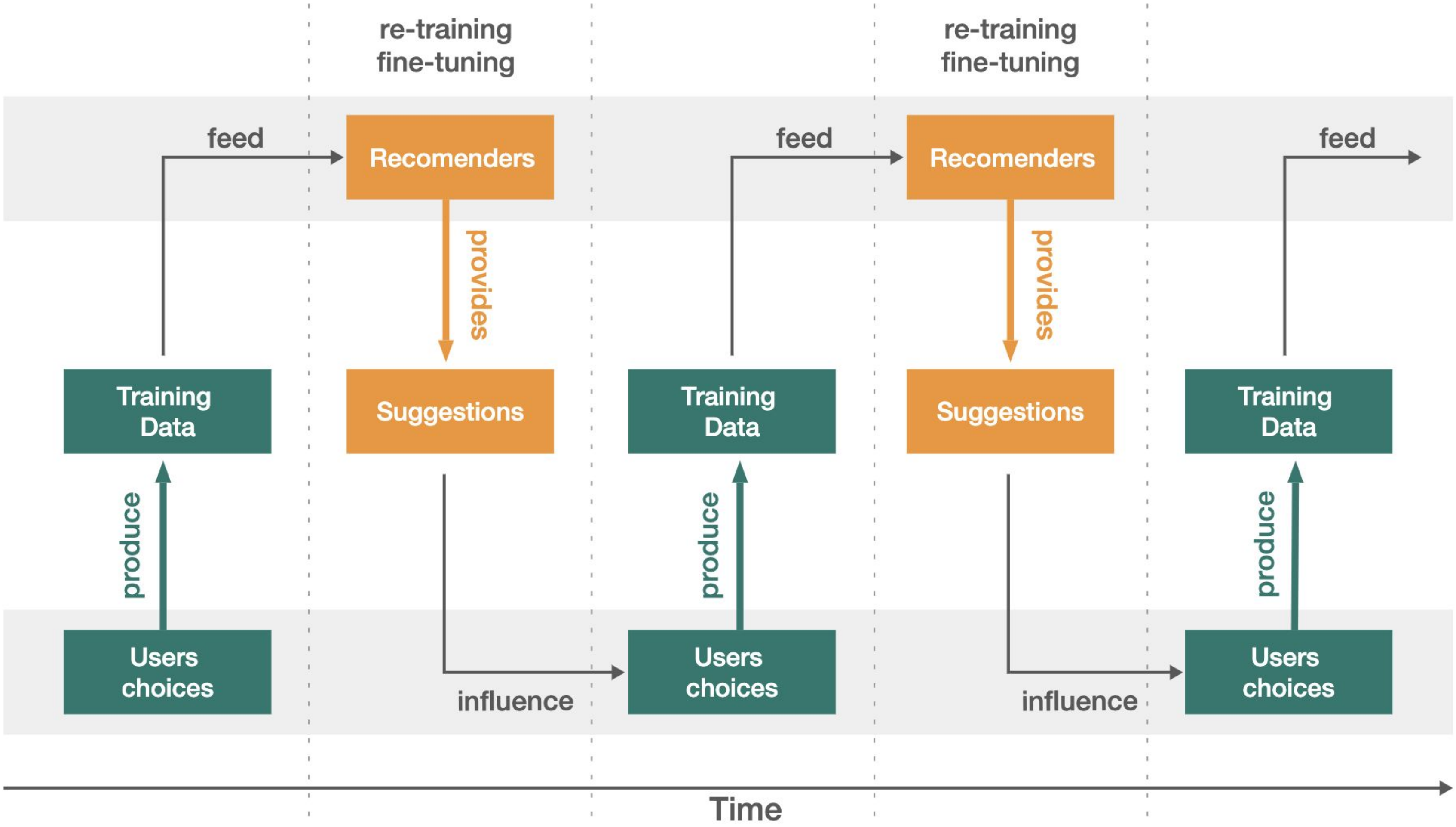

re-training
fine-tuning
feed
Recomenders
Training Data
produce
Users choices



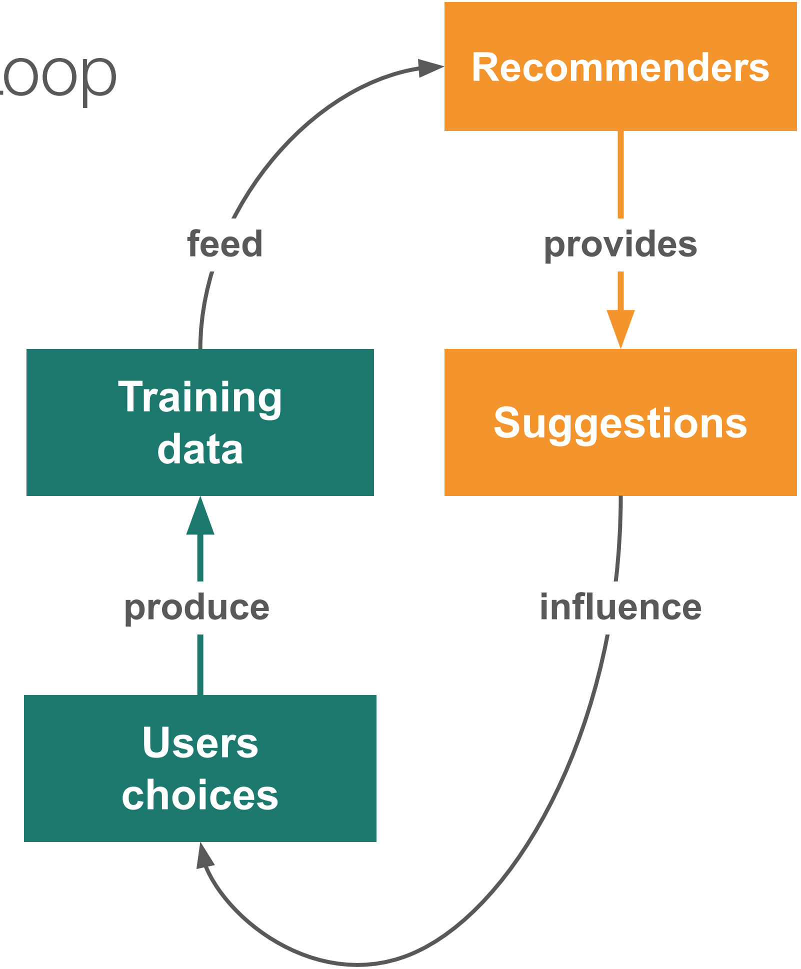
Better Call Saul, The Sopranos
Zen and the art of motorcycle maintenance
@janniksin, @Simone_Biles
“The capital of Latvia is Riga”
Sainte-Chapelle, Musée du Louvre



Better Call Saul, The Sopranos
Zen and the art of …
@janniksin, **@Simone_Biles**
“The capital of Latvia is Riga” 👍
Sainte-Chapelle, **Musée du Louvre**


re-training
fine-tuning
re-training
fine-tuning
feed
Recomenders
feed
Recomenders
feed
provides
provides
Training
Data
Suggestions
Training
Data
Suggestions
Training
Data
produce
produce
produce
Users
choices
influence
Users
choices
influence
Users
choices
Time

The Feedback Loop



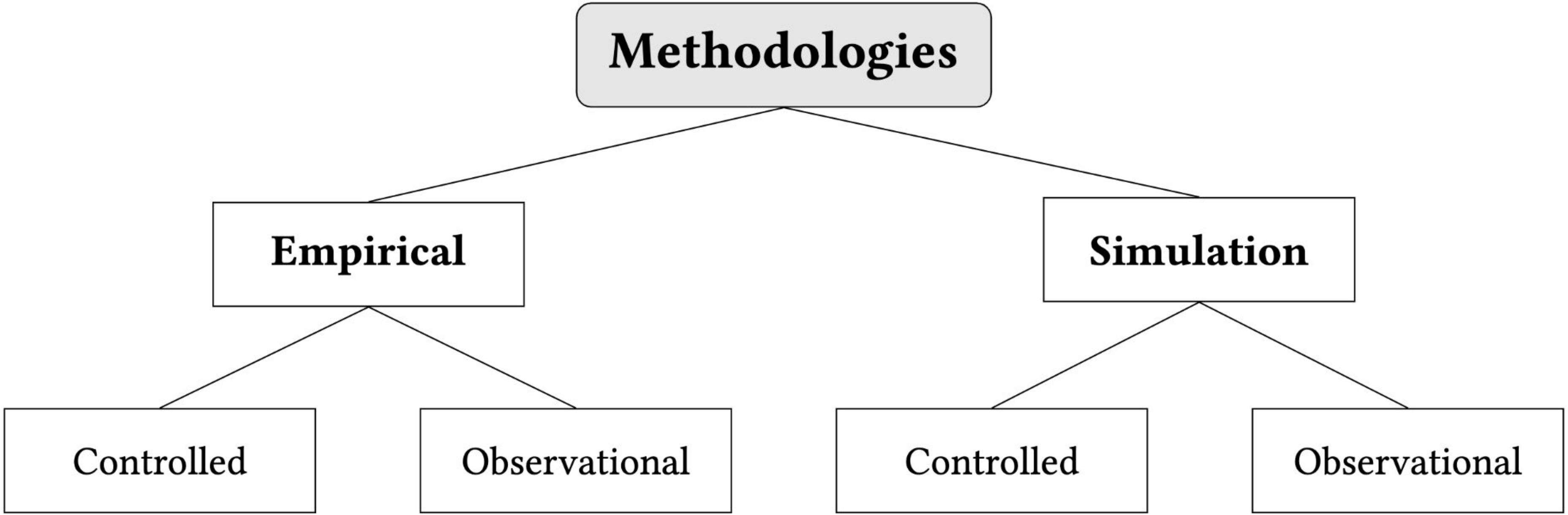
Tracking the feedback loop

- **We need to track all phases:**
  - users’ choices
  - recommendations provided
  - recommendations accepted/interacted
  - details on (re)training

- Data rarely (if never) available

- The impact of recommenders can only be estimated based on the observation of users’ choices

Overview of methodologies



EMPIRICAL STUDIES

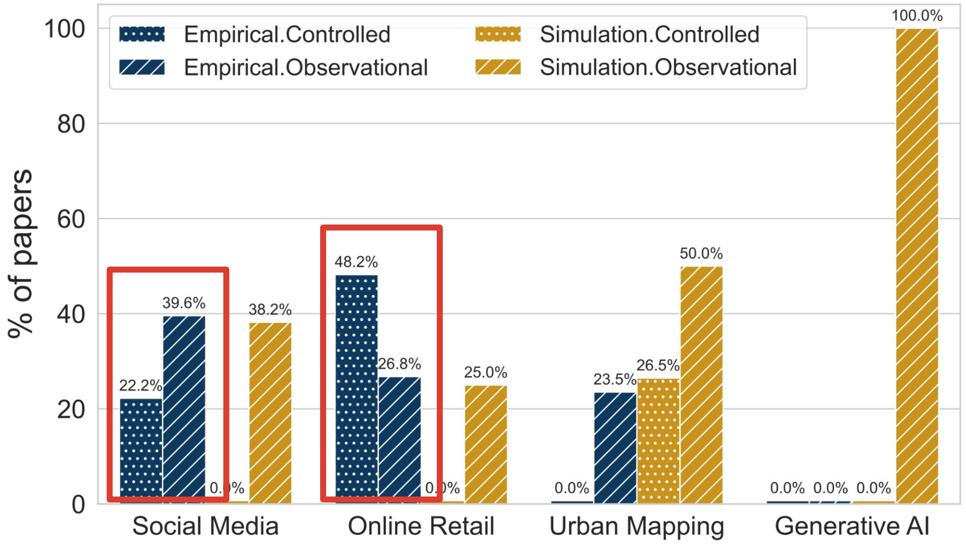
Based on data generated as a by-product of users' activity on VLOPs

- **big data**
- **lab data**
  - *real* users interacting in laboratory settings
  - *bots* that simulate human behaviour

EMPIRICAL STUDIES

Prevalent in **social media** and **online retail**:

- social media: analysis of data from *sock-puppets* simulating users' behaviours
- online retail: analysis of data from e-commerce platforms



A survey of the impact of AI-based recommenders on human behaviour: methodologies, outcomes and future directions, ArXiv, 2024

SIMULATION STUDIES

Based on *synthetic* data generated through a model:

- mechanistic
- AI-based
- digital-twin-based

SIMULATION STUDIES

Mechanistic models:

- based on known physical, biological, or social principles
- incorporate causal relationships

AI-based models:

- rely on *machine learning* to learn patterns from data without explicitly encoding physical or causal mechanisms

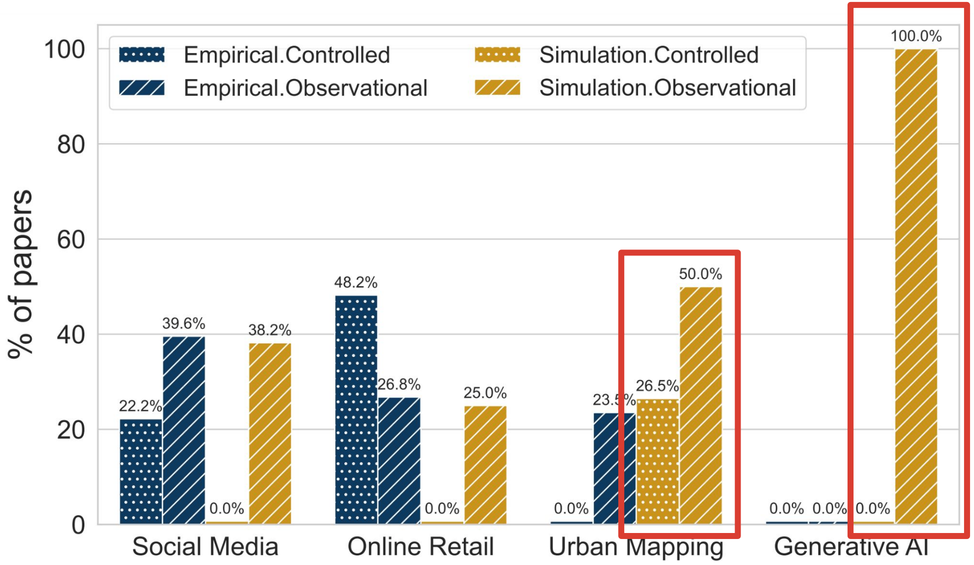
Digital-twins:

- a virtual replica of a physical system that continuously updates based on real-time data from its physical counterpart

SIMULATION STUDIES

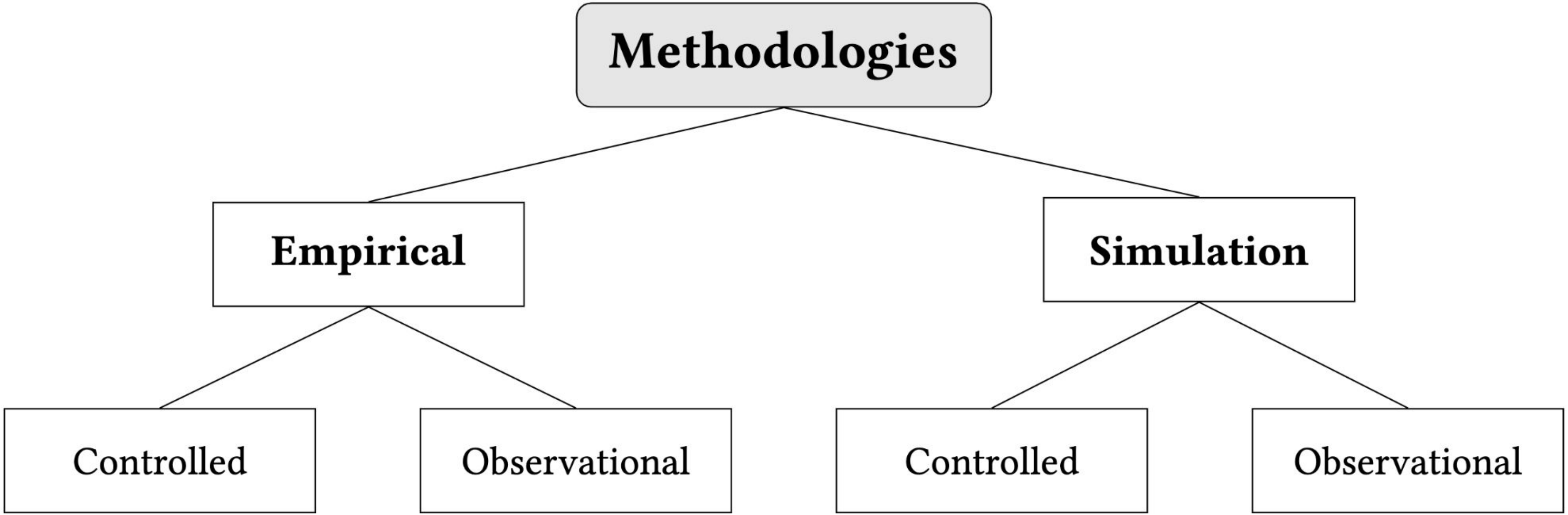
Prevalent in **urban mapping** and **genAI**:

- urban mapping: hard to get detailed data from platforms
- genAI: impossible to get data from platforms



A survey of the impact of AI-based recommenders on human behaviour: methodologies, outcomes and future directions, ArXiv, 2024

How to study Human-AI coevolution?



Daniel in the Bible

597 BC: the king of Babylon sacked the kingdom of Judah

- He brought thousands of captives to Babylon
- He commanded *Ashpenaz* to reeducate children in the language and culture of Babylon,
  - to serve in his court
- As part of the education, they would get to eat royal meat and drink royal wine

Daniel in the Bible

Daniel, refused to touch royal meet.

- He proposed a 10-days experiment
  - to convince Ashpenaz that vegetarian diet is good as well

- Four children will be feed with vegetarian diet (treatment group)
- Four children will be feed with carnivore diet (control group)

CONTROLLED STUDIES

Users are split into:

- **experimental group(s)**
  users do receive a recommendation

- **control group**
  users do <u>not</u> receive a recommendation

Differences in behaviour are analyzed

CONTROLLED STUDIES

Users are split into:

- **experimental group(s)**
  users do receive a recommendation

- **control group**
  users do <u>not</u> receive a recommendation

Differences in behaviour are analyzed

Analogy with medical experiments:

- patients in experimental group(s) receive a drug

- patients in the control group receive a placebo or nothing

CONTROLLED STUDIES

"the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units"
D. R. Cox, Planning of Experiments, 1992

- Known as the Stable Unit Treatment Value Assumption (SUTVA)

On online platforms, users in the control group <u>can never be isolated</u> from the indirect effects of recommendations: they are influenced by choices by users in the treatment group

- This violets SUTVA

OBSERVATIONAL STUDIES

Data describe the behaviour of users **under a single recommendation principle**, without any control

- users are not split into separate groups *at the same time*

- in this definition, *quasi-experiments* are not controlled

OBSERVATIONAL STUDIES

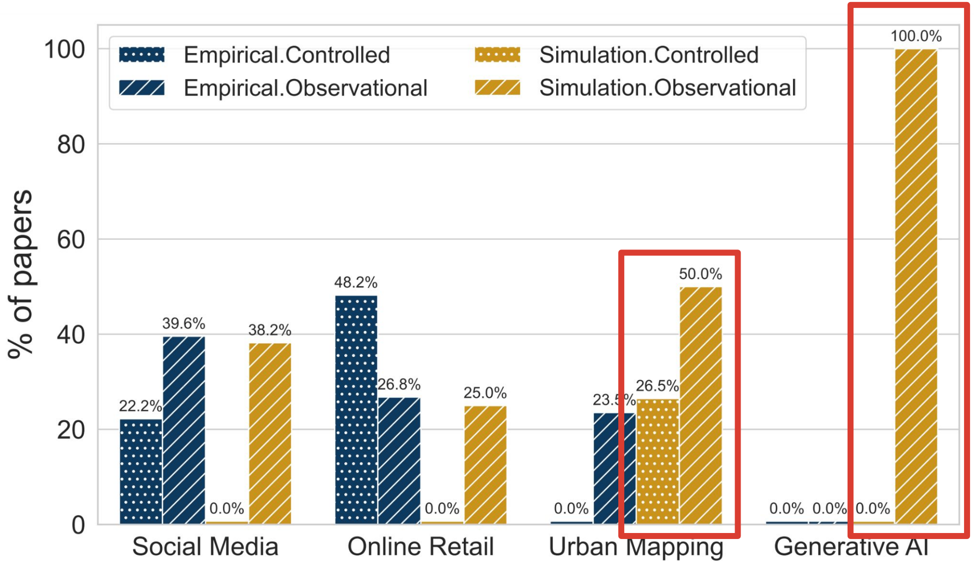
Data describe the behaviour of users **under a single recommendation principle**, without any control

- e.g., behaviour of Facebook users, routes followed by drivers

CONTROLLED vs OBSERVATIONAL

Observational studies are typically more common than controlled ones:

- they are easier to perform
- online retail is an exception



A survey of the impact of AI-based recommenders on human behaviour: methodologies, outcomes and future directions, ArXiv, 2024

Nutrition – Do Saturated Fats Increase Heart Disease Risk

	Observational Studies	Controlled Trials
Example	Framingham Heart Study, others (1960s–90s)	PURE study, RCT meta-analyses (2010s–)
Finding	Higher saturated fat intake → higher heart disease risk	No clear link between saturated fat and heart disease/mortality
Limits	Confounding from lifestyle (e.g., smoking, processed food)	Diet strictly controlled; fewer confounders
Policy Impact	Led to low-fat dietary guidelines	Challenged one-size-fits-all dietary recommendations

Dawber et al. 1951. Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health*, 41(3)
SiriTarino et al 2010. Meta-analysis of prospective cohort studies evaluating the association of saturated fat with cardiovascular disease. American Journal of Clinical Nutrition, 91(3), 535–546

Social Science – Violent Media and Aggression

	Observational Studies	Controlled Experiments
Example	Longitudinal and survey studies (1970s–2000s)	Lab RCTs, meta-analyses (e.g., Anderson & Dill, 2000)
Finding	Correlation between violent content and aggression	Short-term effects in lab; long-term effects unclear
Limitation	Reverse causality; confounding (e.g., parenting)	Controlled exposure and outcome measures
Interpretation Shift	Media blamed for societal aggression	Effects appear weak, contextual, and non-generalizable

Huesmann et al. 2003 Longitudinal relations between children's exposure to TV violence and their aggressive and violent behavior in young adulthood. Developmental Psychology
Anderson and Dill 2000 Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. Journal of Personality and Social Psychology

Quiz session

Experiment:

- A study selects 10 users on a social media platform
- On day 1, the users are exposed to recommender R1
- On day 2, the users are exposed to a recommender R2
- The number of new likes is computed for each user

Is this an empirical or a simulation study?

✓ A. Empirical

✗ B. Simulation

Quiz session

Experiment:

- A study selects 10 users on a social media platform
- On day 1, the users are exposed to recommender R1
- On day 2, the users are exposed to a recommender R2
- The number of new likes is computed for each user

Is this an observational or controlled study?

✓ A. Observational

✗ B. Controlled

Quiz session

Experiment:

- Two groups of users (G1-G2) are selected on an online retail platform
- G1 is exposed to recommender R1
- G2 is exposed to recommender R2
- The diversity of purchased products is measured

Is this an empirical or a simulation study?

✓ A. Empirical

✗ B. Simulation

Quiz session

Experiment:

- Two groups of users (G1-G2) are selected on an online retail platform
- G1 is exposed to recommender R1
- G2 is exposed to recommender R2
- The diversity of purchased products is measured

Is this an observational or controlled study?

✗ A. Observational

✓ B. Controlled

Quiz session

Experiment:

- Two groups of users (G1-G2) are selected on a platform
- G1 is exposed to recommender R1
- G2 is exposed to recommender R2
- The accuracy of R1 and R2 is evaluated

Is this an empirical or a simulation study?

✗ A. Empirical

✗ B. Simulation

Quiz session

In a study evaluating the effect of a new fertilizer on crop yield, farmers are randomly assigned either the new fertilizer (treatment) or their usual fertilizer (control). Some of the treated farmers share surplus fertilizer with neighboring untreated farmers, who then apply it to parts of their fields.

A. Respects SUTVA

B. Violates SUTVA ✓

Articles

- L. Pappalardo et al. **A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions**, 2024, https://doi.org/10.48550/arXiv.2407.01630
  - Section 2.3 Methodologies
  - Section 7.1 Methodologies
  - Section 7.2 Methodologies

- H. O. Stolberg et al. **Randomized controlled trials**. American Journal of Roentgenology 2004, https://doi.org/10.2214/ajr.183.6.01831539

Books & articles

- J. Pearl, **The Book of Why: The New Sciences of Cause and Effect,** Basic Books, 2018
- D. R. Cox, **Planning of Experiments**, John Wiley & Sons, 1992

- C. Haney, W. C. Banks, P. G. Zimbardo, **A study of prisoners and guards in a simulated prison**, Naval Research Review 30, 1973
- J. W. Treece Jr., **Daniel and the Classic Experimental Design**, ICR.org
- A. Huxley, **Brave New World**, Chatto & Windus, 1932

The Experiment
2010

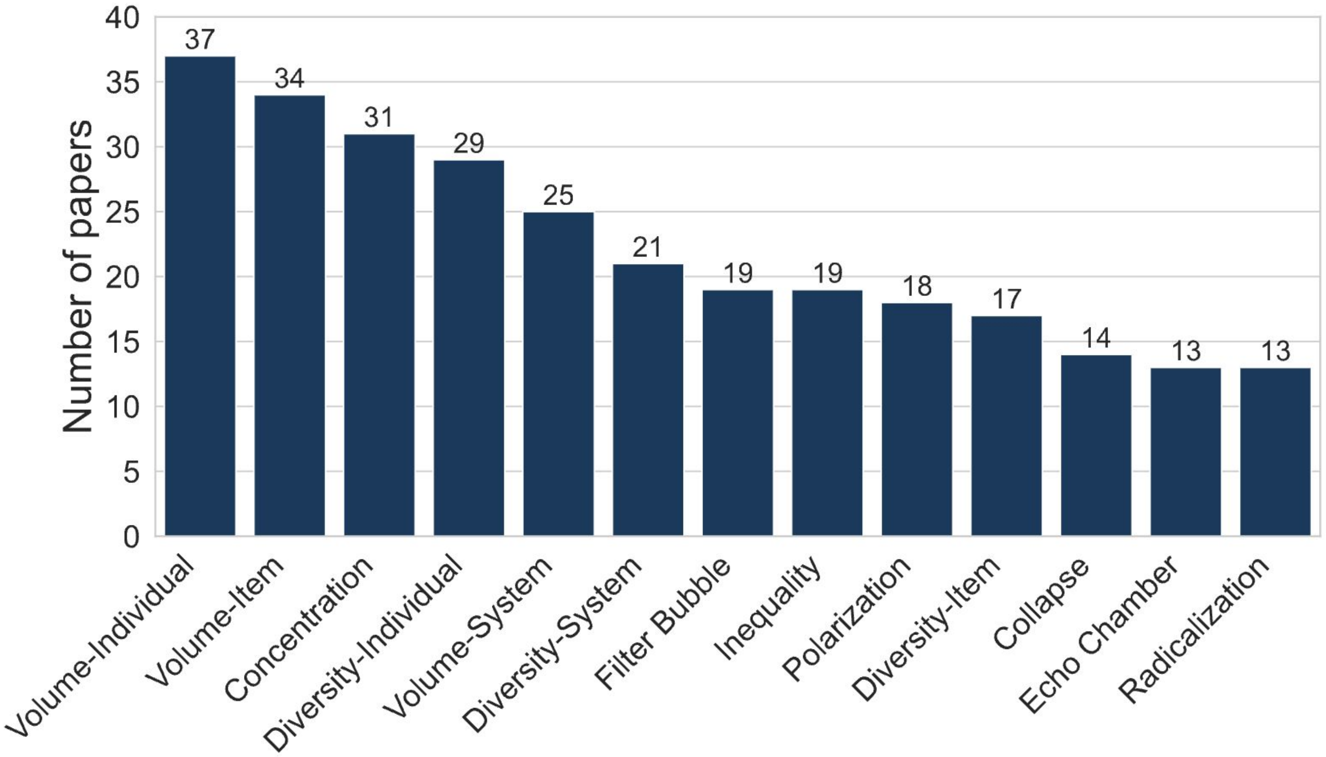
Unintended Consequences

- Personalised recommendations on social media may artificially amplify **echo chambers**, **filter bubbles**, and **radicalisation**

- Profiling and targeted advertising may increase **inequality** and monopolies, accruing **biases** and **discriminations**

- Navigation services suggest routes that may create **congestion** if too many drivers are sent to the same roads

Level	Outcome	Description	Ecosystems
Individual	Diversity	Variety of users' behaviour, items consumed and users followed	SM, OR, UM
	Filter Bubble	Conformation of items or contents with own preferences or beliefs	SM, OR
	Radicalization	Items or individual attributes going towards an extreme	SM
	Volume	Quantity value of some users' attribute	SM, OR, UM
Item	Diversity	Variety of users that consume the item	SM, OR, GAI
	Volume	Quantity value of some items' attribute	SM, OR, UM
Model	Collapse	AI model degradation over time	GAI
Systemic	Concentration	Close gathering of people or things	SM, OR, UM
	Diversity	Aggregate diversity of users or items	SM, OR, UM
	Echo Chamber	Environment reinforcing opinions or item choices within a group	SM, OR, UM
	Inequality	Uneven distribution of resources/opportunities among group members	SM, OR, UM
	Polarization	Sharp separation of users/items into groups based on some attributes	SM
	Volume	Aggregate volume of users' or items' attributes	SM, OR, UM



Number of papers
40
35
30
25
20
15
10
5
0
37
34
31
29
25
21
19
19
18
17
14
13
13
Volume-Individual
Volume-Item
Concentration
Diversity-Individual
Volume-System
Diversity-System
Filter Bubble
Inequality
Polarization
Diversity-Item
Collapse
Echo Chamber
Radicalization

SOCIAL MEDIA

Recommender systems shaping our digital discourse and information consumption

Helping users navigate everyday choices such as **what content** to engage with, **who to follow**, and **which communities** to join.

THE WALL STREET JOURNAL.

the facebook files

A Wall Street Journal investigation

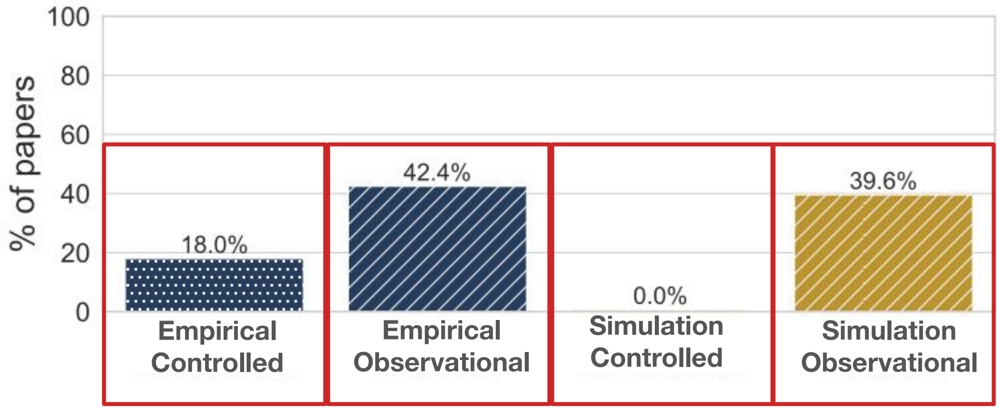
- XCheck **whitelists VIP users**, allowing them to post rule-violating material (e.g., harassment or incitement to violence)
- Instagram is **toxic for teen girls**, increasing anxiety and depression
- The new algorithm introduced in 2018 **made people angrier**

"You see a theme in all these documents that Facebook and its top executives know what their problems are, but in many instances, can't, or won't address them sometimes because it fears hurting the business or growth."

The Facebook Files Podcast, The Wall Street Journal

Employed Methodologies

A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630



- Predominance of empirical over simulation studies
- Controlled experiments are mainly conducted internally by the platforms themselves

Main Outcomes

A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630

Filter Bubble

A state where users are repeatedly exposed to content aligning with their prior beliefs or preferences.

Echo Chamber

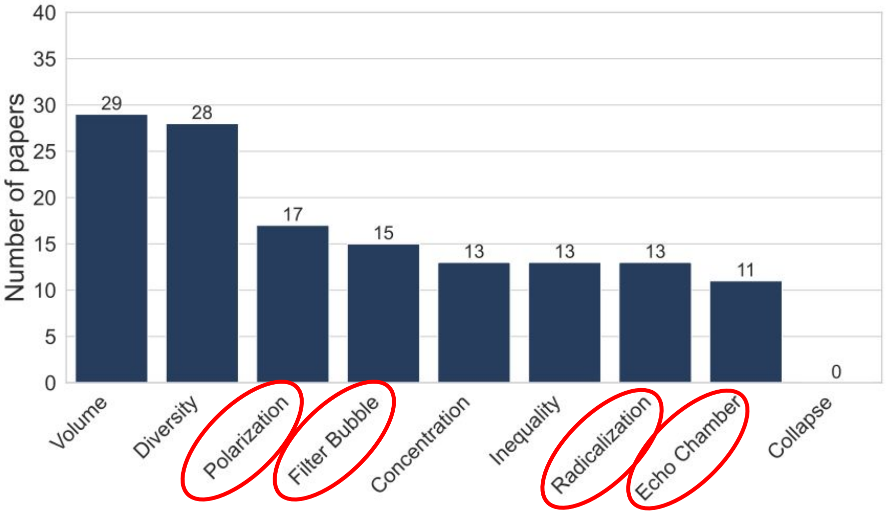
A reinforcing environment where opinions or choices are amplified within a group, limiting outside exposure.

Polarization

Sharp division of users or content into opposing ideological groups, reducing middle-ground visibility.

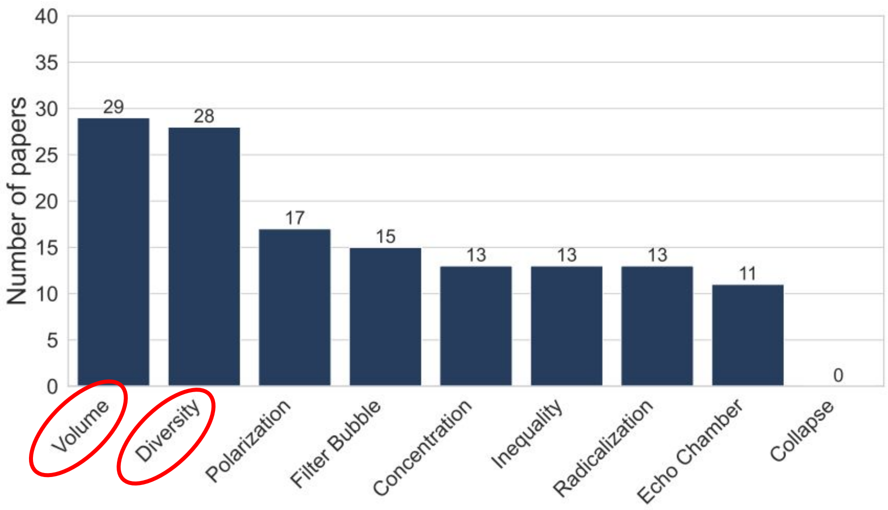
Radicalization

Exposure or drift towards more extreme views or content over time.



Main Outcomes

A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630



- Main Outcomes:
    - Systemic-level focus: **volume**, **diversity**, **concentration**
    - **Volume** and **diversity** are the primary metrics across studies
    - Common **targets**: Polarization, Filter bubble, Radicalization and Echo Chamber

Empirical Studies

Amplification of politics on Twitter

Huszár et al., PNAS 2021, https://doi.org/10.1073/pnas.2025334119

Type: Empirical controlled

VLOP: Twitter

Outcomes:
volume (increase)
inequality (increase)

Amplification of politics on Twitter

- Original Twitter's recommender: users obtain content from accounts they followed in **reverse chronological order**
- In 2016, a ***content-based filtering*** recommender was introduced:
    - users see tweets deemed relevant
      (both older ones and from accounts they do not follow)

POLITICS

When Twitter users hear out the other side, they become more polarized

Echo chambers aren't what's polarizing America.

by **Ezra Klein**
Oct 18, 2018, 2:30 PM GMT+2

Vox

Amplification of politics on Twitter

Does Twitter's recommender systematically prioritize certain political content * by giving them greater visibility in users' feeds and recommendations?

* such as left vs. right, center vs. extremes, specific parties, or news sources with particular ideological leanings

Experimental Setup

- **Control group**: 1% of global users (randomly chosen) excluded from the personalized Home Timeline
    - which still displays tweets in reverse chronological order
- **Treatment group**: 4% of users (randomly chosen) that experience the personalised Home Timeline
- This assignment is maintained over the lifespan of accounts

Tens of millions of users considered

Measuring amplification

The **reach** of a set *T* of tweets in a set *U* of users is the total number of users in *U* who *encountered* a tweet from *T*

Example:

- *T* can be the set of tweets from politicians of Socialist Party in France
- *U* can be the set of French Twitter users in the control group
- the reach of *T* is how many of French users in the control group encounter tweets from the politicians in the Socialist Party

Measuring amplification

The **amplification ratio** of a set T of tweets is defined as:

$$\frac{\text{reach of T in the treatment group}}{\text{reach of T in the control group}}$$

The ratio is normalized so that:

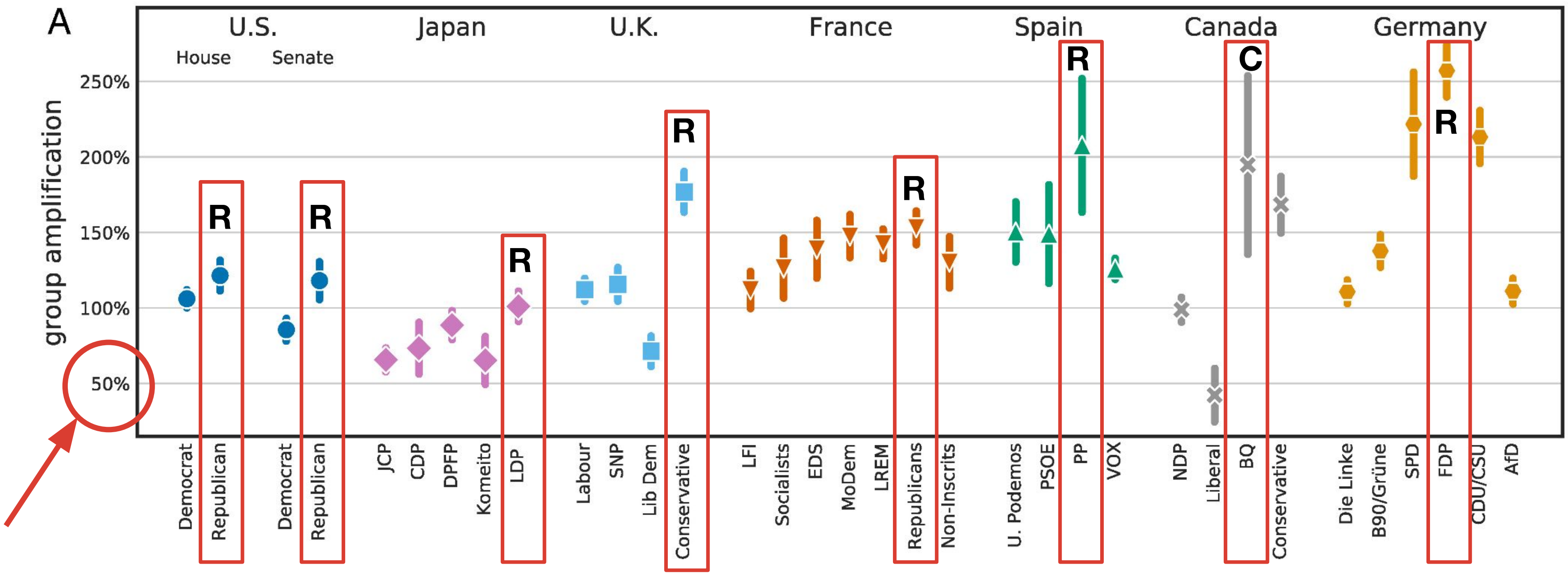
- 0%: equal proportional reach in treatment and control groups
- 50%: the treatment group is 50% more likely to encounter a tweet

Experimental Setup

- **3,643** Twitter accounts related to *currently serving* **legislators**
  - US, Canada, Japan, UK, France, Germany, Spain
    (>100k users in the control group)

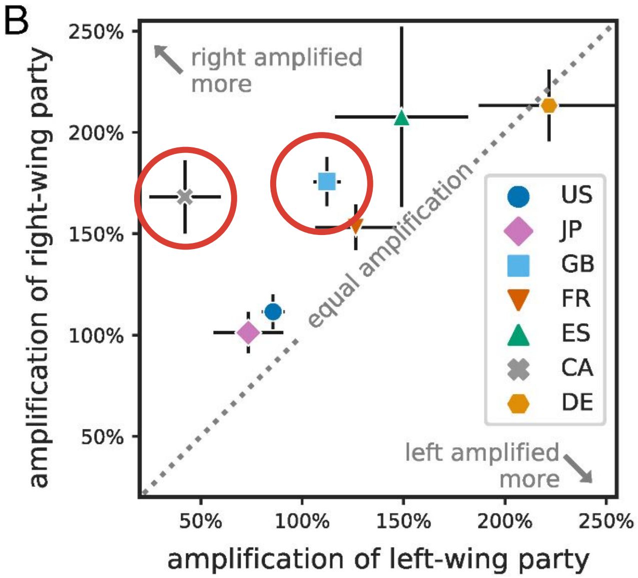
- all tweets, replies and quote tweets are considered
- the **reach** of tweets is computed in the respective country only

Group amplification: All tweets of legislators' accounts of a party



- **Amplification > 50%**
- in some cases > 200%
    - tweets exposed to an audience 3 times larger than that reached with the reverse chronological recommender

Group amplification: All tweets of legislators' accounts of a party



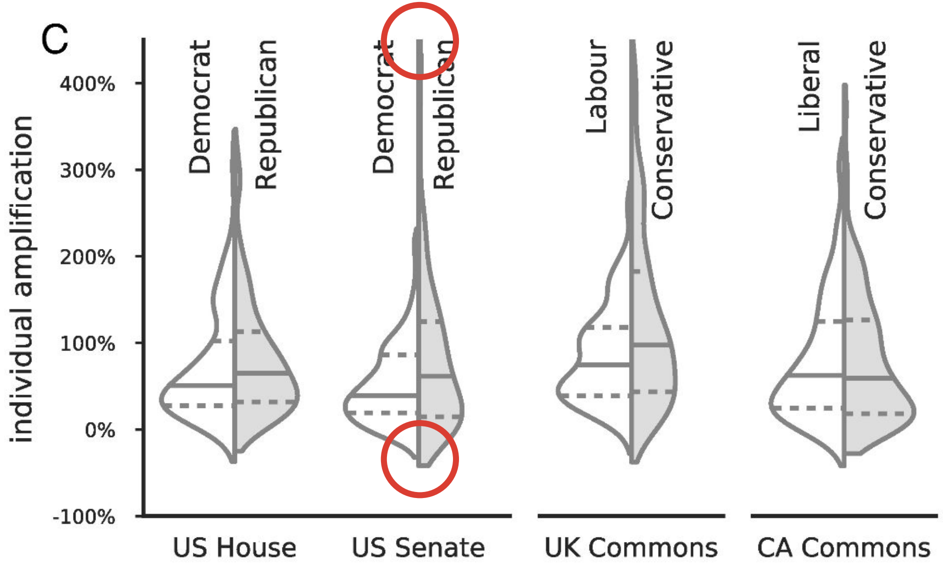
- The ***largest mainstream*** (center-)left and (center-)right parties are selected
- Statistical significant difference **favouring tweets from the political right wing** (except for Germany)

- **Canada**: Left 43% vs Right 167%
- **UK**: Left 112% vs Right 176%

Individual amplification: tweets of individual politicians

Amplification varies:

- Some politicians' amplification is up to **400%**
- for others, it is below **0%**



When comparing individual amplification between parties:

- **no significant association** between an individual's party affiliation and amplification

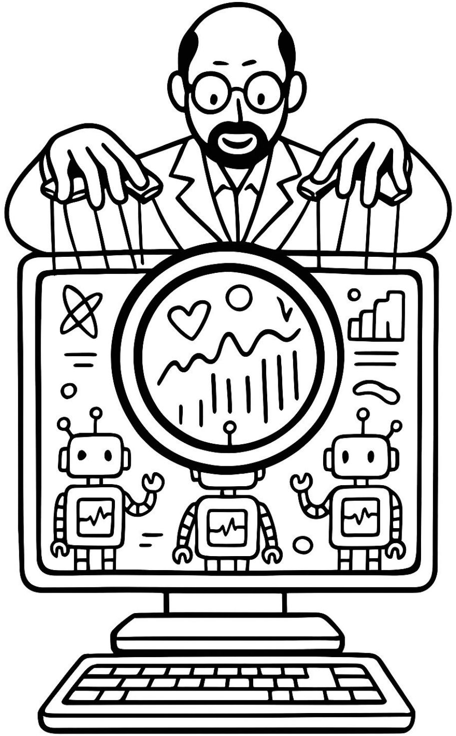
...Two truths and a lie...

Which of these statements is NOT supported by the study?

A. Tweets from political **right-wing** parties were generally **amplified more** than those from left-wing ones

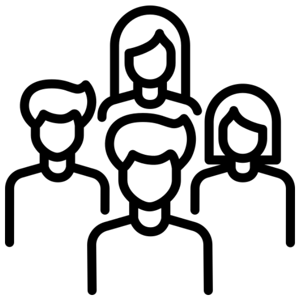
B. **Individual** politicians' amplification was s**ignificantly correlated** with their **party affiliation**

C. The **control group** saw tweets in **reverse-chronological** order, without personalization



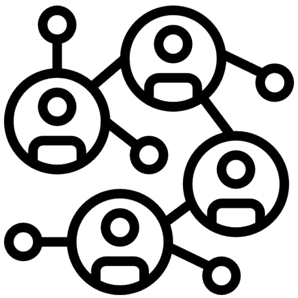
Simulation Studies

SIMULATION INGREDIENTS



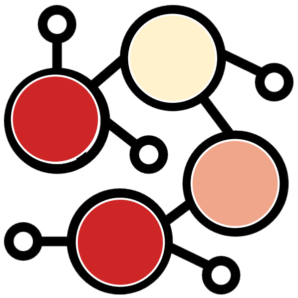
Agents

How many?
How complex?



Social Network

From mean-field to complex, adaptive, higher order topologies

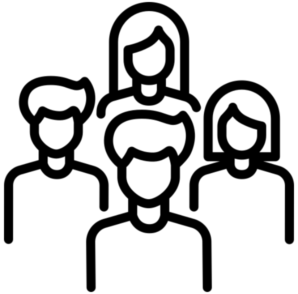


Social Dynamic

Opinion evolution, information diffusion, relationships evolution, content engagement…

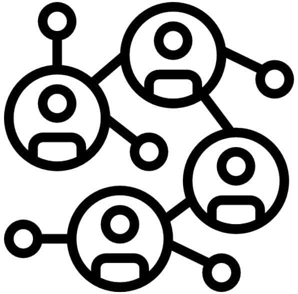
Simulation: a sequence of micro-actions and -interactions *among agents between agents and their environment* (e.g., news) where each interaction "changes something"

SIMULATION INGREDIENTS



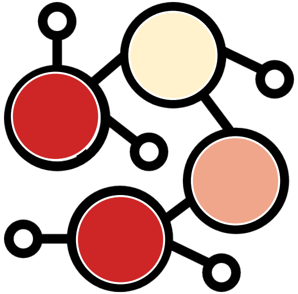
Agents

How many?
How complex?



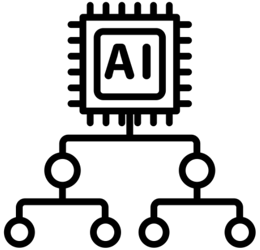
Social Network

From mean-field to complex, adaptive, higher order topologies



Social Dynamic

Opinion evolution, information diffusion, relationships evolution, content engagement…



Recommender Systems

From simplified abstractions to state of the art algorithms

Recommender Systems: *mediate* these interactions.

Algorithmic bias amplifies opinion polarization: A bounded confidence model

Sirbu et al., PLOSOne, 2019

Type: Simulation Observational

VLOP: None

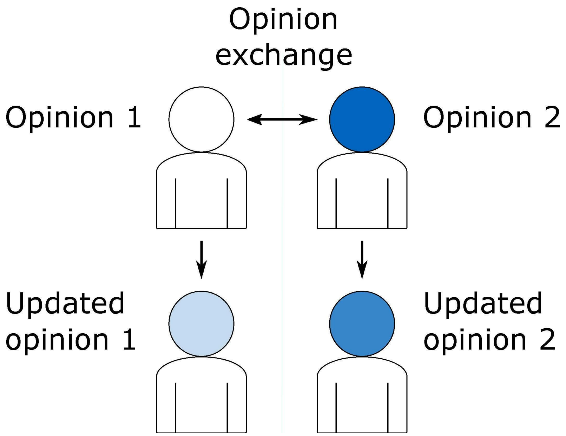
Outcomes: Polarization

ALGORITHMIC BIAS: the design

A recommender system will create an **algorithmic bias** that will skew interactions towards like-minded individuals creating “echo chambers” or “filter bubbles” and thus fostering **polarization**.

How can we test this hypothesis?

By means of opinion dynamics **simulations**!

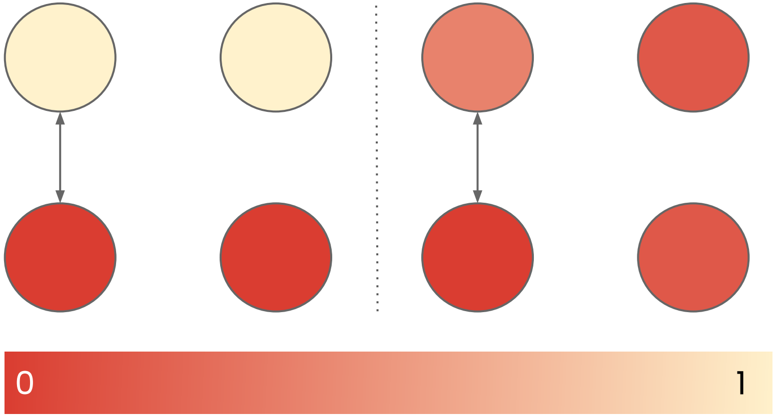


ALGORITHMIC BIAS: the design

Model without Recommender: Deffuant-Weisbuch Model

Fully mixed population of N individuals
Opinions $x_i \in [0,1]$ uniformly distributed

Two random agents *i* and *j* interact with **bounded confidence ϵ**



$$x_i(t+1) = \begin{cases} x_i(t) + \mu(x_j(t) - x_i(t)) \text{ iff } \boxed{|x_i - x_j| < \epsilon} \\ x_i(t) \end{cases}$$

ALGORITHMIC BIAS: the design

Model with Recommender: γ

Fully mixed population of N individuals
Opinions $x_i \in [0,1]$ uniformly distributed

Two ~~random agents *i* and *j* interact with~~
bounded confidence ϵ

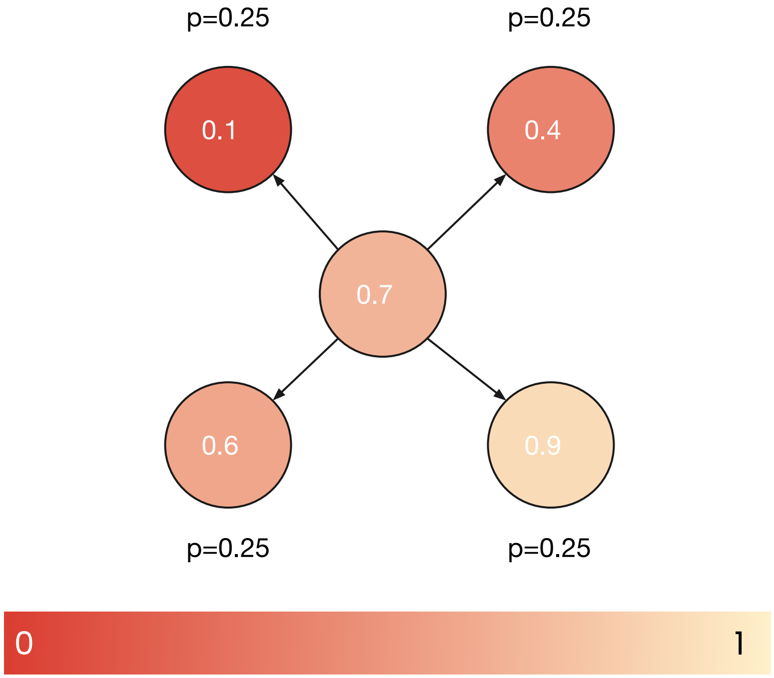
Biased interactions

The higher γ (algorithmic bias) the higher the probability to interact with similar individuals:

$$p_i(j) = \frac{d_{ij}^{-\gamma}}{\sum_{k \neq i} d_{ik}^{-\gamma}}$$

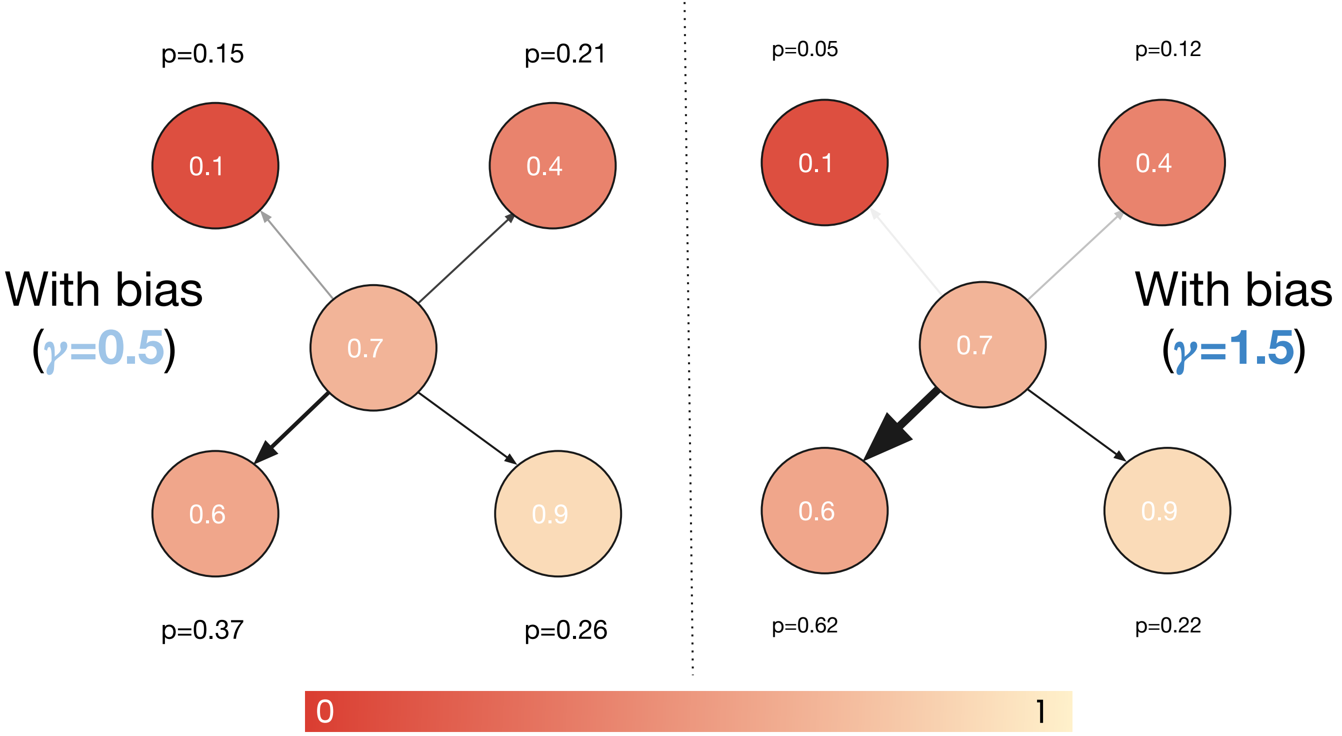
ALGORITHMIC BIAS: the design

Model without Recommender: interaction probability



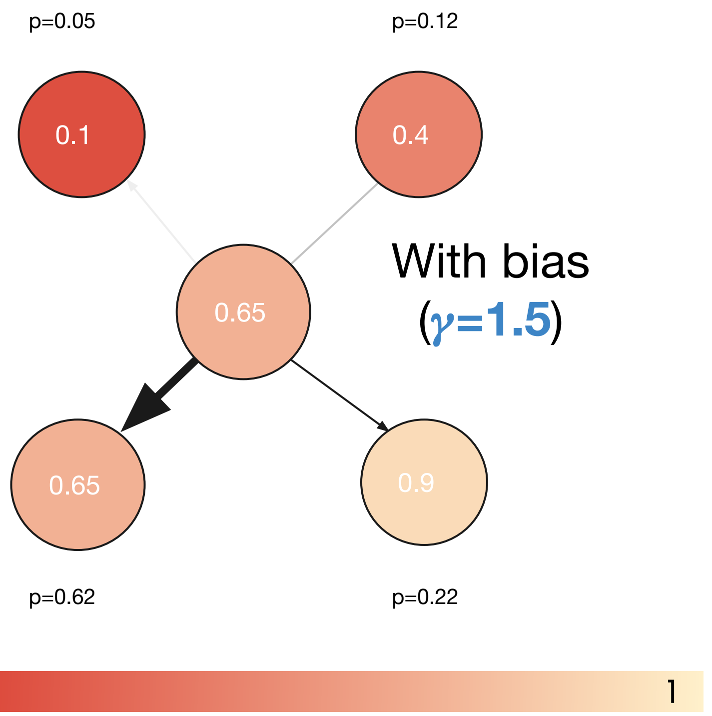
ALGORITHMIC BIAS: the design

Model with Recommender: interaction probability



ALGORITHMIC BIAS: the design

Model with Recommender: interaction probability

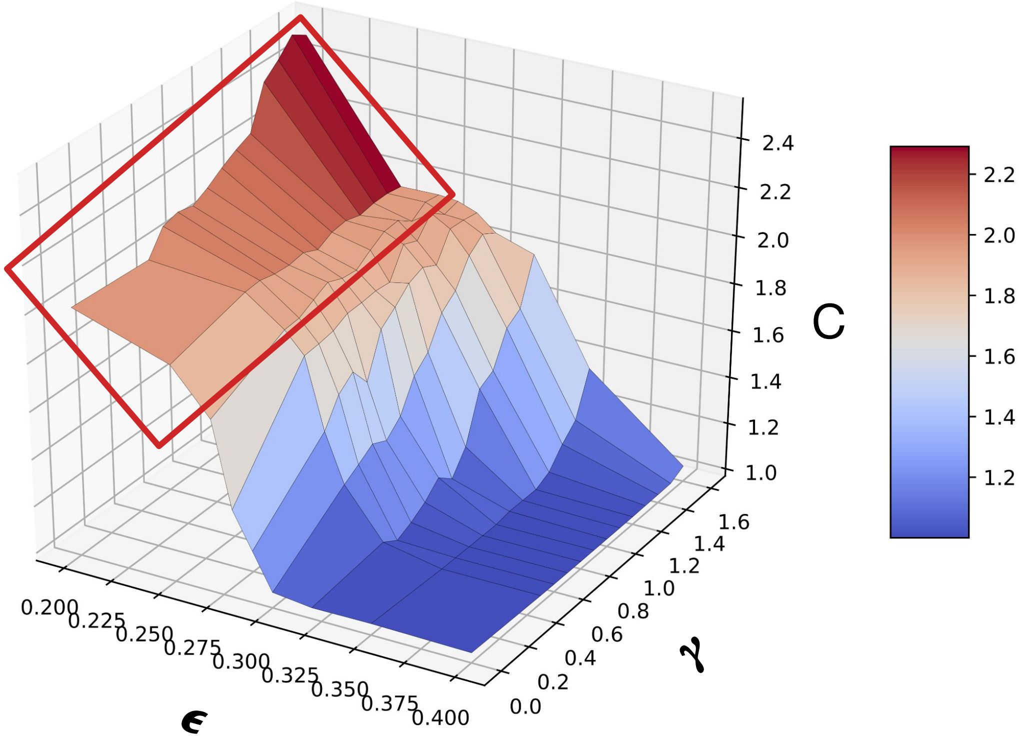


ALGORITHMIC BIAS: results

Effective Number of Clusters*

In the final opinion distribution

$$C = \frac{\left(\sum_i c_i\right)^2}{\sum_i c_i^2}$$



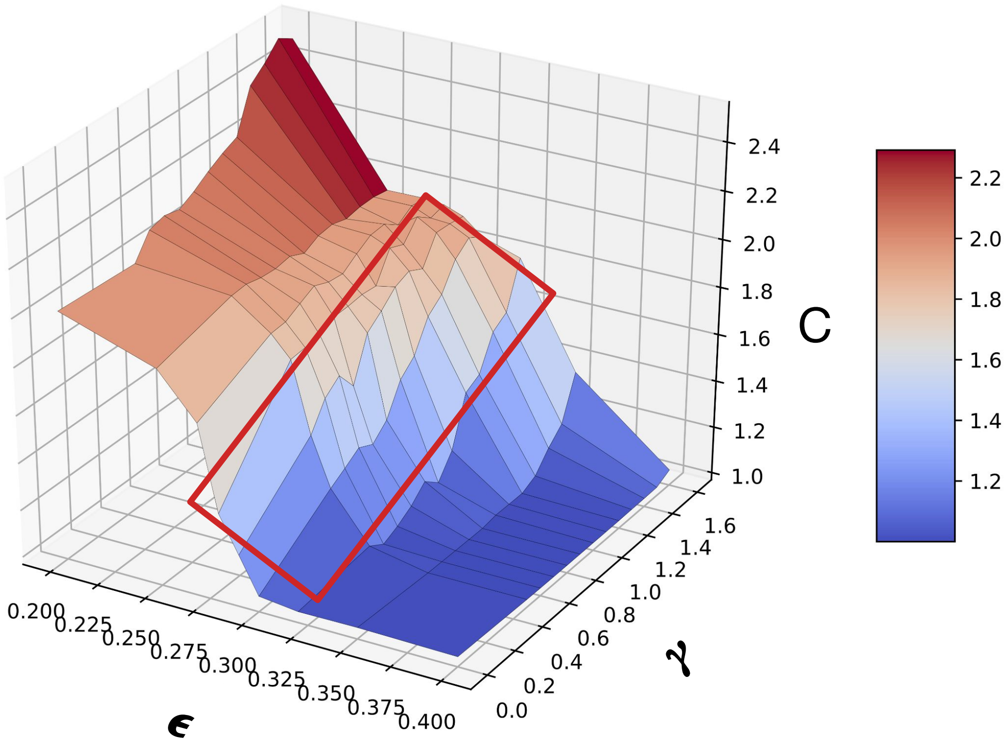
* results are always averaged over 500 independent Monte Carlo simulations

ALGORITHMIC BIAS: results

Effective Number of Clusters*

In the final opinion distribution

$$C = \frac{\left(\sum_i c_i\right)^2}{\sum_i c_i^2}$$



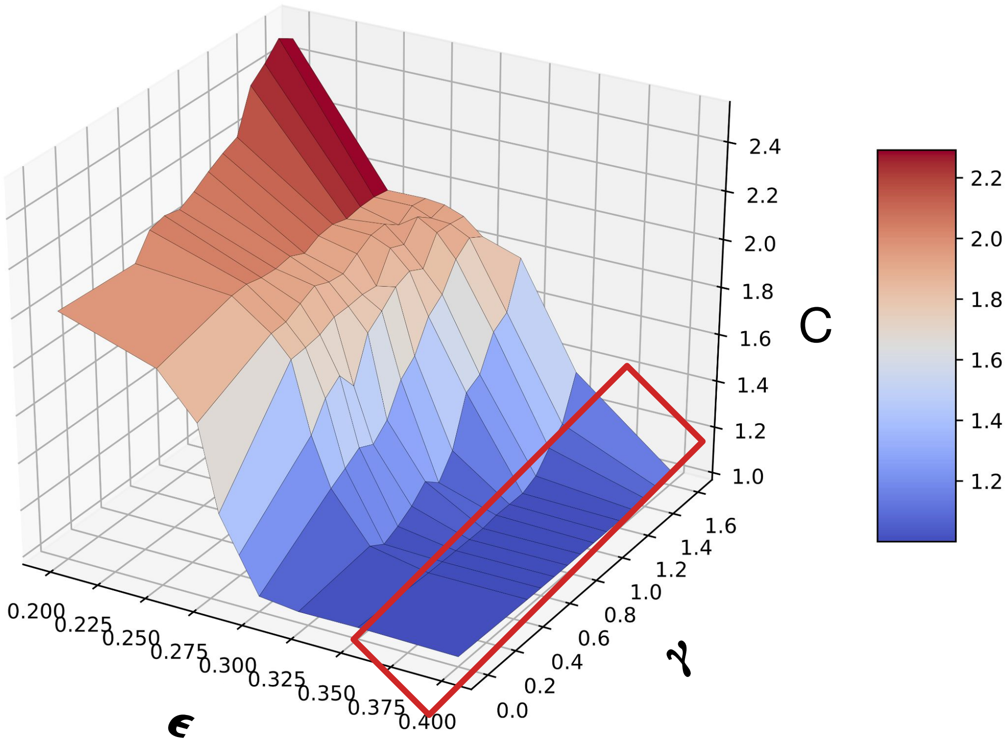
* results are always averaged over 500 independent Monte Carlo simulations

ALGORITHMIC BIAS: results

Effective Number of Clusters*

In the final opinion distribution

$$C = \frac{\left(\sum_i c_i\right)^2}{\sum_i c_i^2}$$

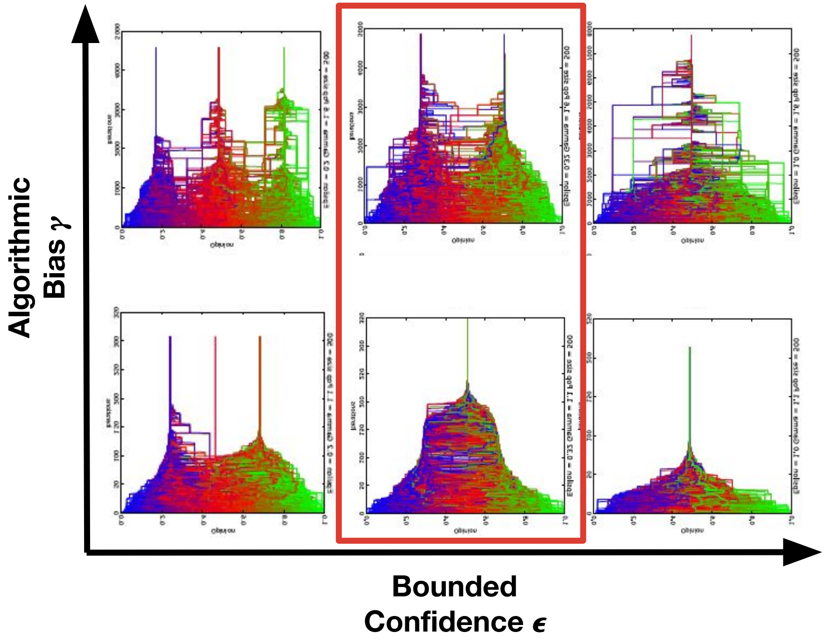


* results are always averaged over 500 independent Monte Carlo simulations

ALGORITHMIC BIAS: results

A “semantic” algorithmic bias exacerbates polarization and fragmentation in the long term:

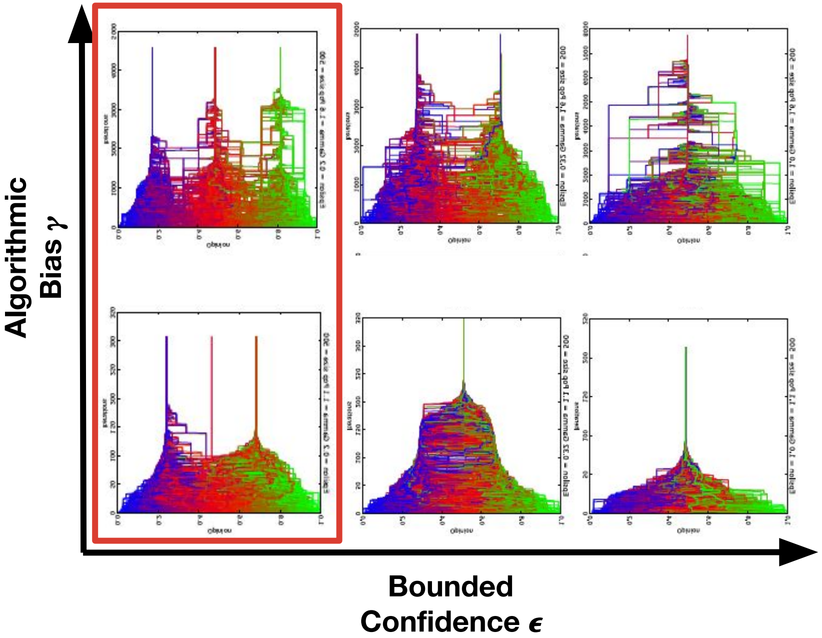
- **More opinion clusters**: the effective number of clusters increases with gamma (for a fixed epsilon)



ALGORITHMIC BIAS: results

A “semantic” algorithmic bias exacerbates polarization and fragmentation in the long term:

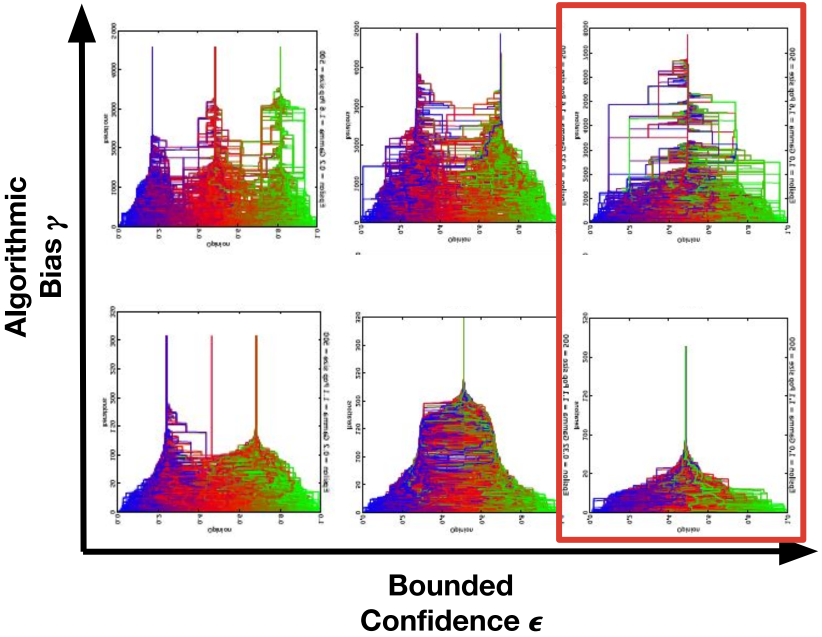
- **More opinion clusters**
- **Further apart opinion clusters**: the average pairwise distance increases with gamma (for a fixed epsilon)



ALGORITHMIC BIAS: results

A "semantic" algorithmic bias exacerbates polarization and fragmentation in the long term:

- **More opinion clusters**
- **Further apart opinion clusters**
- **Longer time necessary to reach consensus**: the number of interactions necessary to reach consensus increases with gamma (for a fixed epsilon)



ALGORITHMIC BIAS: the results

Is this the whole story?

ALGORITHMIC BIAS: the results

Is this the whole story?

NO!

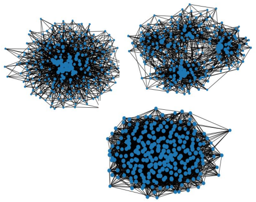
ALGORITHMIC BIAS AND…

Extensions of Sirbu et al. incorporating:

External Effects

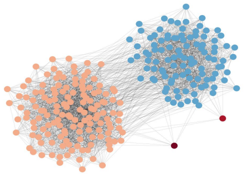
Algorithmic Bias counters homogenization due to propaganda and favors opinion-based clustering

Network effects



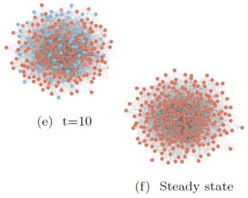
The sparser the network the easier is to have a fragmented population even with lower personalization

Co-evolving topology



Echo chamber formation is slowed down by RecSys under this model

Peer-pressure



Peer pressure mechanisms within group interaction enhance consensus: RecSys don't break strong communities

Pansanella, Valentina, et al. (2023). "Mass Media Impact on Opinion Evolution in Biased Digital Environments: a Bounded Confidence Model." Sci. Rep..
Pansanella, V., et al. (2021). From mean-field to complex topologies: network effects on the algorithmic bias model. Proceedings of Complex Networks XI
Pansanella, V., et al. (2022). Modeling Algorithmic Bias: Simplicial Complexes and Evolving Network Topologies. Applied Network Science

…Two truths and a lie…

Which statement is NOT a correct outcome of increasing algorithmic bias (γ)?

A. It **increases** the number of **opinion clusters** in the population

B. It **accelerates consensus** by connecting like-minded individuals more effectively

C. It **increases** the average **opinion distance** between groups

WHAT'S NEXT?

SIMULATING SOCIETIES

Creating more realistic simulations (leveraging e.g. LLMs) allowing us evaluating other outcomes and other models

Qualify as vetted researchers and directly study VLOPs with controlled experiments

ENFORCING THE DSA

SIMULATING SOCIETIES: YSocial

Rossetti et al., Arxiv (2024)

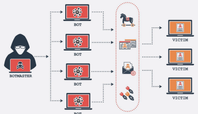
A **replica** of an online social platform that allows for the design of **realistic social simulations** in a controlled environment

LLM-Agents	**Recsys**	**Analyst**
Generate (and discuss) content/news	Filter Interactions Alg. Bias	Studies emerging behaviors

Scenarios

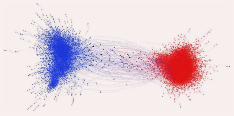
D/Misinformation

Alg. Bias

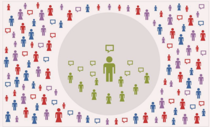


Bots

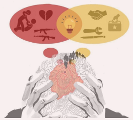
Observable Effects



Echo Chambers



Filter Bubbles



Radicalization

Goals:

- Support for the definition of data-driven scenarios and simulations
- Understand of the impact of recommendation systems on user behavior
- Study online debate outcomes

TAKE HOME MESSAGES

TAKE HOME MESSAGES

- Recommenders and feedback loop investigated within the social media environment.

TAKE HOME MESSAGES

- Recommenders and feedback loop investigated within the social media environment.
- The POV is always the user or the system, rarely item or model.

TAKE HOME MESSAGES

- Recommenders and feedback loop investigated within the social media environment.
- The POV is always the user or the system, rarely item or model.
- Great focus on *filter bubbles*, *echo chambers* and *polarization/radicalization*, with non conclusive results. **Is it time to prioritize other outcomes?**

TAKE HOME MESSAGES

- Recommenders and feedback loop investigated within the social media environment.
- The POV is always the user or the system, rarely item or model.
- Great focus on *filter bubbles*, *echo chambers* and *polarization/radicalization*, with non conclusive results. **Is it time to prioritize other outcomes?**
- Why non conclusive? A lot of observational studies, results depend on time and other contextual elements, **lack of generalizable and universal results.**

TAKE HOME MESSAGES

- Recommenders and feedback loop investigated within the social media environment.
- The POV is always the user or the system, rarely item or model.
- Great focus on *filter bubbles*, *echo chambers* and *polarization/radicalization*, with non conclusive results. **Is it time to prioritize other outcomes?**
- Why non conclusive? A lot of observational studies, results depend on time and other contextual elements, **lack of generalizable and universal results.**
- Controlled studies are almost impossible to perform for external researchers, **what do platforms know that we ignore?**

References

[Section 3] Pappalardo, L., Ferragina, E., Citraro, S., Cornacchia, G., Nanni, M., Rossetti, G., ... & Pedreschi, D. (2024). **A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions.** arXiv preprint arXiv:2407.01630.

- Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). **Algorithmic amplification of politics on Twitter.** Proceedings of the national academy of sciences, 119(1), e2025334119.
- Kertesz, J., Sirbu, A., Gianotti, F., & Pedreschi, D. (2019, April). **Algorithmic bias amplifies opinion polarization: A bounded confidence model.** In StatPhys 27 Main Conference.
- Rossetti, G., Stella, M., Cazabet, R., Abramski, K., Cau, E., Citraro, S., ... & Pansanella, V. (2024). **Y social: an llm-powered social media digital twin.** arXiv preprint arXiv:2408.00818.
- Pansanella, V., Sîrbu, A., Kertesz, J., & Rossetti, G. (2023). **Mass media impact on opinion evolution in biased digital environments: a bounded confidence model.** Scientific Reports, 13(1), 14600.
- Pansanella, V., Rossetti, G., & Milli, L. (2021, November). **From mean-field to complex topologies: network effects on the algorithmic bias model.** In International Conference on Complex Networks and Their Applications (pp. 329-340). Cham: Springer International Publishing.

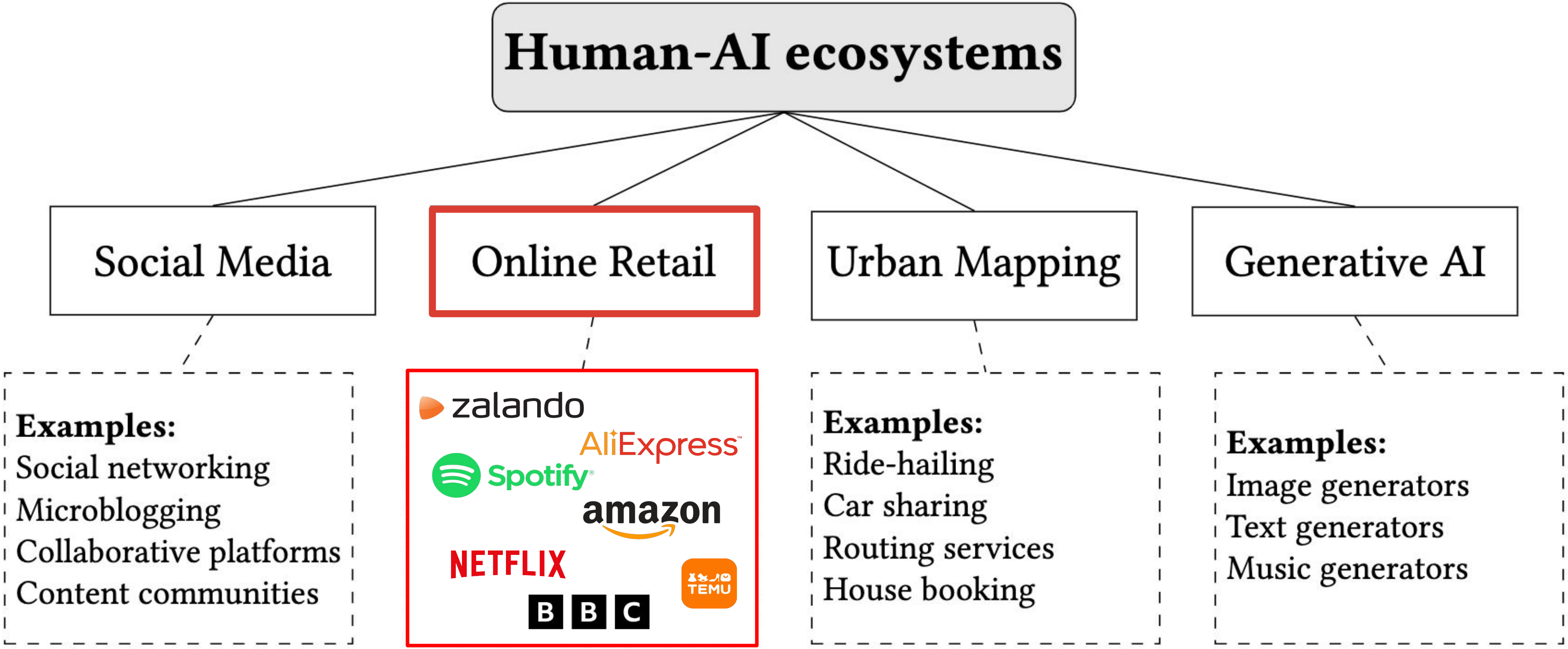
Limitations of the study

Huszár et al., PNAS 2021

The experiment violates SUTVA (Stable Unit Treatment Value Assumption):

- the control group is not isolated from indirect effects of personalization
- the experiment cannot provide unbiased estimates of causal quantities

The study just present findings based on simple comparison of measurements between the treatment and control groups


Human-AI ecosystems
Social Media
Online Retail
Urban Mapping
Generative AI
Examples:
Social networking
Microblogging
Collaborative platforms
Content communities
zalando
AliExpress
Spotify
amazon
NETFLIX
TEMU
BBC
Examples:
Ride-hailing
Car sharing
Routing services
House booking
Examples:
Image generators
Text generators
Music generators

ONLINE RETAIL

Recommender systems to alleviate choice-overload of consumers

Helping individuals to find the most **appropriate products** or discover **interesting content**.

Shopping addiction

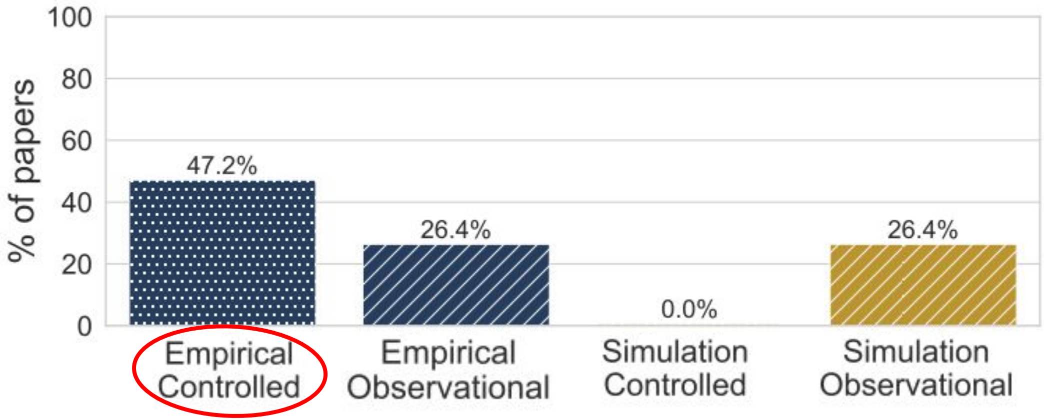
- Online retailers are increasingly using *psychological techniques* to keep shoppers spending money

- >1,000 people in Switzerland grouped into categories of shoppers:
  - **3% addicted** to online shopping
  - **11% at risk**
    - "I think about shopping/buying things all the time"
    - "I shop/buy things in order to change my mood"

E. Marris, The science of shopping addiction: what makes people buy loads of stuff? Nature 639, 26-28 (2025)
Augsburger et al., The concept of buying-shopping disorder, J. Behav. Addict. 9, 808–817 (2020).

Experiments

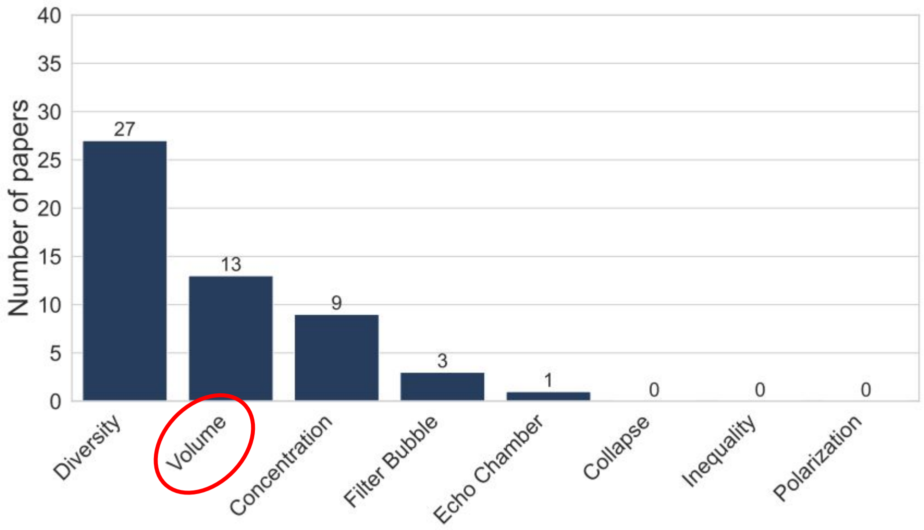
A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630



- Predominance of empirical studies over simulations
- **Larger amount of empirical controlled experiments** with respect to other ecosystems (either large analysis conducted inside organisations or small experiments conceived in collaborations with universities)

Main outcomes

A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630

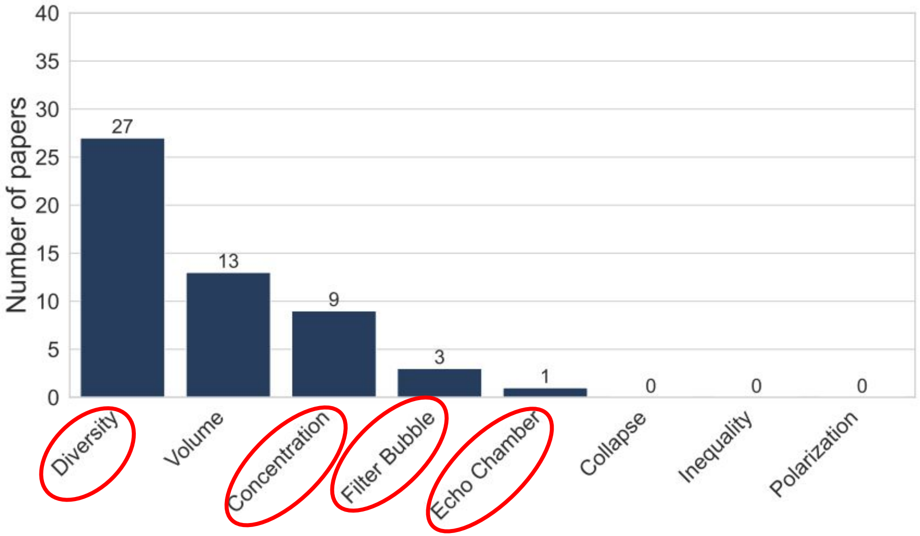


At first it was **volume** - monolithic agreement: RS increase volumes significantly (it is not just a matter of choice overload, they indeed push individuals to buy more)

- Volume of **sales**, **clicks**, **views**, **ratings**, **retention time**…

Main outcomes

A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630



Anderson's hypothesis: "main effect of recommenders will be to help people move from the world of hits to the world of niches" [1] (long-tail effect)

Diversity hypothesis: recommenders will reinforce the world of hits making niches disappear (rich-get-richer effect)

[1] Anderson, Chris, "The Long Tail: why The Future of Business Is Selling Less Of More", Hyperion Books (2008).

Online Retail		Empirical		Simulation	
		Observational	Controlled	Observational	Controlled
Individual	**Filter Bubble**	[116]	[27]	[117]	
	Radicalization				
Model	**Collapse**				
Systemic	**Concentration**	[55, 75]	[93, 94, 155]	[54, 56, 105, 152]	
	Echo Chamber	[60]			
	Inequality				
	Polarization				
Individual Item Systemic	**Diversity**	individual: [8, 60, 116], systemic: [26, 122]	individual: [74, 93–95, 97, 155], item: [111], systemic: [44, 74, 92, 110, 111, 155]	individual: [9, 54, 56, 117], item: [71], systemic: [9, 24, 71, 105]	
	Volume	individual: [55, 75, 116], item: [26, 122]	individual: [27, 44, 74, 93, 95], item: [92, 94], systemic: [8]		

Selected study:

[94] D. Lee and K. Hosanagar, *How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment.* Information Systems Research 30, 1 (2019)

Empirical Studies

How do recommender systems affect sales diversity? A cross-category investigation via randomized field experiment

Lee and Hosanagar et al., Information Systems Research, 2019

Type: Empirical controlled

VLOP: Canadian online retail platform

Outcomes: aggregate diversity loss

Experimental Setup

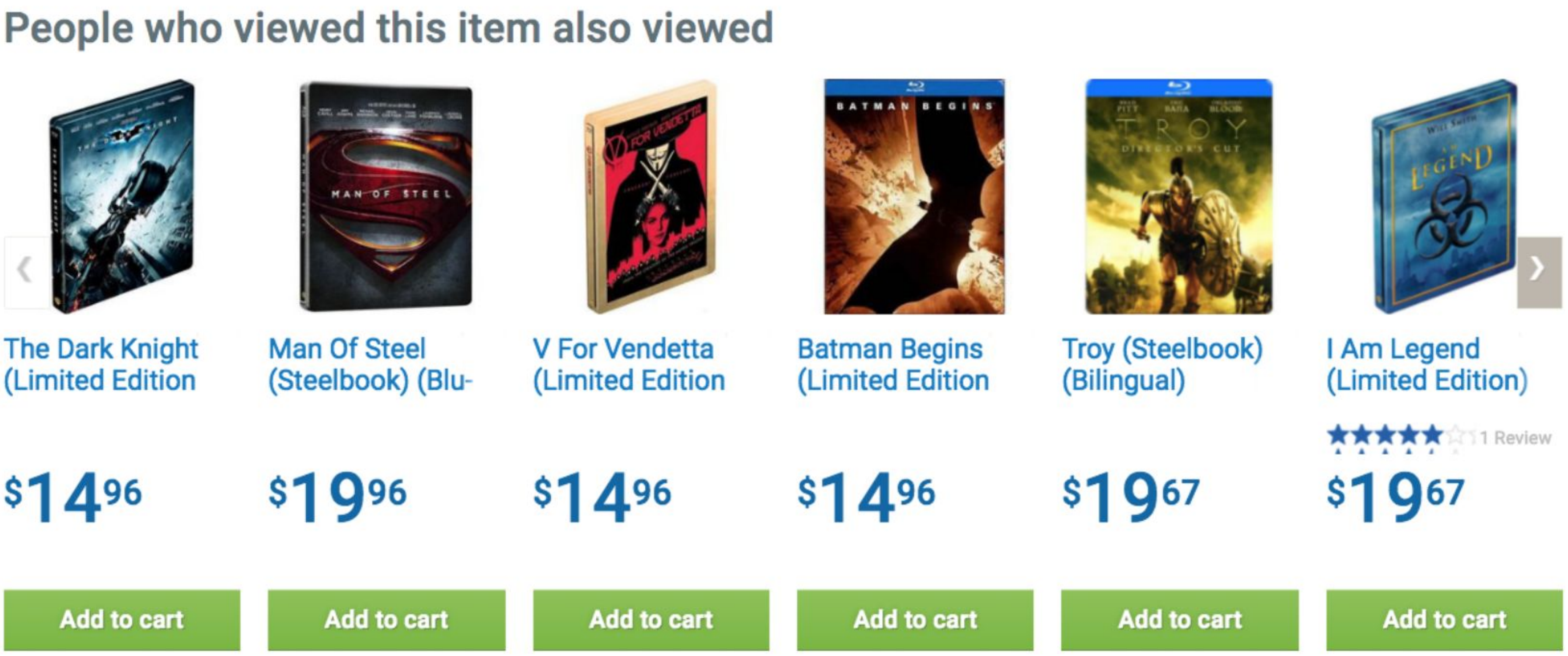
Consumers on a Canadian online retail website

- **Two weeks**: August 8 to 22, 2013
- **A/B/n testing platform** that tracks users' behaviour during the experiment
- View and purchase logs are collected:
  - views and purchases of 1M users
  - 82K Stock-Keeping-Units (products)
  - 2.8M rows of individual-level data

Experimental Setup

- **Control group**: no recommendation
- **Treatment** (20% of users):
    - **Treatment group 1**: visualizes recommendations from a view-based collaborative filtering (VBCF)
        - "People who viewed this item also viewed"
    - **Treatment group 2**: visualizes recommendations from a purchase-based collaborative filtering (PBCF)
        - "People who purchased this item also purchased"

Experimental Setup



Quiz

Which kind of recommender is this?

user-based CF

item-based CF ✅

Experimental Setup

The recommender takes as input:

- the **focal item** (the product a user is viewing)
- the user's **past purchases**
  - data about 60 days before the experimentation starts
  - recommender retrained every 3 days

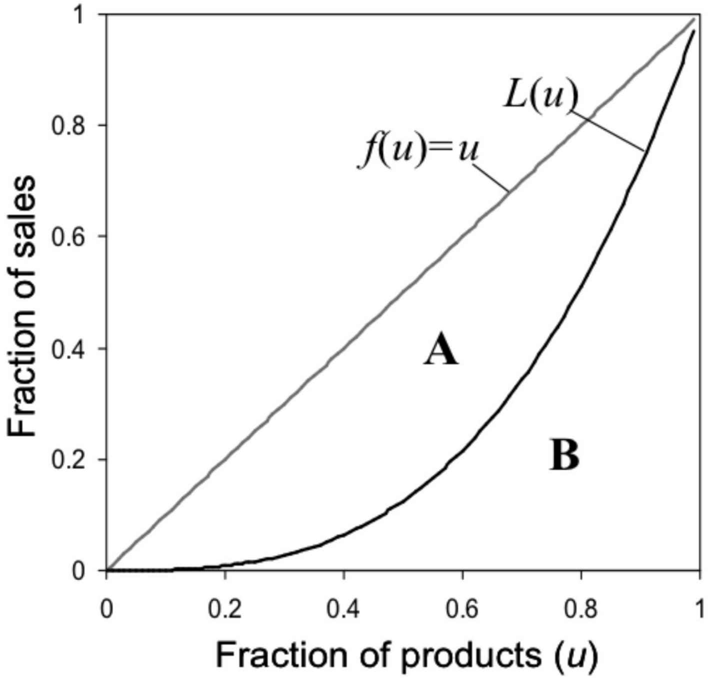
The top N candidate products that are not yet purchased/viewed by the consumer are recommended

Sales diversity

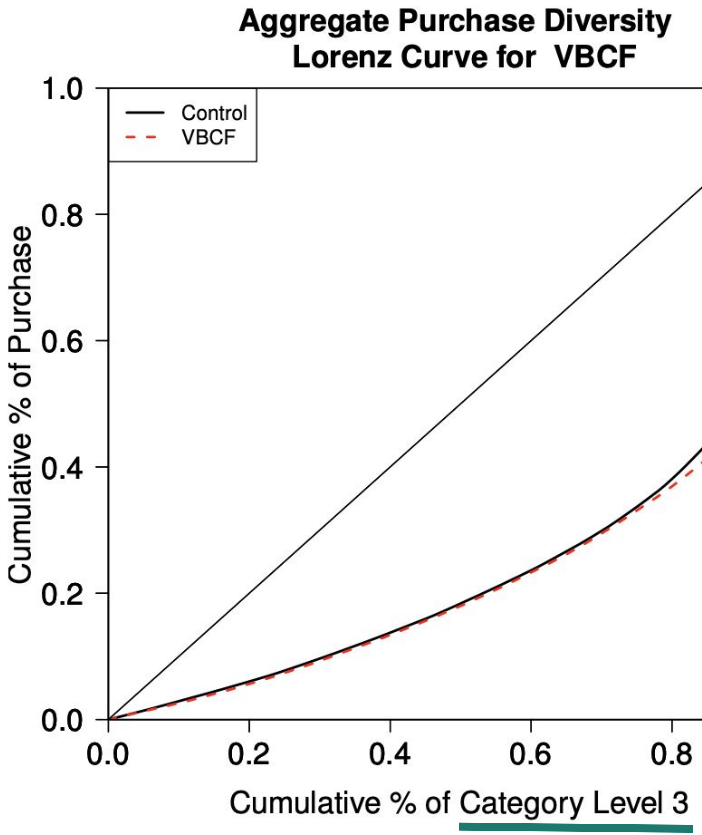
Gini coefficient:

- **0** is the least amount of concentration (**highest diversity**, equal sales)
- **1** represents the highest amount of concentration (**lowest diversity**, a few broad-appeal blockbuster items account for most of the sales)

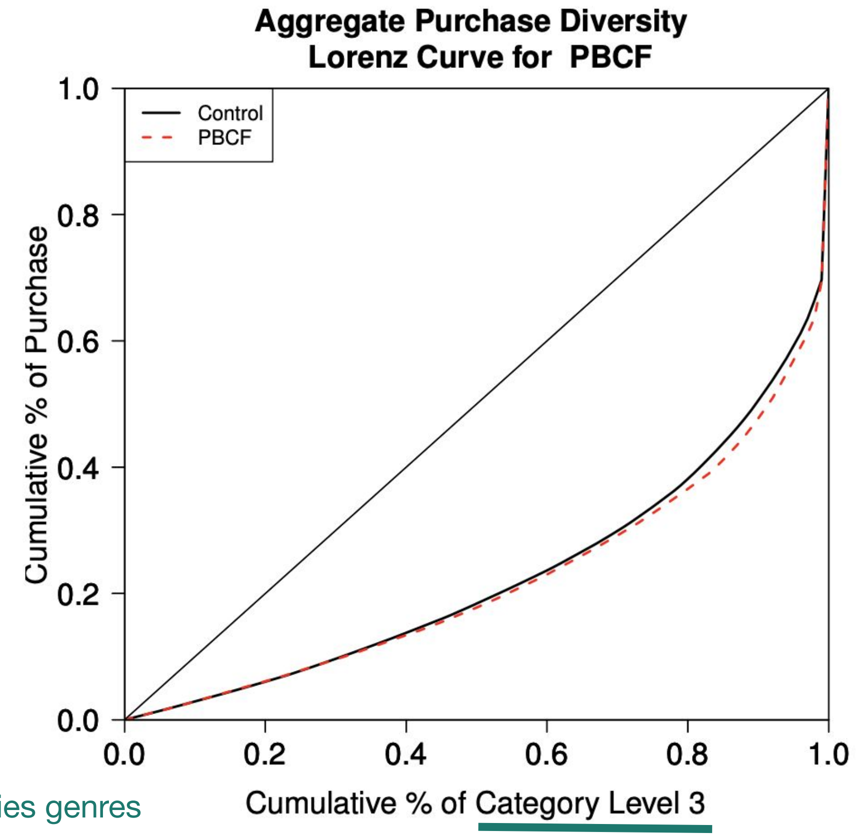
$$G = \frac{A}{A+B}$$



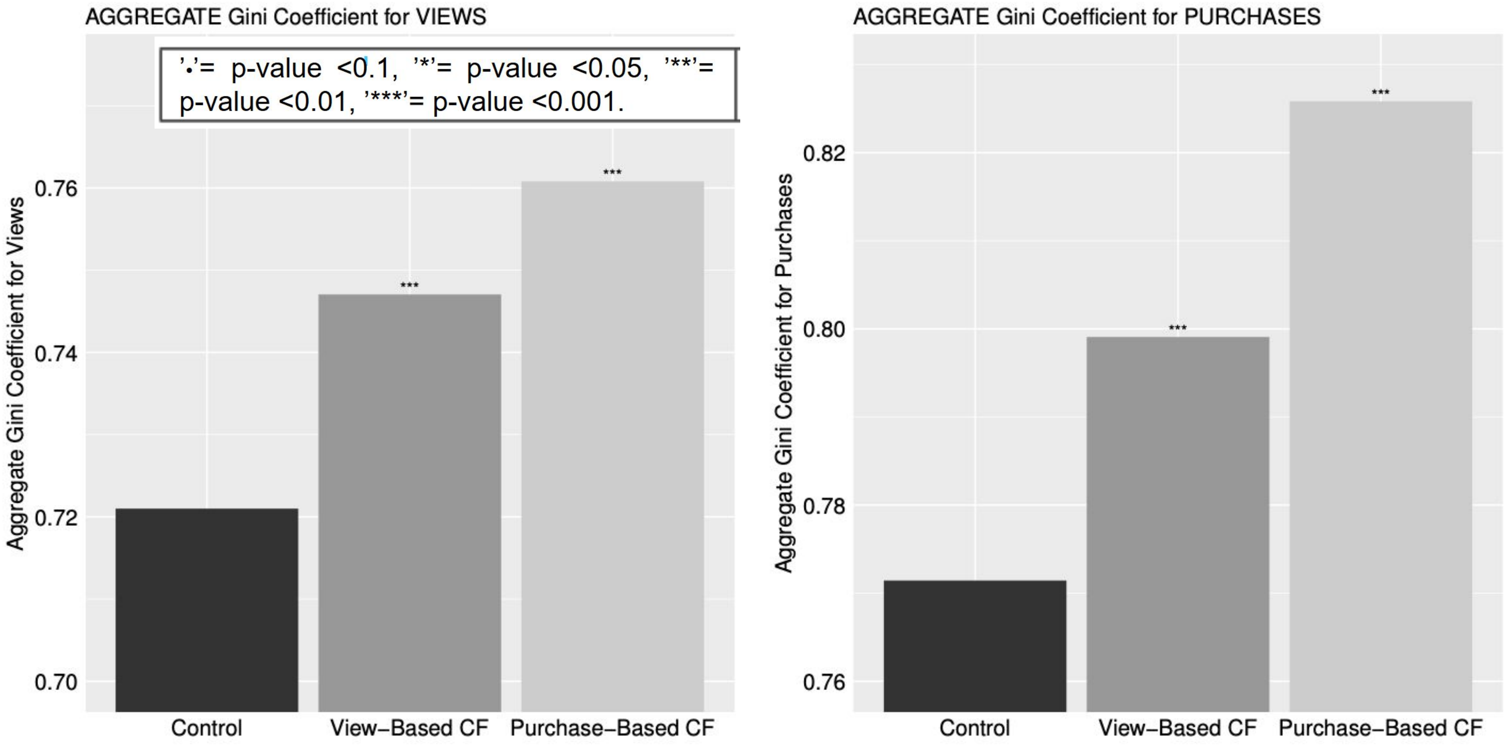
Aggregated diversity



e.g. Movies genres

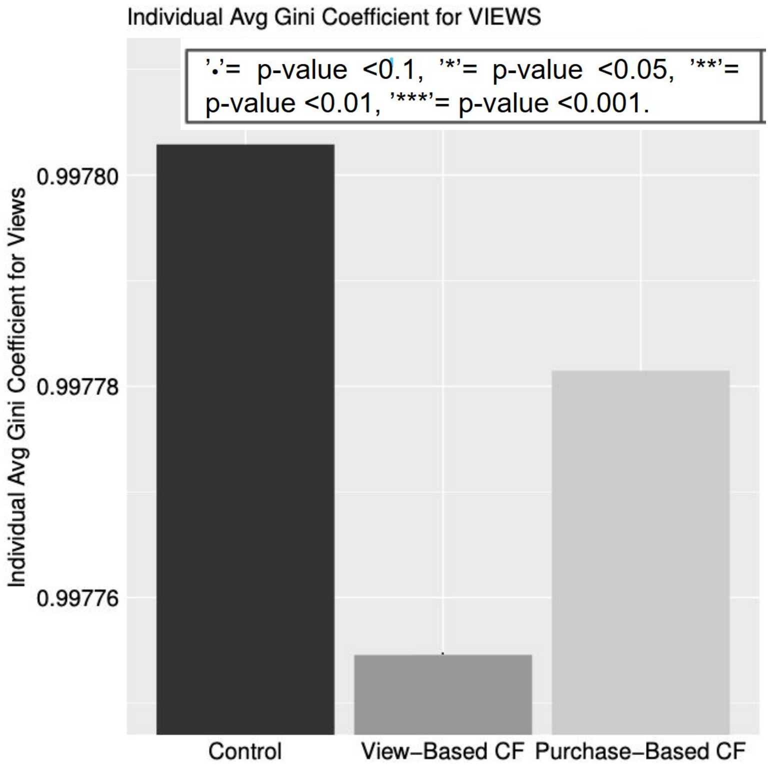


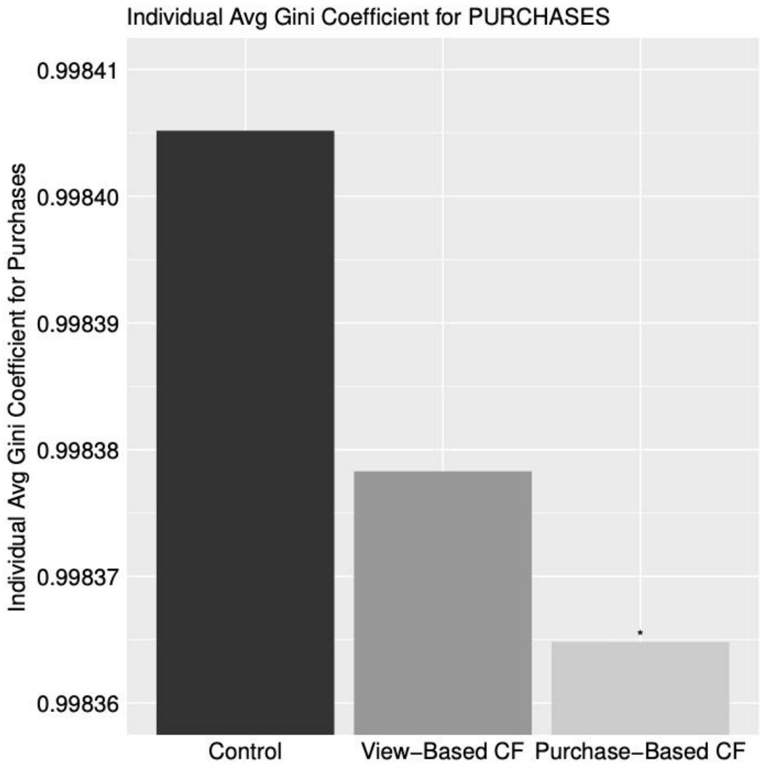
Aggregated diversity:



- both VBCF and PBCF are causing consumers to view and purchase less variety of products

Individual diversity





- no concentration bias (Gini is lower for CF but not significantly)

Notes on Gini: small differences can convey large consequences

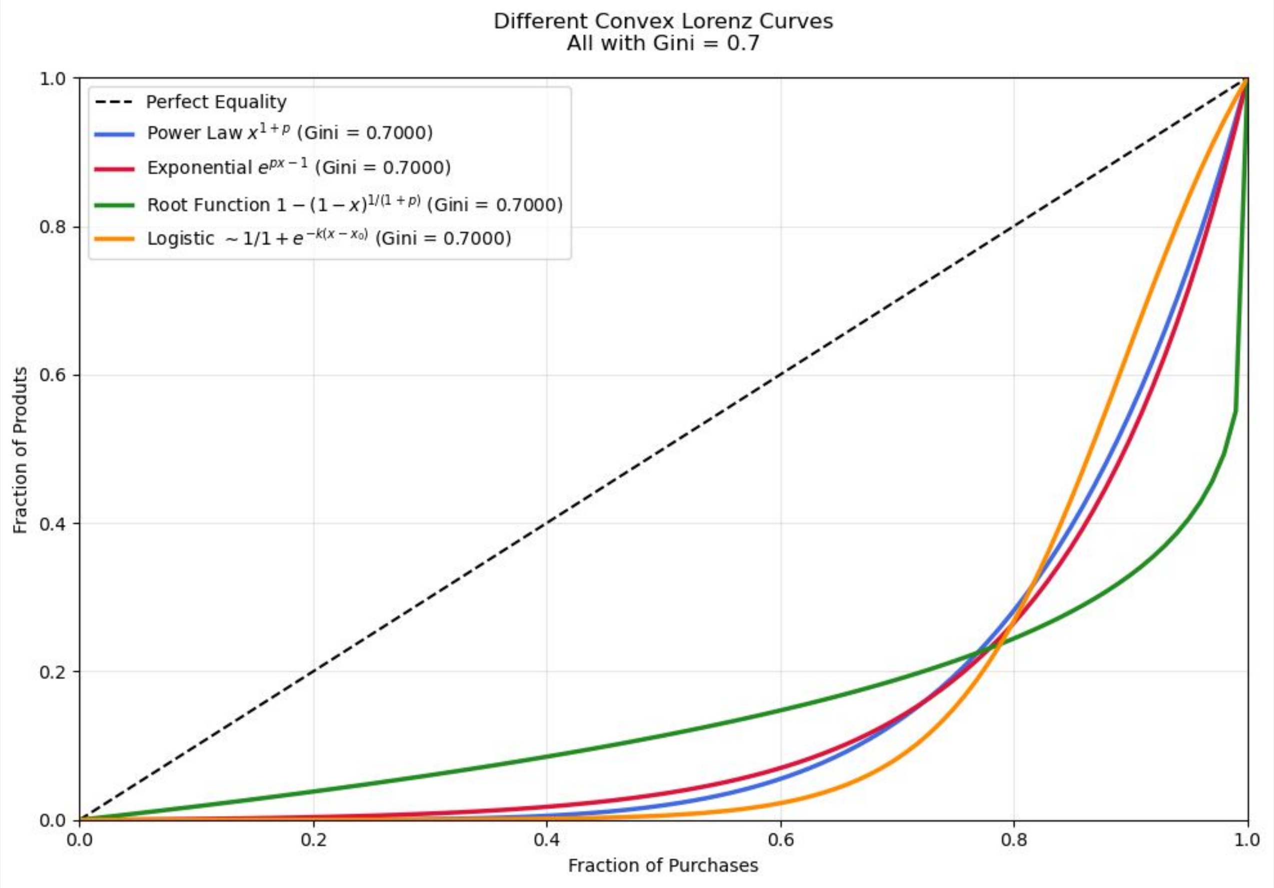
'•'= p-value <0.1, '*'= p-value <0.05, '**'= p-value <0.01, '***'= p-value <0.001.	Control	VBCF	PBCF
Aggregate View	0.720997	0.747055***	0.760807***
Aggregate Purchase	0.771437	0.799075***	0.825829***
Individual Avg View	0.997803	0.997755•	0.997781
Individual Avg Purchase	0.998405	0.998378	0.998365*

Aggregate PBCF: 0.825829 - 0.771437 = **0.054**

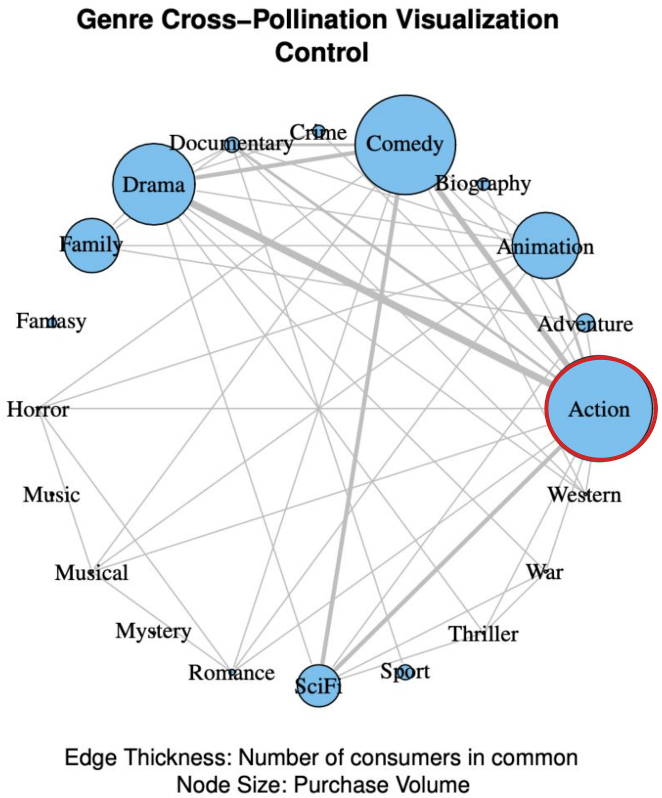
*"Increasing the Gini coefficient of DVD rentals by **0.0029** translates to increasing the market share of the top 1% of DVDs by 1.96% and the market share of the top 10% of DVDs by 0.58%. At the same time, the market share of the bottom 1% of DVDs is reduced by 21.29%, while the market share of the bottom 10% of DVDs is reduced by 5.28%."*

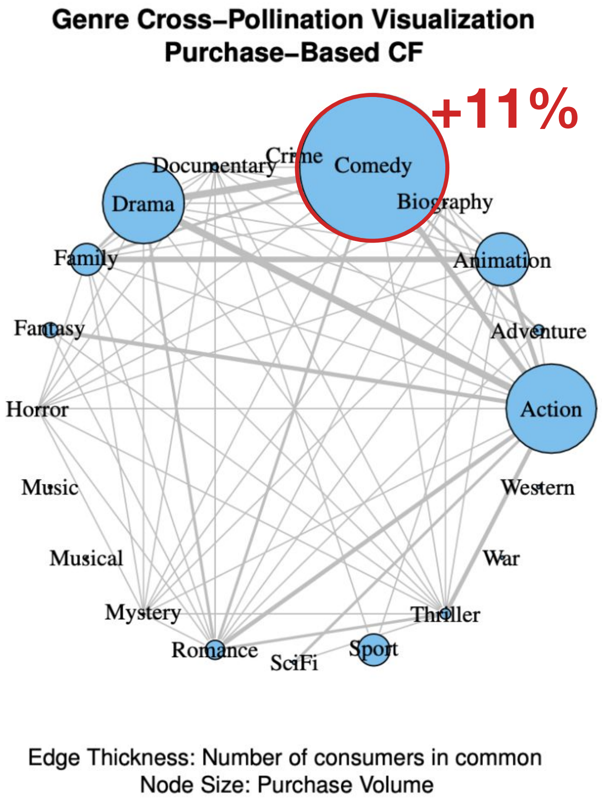
Tan et al., 2017, 'Is Tom Cruise Threatened? An Empirical Study of the Impact of Product Variety on Demand concentration'. Information Systems Research 28(3), 643–660.

Notes on Gini: Different Lorenz curves can have identical Gini value

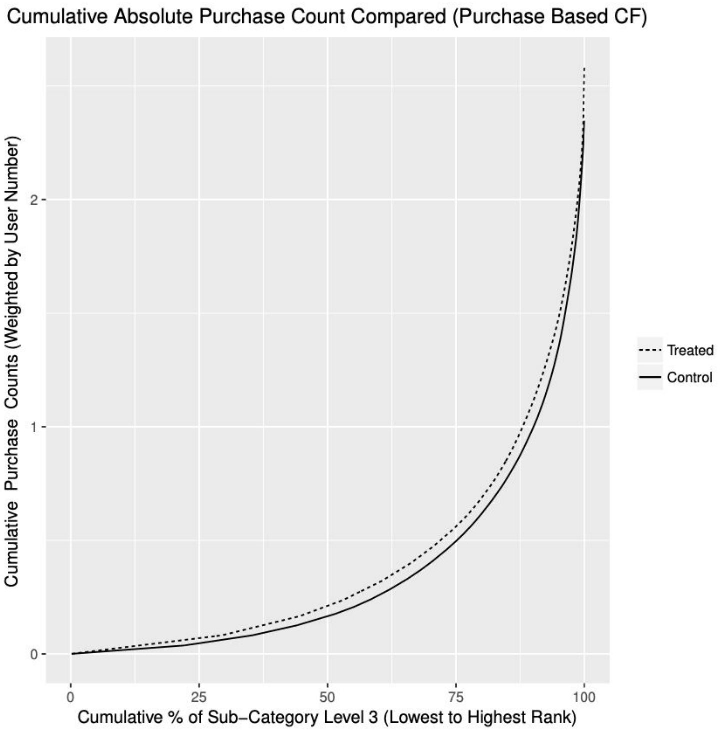


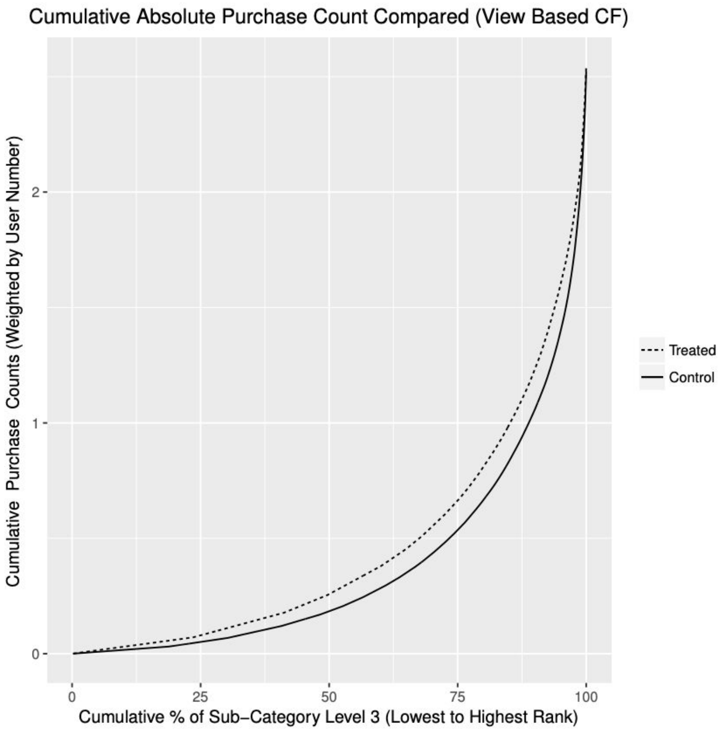
Co-Purchase networks (to understand more)





Niche products





All products are sold more, regardless of their popularity!

In summary

- Consumers cross-purchase more;
- At the same time, their explorations are highly correlated due to the nature of CF;
- Therefore, **the market share for the top-selling products keeps increasing**, creating a *rich-get-richer* bias;
  - However, nich items do not necessarily lose as CF increases absolute sales volumes for all items.

Is this the whole story?

Diversity	Individual	Systemic
Increased	[116] [93–95, 97], [155] (only views)	[26, 122] [44, 74, 110], [92] (only views)
Reduced	[8, 60], [155] (only purchases), [27, 74]	[56] [93, 94, 155], [111] (only without cold start)

- Empirical Observational
- Empirical Controlled

M.C. Escher, *Relativity*, 1953. Lithograph.

The Engagement-Diversity Connection: Evidence From a Field Experiment on Spotify

DAVID HOLTZ, MIT Sloan School of Management, USA
BEN CARTERETTE, PRAVEEN CHANDAR, ZAHRA NAZARI, and HENRIETTE CRAMER, Spotify, USA
SINAN ARAL, MIT Sloan School of Management, USA

We present results from a large-scale, randomized field experiment on Spotify testing the effect of personalized recommendations on consumption diversity. In the experiment, both control and treatment users were given podcast recommendations, with the sole aim of increasing podcast consumption. However, the recommendations provided to treatment users were personalized based on their music listening history, whereas control users were recommended the most popular podcasts among users in their demographic group. Consistent with previous studies, we find that the treatment increased the average number of podcast streams per user. However, we also find the treatment decreased the average individual-level diversity of podcast streams and increased the aggregate diversity of podcast streams, indicating that personalized recommendations have the potential to create consumption patterns that are homogenous within users and diverse across users. Our results provide evidence of an "engagement-diversity trade-off" when optimizing solely for increased consumption: while personalized recommendations increase user engagement, they also affect the diversity of content that users consume. This shift in consumption diversity can affect user retention and lifetime value, and also impact the optimal strategy for content producers. Additional analyses suggest that exposure to personalized recommendations can also affect the content that users consume organically. We believe these findings highlight the need for both academics and practitioners to continue investing in approaches to personalization that explicitly take into account the diversity of content recommended to users.

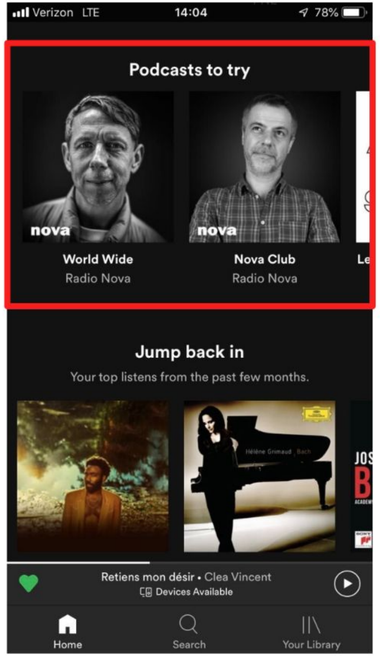
CCS Concepts: • **Information systems**; • **Applied computing** → *Electronic commerce*; • **Computing methodologies** → Machine learning;

The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify

D. Holtz et al., Proceedings of the 21st ACM Conference on Economics and Computation, New York, 2020.

Type:	Empirical Controlled
VLOP:	Spotify
Outcomes:	Increased aggregate diversity, decreased individual diversity

A different experiment



- **Podcast streamings** of 800K premium users on Spotify, across 17 countries (US, IT, AR..);
- **Two weeks**: April 18 to May 2, 2019;

- **Control**: Recommended the 10 most popular podcasts among users in their demographic group;
- **Treatment**: Recommended 10 podcasts based on an NN classifier fed with music listening history and demographic info [1];
  - No retraining
  - Stop recommending once a user streams their first podcast.

[1] Nazari, Zahra, et al., *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Similar research questions

Assess potential **impacts on streams diversity**:

- At the **individual level**, through the average Shannon entropy of individuals:

$$h_i = -\sum_{c \in C} s_{ci} \ln(s_{ci})$$

  where s_ci is the fraction of streams of user i from category c;

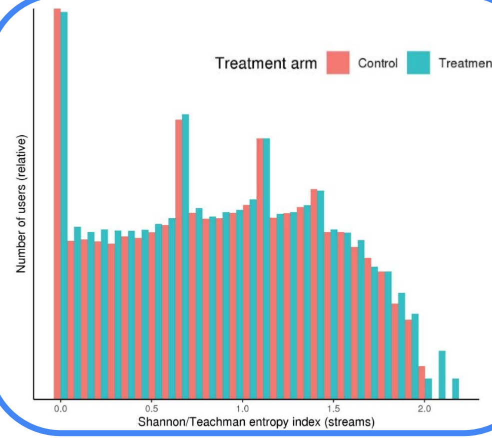
- At the **aggregate level**, by the intragroup diversity:

$$ID = \frac{1}{n_c} \sum_{j=1}^{n_c} \left[1 - \cos\left(\Gamma_j, \bar{\Gamma}\right)\right]^2$$

  where n_c is the number of categories and Gamma_j is the vector interactions between user j and category c.

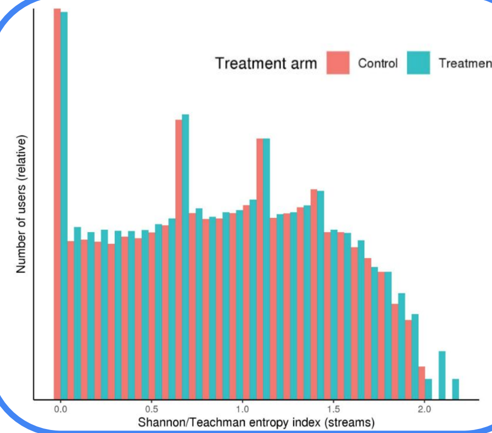
A different outcome

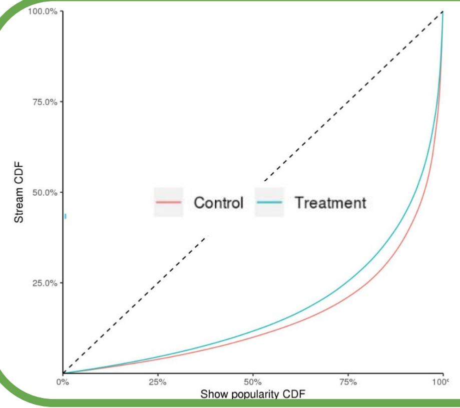
- “Recommender systems can create an engagement-diversity trade-off for firms when optimizing solely for engagement”
    - **Increase the amount of content** users consume by 28%
    - **Increase the homogeneity of content** that individual users consume: average Shannon entropy of podcast streams 11% lower in the treatment group.



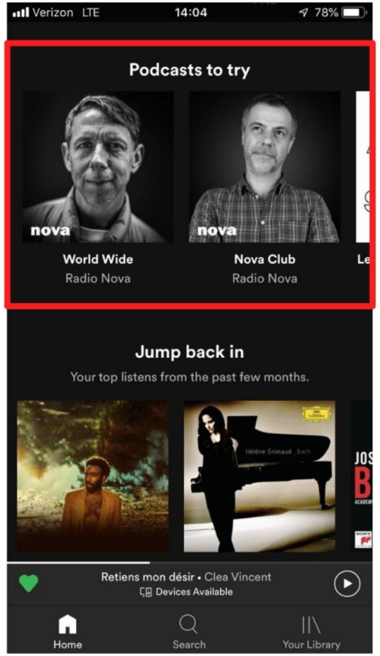
A different outcome

- “Recommender systems can create an engagement-diversity trade-off for firms when optimizing solely for engagement”
  - **Increase the amount of content** users consume by 28%
  - **Increase the homogeneity of content** that individual users consume: average Shannon entropy of podcast streams 11% lower in the treatment group.
  - **Increase the dissimilarity between** what **different users** consume: the intragroup diversity for podcast streams is increased by 5.96%.
- Exposure to personalized recommendations affects recommended consumption and “organic” consumption





Quiz: which factors could explain the discrepancy?

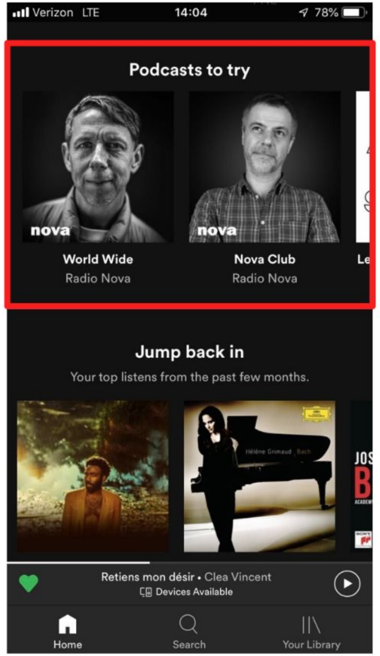


- **Podcast streamings** of 800K premium users on Spotify, across 17 countries (US, IT, AR..);
- **Two weeks**: April 18 to May 2, 2019;

- **Control**: Recommended the 10 most popular podcasts among users in their demographic group;
- **Treatment**: Recommended 10 podcasts based on an NN classifier fed with music listening history and demographic info [1];
  - No retraining
  - Stop recommending once a user streams their first podcast.

[1] Nazari, Zahra, et al., *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

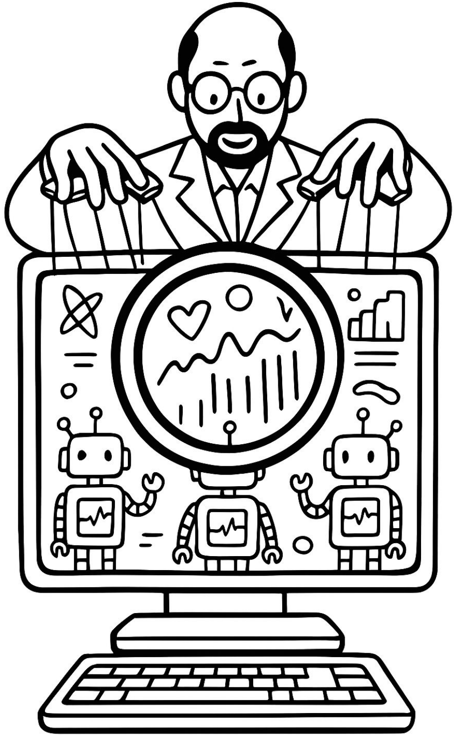
A different experiment



- **Podcast streamings** of 800K premium users on Spotify, across 17 countries (US, IT, AR..);
- **Two weeks**: April 18 to May 2, 2019;

- **Control**: Recommended the 10 most popular podcasts among users in their demographic group;
- **Treatment**: Recommended 10 podcasts based on an NN classifier fed with music listening history and demographic info [1];
    - No retraining
    - Stop recommending once a user streams their first podcast.

[1] Nazari, Zahra, et al., *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.



Simulation Studies

Experimental Setup in simulations

What type of consumption?
Which items?

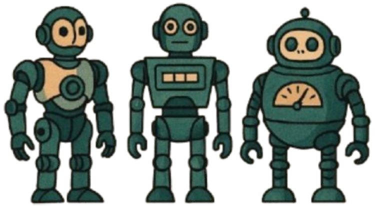
Experimental Setup in simulations

What type of consumers?
heterogeneity in tastes, habits, specific preferences…

Experimental Setup in simulations

How do they choose specific items?
(choice model)

Experimental Setup in simulations



which family of recommenders? which hyperparameters?

Recommender Systems Effect on the Evolution of Users’ Choices Distribution

Naieme Hazrati, Francesco Ricci, Information Processing and management, International journal, 2022

Type: Simulation observational

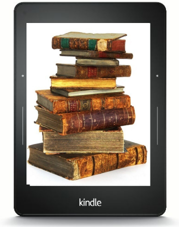
VLOP: Amazon

Outcomes: aggregate diversity

Dataset

Time-stamped rating log data from three Amazon collections [1]:

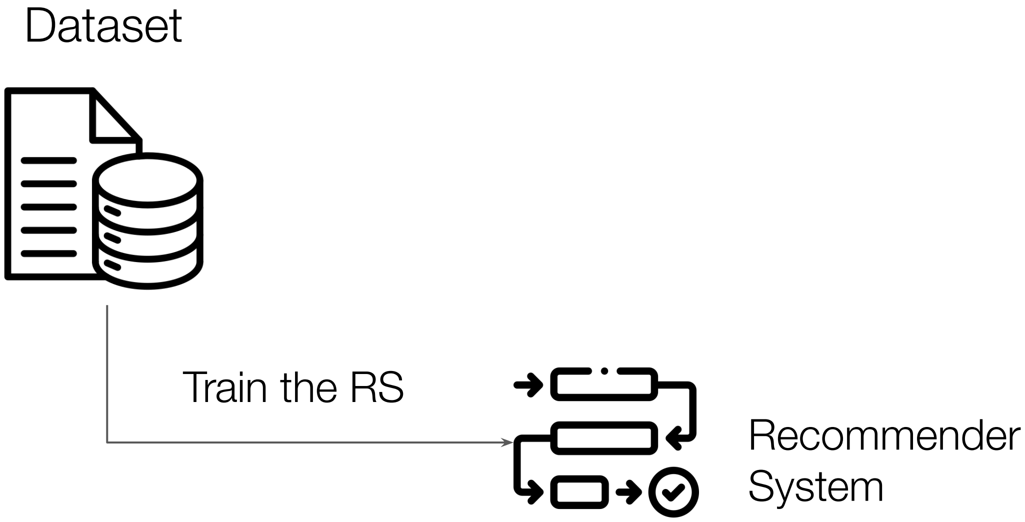
- **Apps** *(42+10 months)*
  - 5K users
  - 24K items
  - 154K interactions
- **Games** *(169+10 months)*
  - 2K users
  - 20K items
  - 80K interactions
- **Kindle** *(134+10 months)*
  - 3K users
  - 16K items
  - 28K interactions



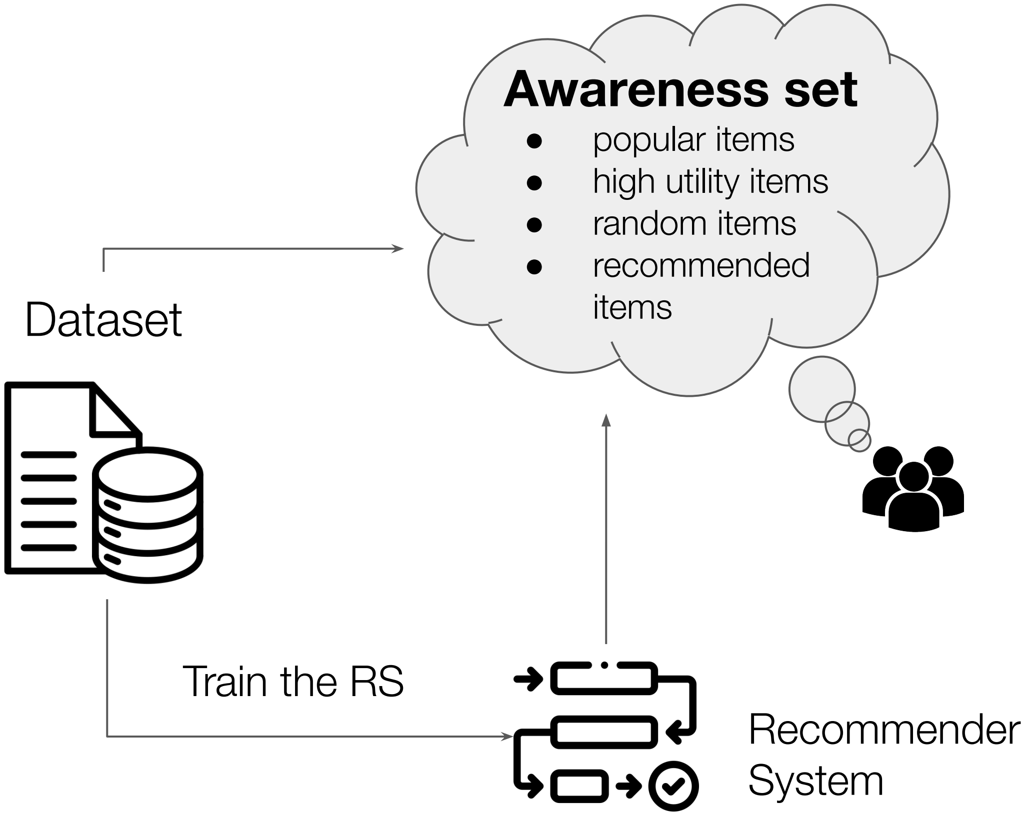
Note: users do not make repeated choices for a single item

[1] https://amazon-reviews-2023.github.io/

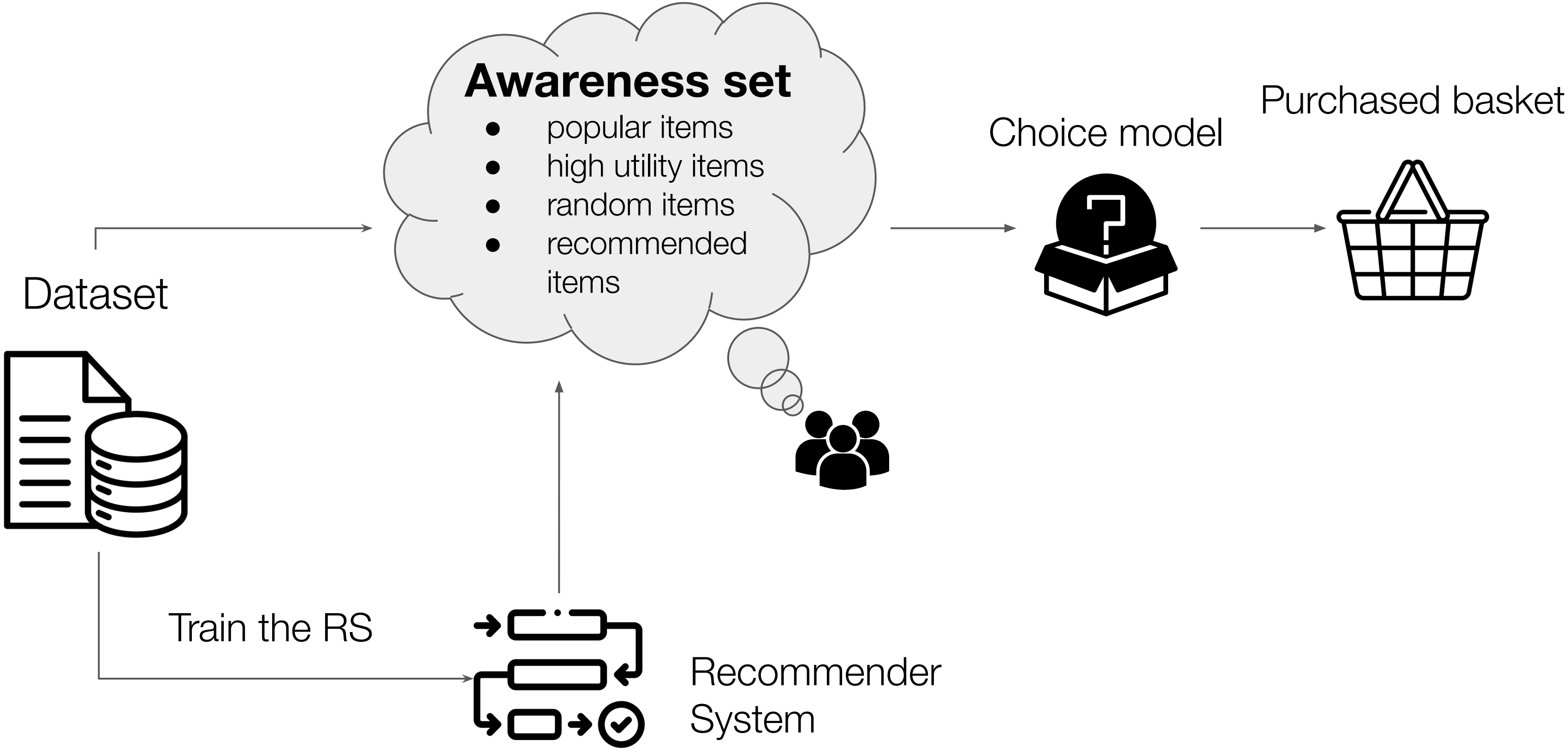
Simulated purchase



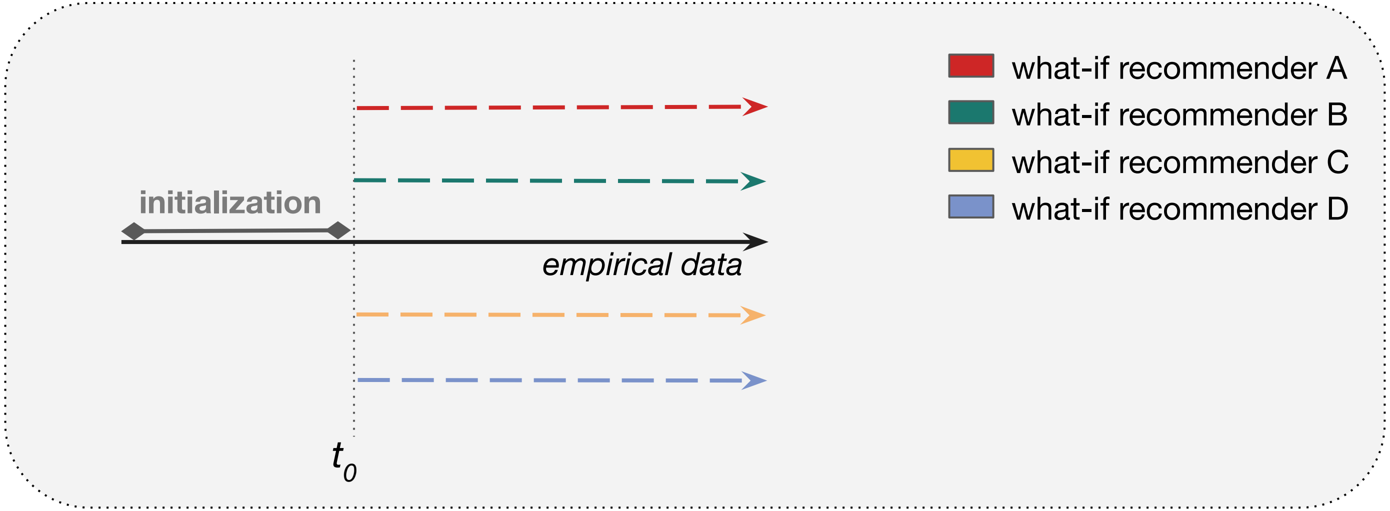
Simulated purchase



Simulated purchase



Simulated sequence of purchases



Considered Recommenders

A. ***Popularity-based CF***: suggest the most popular items purchased by the most similar customer (in terms of cosine similarity between interaction vectors)

B. ***Low Popularity-based CF***: suggest items from the set of the PBCF while discounting for their popularity (divide the score by the popularity)

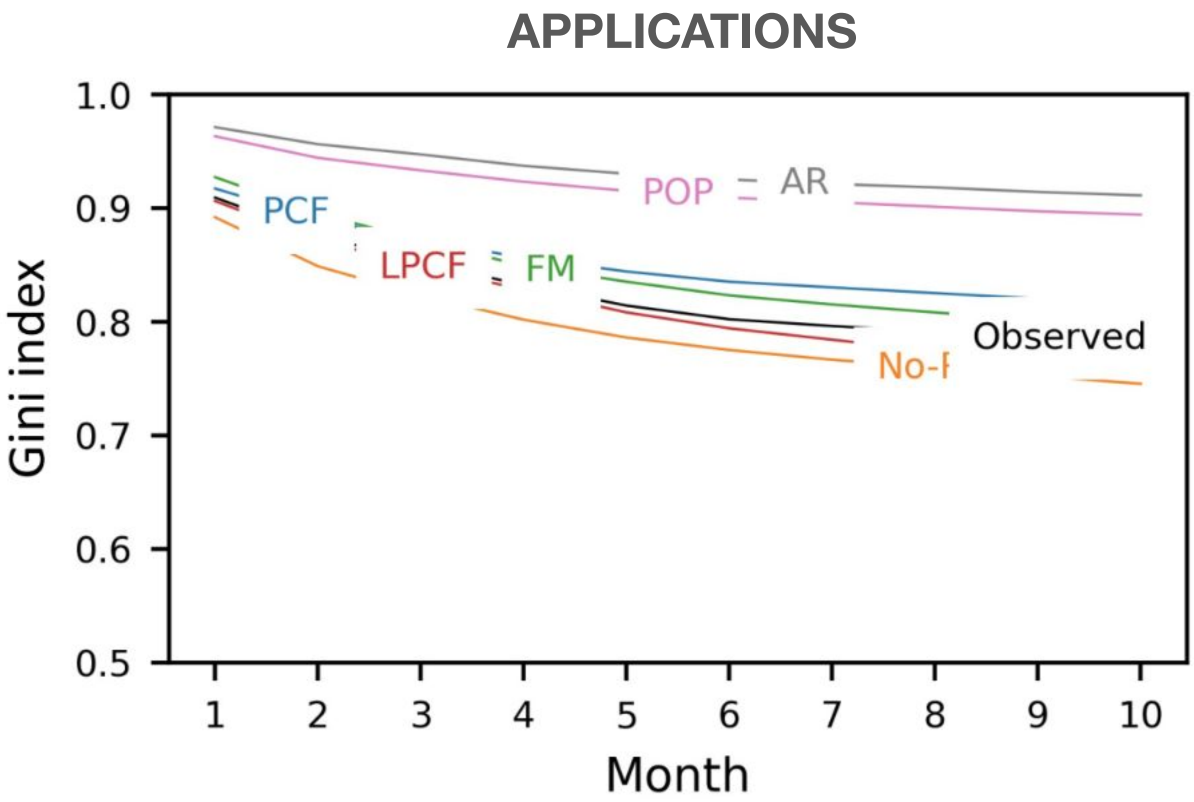
C. ***Factor model***: variation of a Matrix Factorization model for implicit feedback interactions [3].

D. **Popularity based**: suggest the most purchased items

E. ***Average rating***: suggest items with highest average predicted rating

[1] Hu et al. *Collaborative Filtering for Implicit Feedback Datasets*, *8th IEEE International Conference on Data Mining*, Pisa, 2008.

Aggregate diversity over time



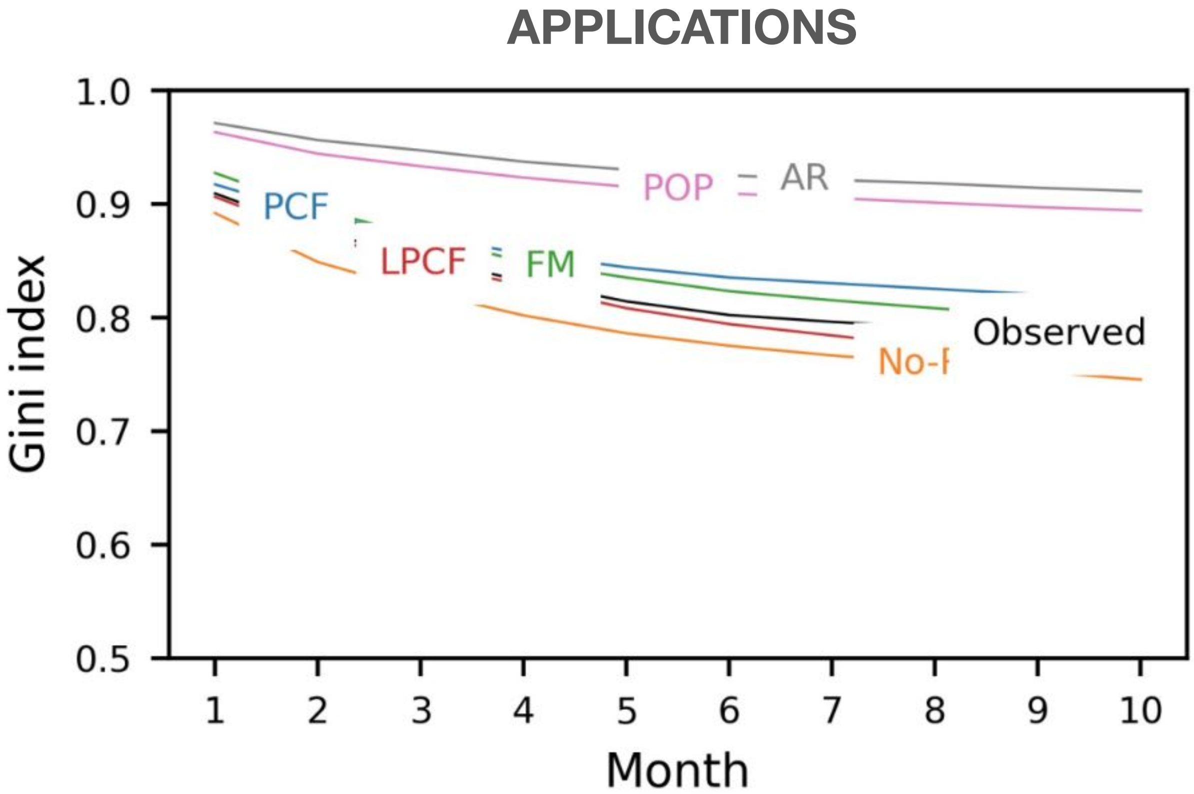
Non-personalised RS

Personalised RS

& observed data

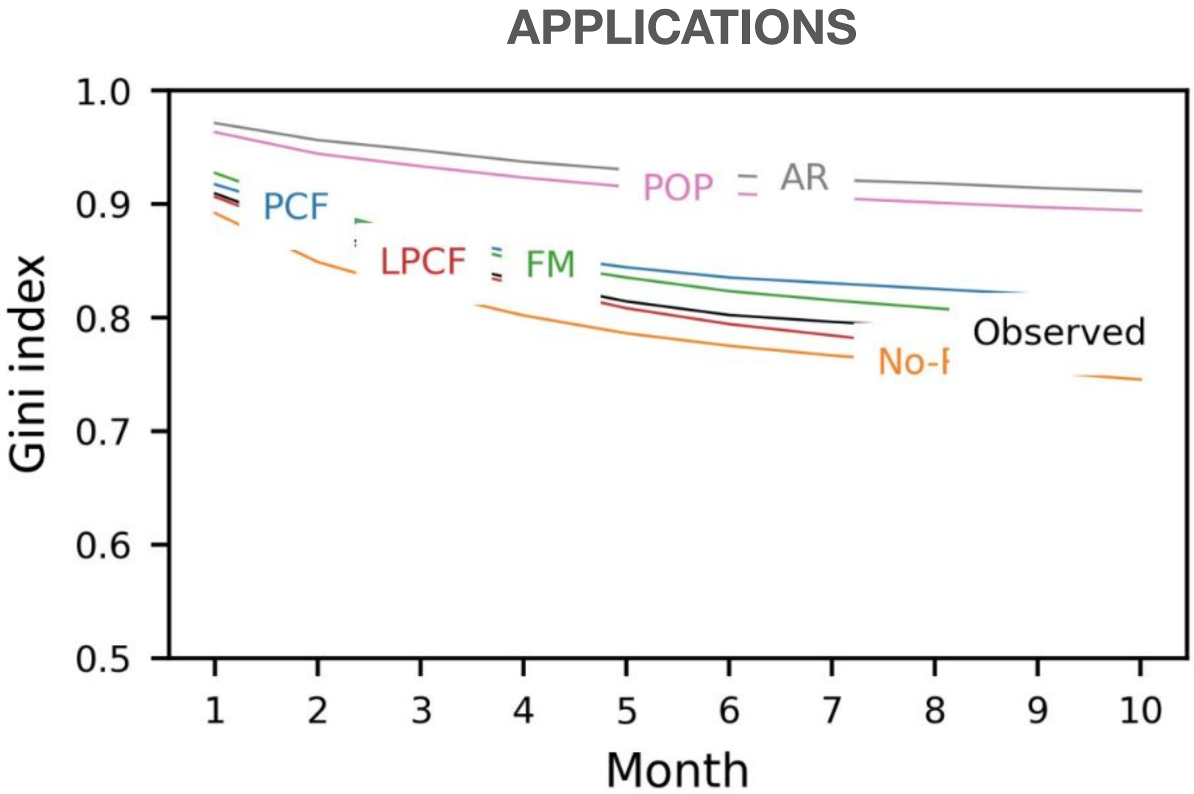
personalisation matters!

Aggregate diversity over time



also specific implementation matters!

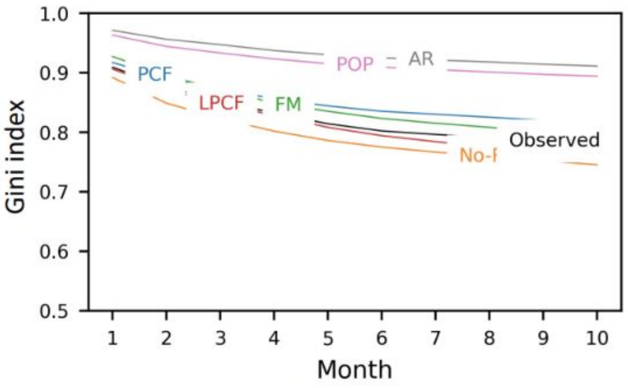
Aggregate diversity over time



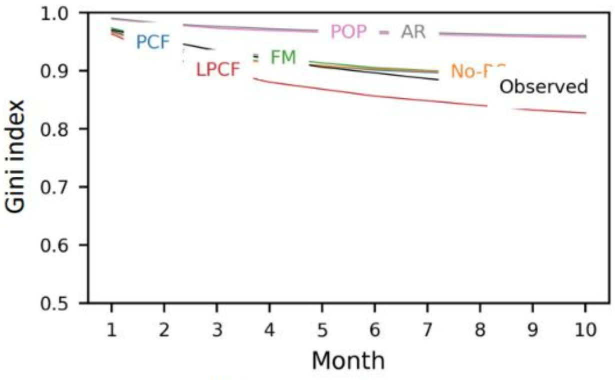
and the specific baseline you choose…

Aggregate diversity over time

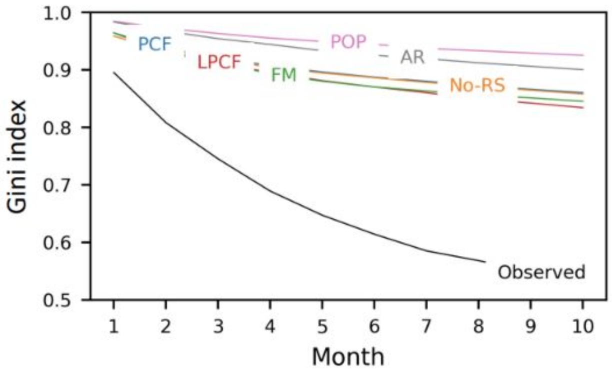
APPLICATIONS



GAMES



KINDLE BOOKS



<u>as well as the domain</u>

In summary

- Personalised RSs can increase aggregate diversity much more than non-personalised ones
- Non personalised RSs suggest items with larger predicted rating compared to personalised RSs
- Increasing the recommendation set size has a marginal effect on diversity choices wrt user “awareness set”

In summary

The impact of Recommender Systems on purchases diversity depends on:

- The **family of recommenders** (popularity based, content based, collaborative filtering)
- The specific **algorithm** deployed
- The **dataset** considered
- The **baseline**
- The **size** of the awareness-set

Main takeaways from Online Retail ecosystem

Main methodologies

- **Abundance of empirical controlled studies**:
  - PRO: disclosing the real behavior of individuals
  - CONS: lack of generalizability and reproducibility
- **Increasing reliance on simulation studies**:
  - PRO: flexible and reproducible
  - CONS: outcomes highly depend on modeling assumptions

Main outcomes

- **Volume of engagement** metrics (empirical studies only): solved
- **Diversity**: a nuanced result - it depends on the recommender, on its hyperparameters, on the metrics employed…

WHAT'S NEXT?

Systematic framework development

➔ Create unified methodologies to synthetize existing results and enable cross-study comparisons

A mechanistic model for users' consumption in simulations

➔ To implement reliable comparative baselines (akin to a control group for simulations) to overcome the need for platform-sourced data

References

[Section 4] Pappalardo, L., Ferragina, E., Citraro, S., Cornacchia, G., Nanni, M., Rossetti, G., ... & Pedreschi, D. (2024). **A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions.** arXiv preprint arXiv:2407.01630.

- D. Lee and K. Hosanagar (2019), **How Do Recommender Systems Affect Sales Diversity? A Cross-Category Investigation via Randomized Field Experiment**, Information Systems Research 30 (1): 239-259.
- D. Holtz, B. Carterette, P. Chandar, Z. Nazari, H. Cramer, and S. Aral (2020), **The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify**, In Proceedings of the 21st ACM Conference on Economics and Computation, Association for Computing Machinery, New York, NY, USA, 75–76.
- Hazrati, F. Ricci (2022), **Recommender systems effect on the evolution of users' choices distribution**, Information processing & Management.

WHAT'S NEXT?

1. **Systematic framework development**
   - Create unified methodologies to synthetize existing results and enable cross-study comparisons
2. **A mechanistic model for users' consumption in simulations**
   - To implement reliable comparative baselines (akin to a control group for simulations) to overcome the need for platform-sourced data

3. *Evaluation metrics beyond volume and diversity*
4. *Include in the framework composite effects of different marketing strategies and commercial objectives*

Experimental Setup: the awareness set A

Naieme, Ricci 2022

- First an aggregation L of two ranked lists is built:
  - *Pop_u*: Items which have not been chosen by *u,* sorted w.r.t. **their popularity**
  - Hut_u: Items which have not been chosen by *u,* sorted w.r.t. **their utility (critical for bias management)**

|A| is the same for every user

- Then, A is obtained by including randomness in such aggregation of lists:
  - The top α*A are taken from L, where α = 0.9;
  - The remaining (1-α)*A are **random items** from the entire collection.

Experimental Setup: the choice model

Naieme, Ricci 2022

The user u chooses an item i in his awareness set with probability:

$$p(u\ chooses\ i) = \frac{e^{v_{ui}}}{\sum_{j \in A_u^l} e^{v_{uj}}}$$

Items with larger utilities are more likely to be chosen but the user don't necessarily select the item with largest utility!

- **v_{ui} - utility of item i for user u: proportional to the predicted rating $\hat{r}_{ui}$, i.e. user taste predicted through a debiased MF approach.**
  - **IPS-MF** (Inverse-Propensity Score) to predict missing ratings: variation of a MF which modify the loss to face **selection bias** in the interaction dataset (the fact that a user may avoid rating an item because he didn't experience it but also because he was non interested: Missing Not At Random): **the idea is to weight each observed user-item interaction in the loss by the inverse of its propensity score, that is the probability that the interaction was observed**
- **The utility of recommended items is adjusted with the level of acceptance;**

FACTOR MODEL

- **Matrix factorization model developed for implicit feedback dataset**
- All interactions are considered as positive feedbacks
  - The confidence level varying w.r.t. the volume of feedback (how many times a user interacted with an item)
- One of the main goal is to handle large datasets efficiently (due to the common sparsity of such a datasets)
- At the time (2008) most of the algorithms work with explicit feedback, so the paper is one of the first that consider implicit feedback efficiently
- The idea is the following:
  - A user may watch a TV show just because she is staying on the channel of the previously watched show. Or a consumer may buy an item as gift for someone else, despite not liking the item for himself
  - As the same, the user might be unaware of the existence of the item, or unable to consume it due to its price or limited availability
- **So they assign a confidence level to each pair user-item**
- **Then use the standard matrix factorization, proposing an optimization for large datasets**
- What it does means?
  - **Find a vector u for a user and a vector i for an item and project the vectors into a common latent factor space where they can be compared directly**
  - Since they use confidence, there are observations for every pairs, not only the positive ones
  - Observations >1 are set to 1, otherwise 0
  - They propose a confidence calculation as cui = 1 + αrui where rui is the values of the interaction matrix (for user u and item i) and α is a tunable parameter to determine the level. So there is a minimum level (1) also for negative (or unknown) observations

Metrics

Simulation - Observational

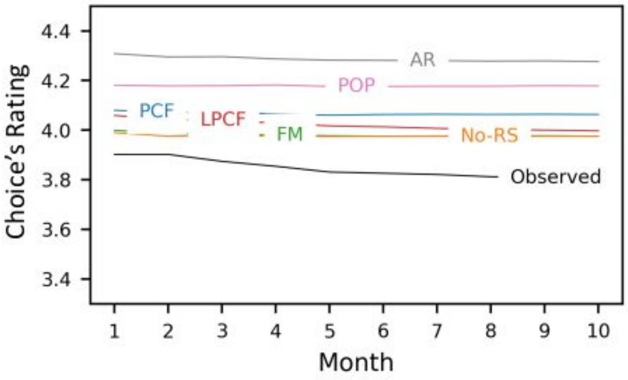
N. Hazrati, F. Ricci, Information Processing and management, International journal, 2022

- **Choice's rating (individual-level)**: for each user, the average of the predicted rating of the chosen items
  - To predict the ratings, the IPS-MF model is used [4]
  - The model predict what would be the rating for each item, for a given user

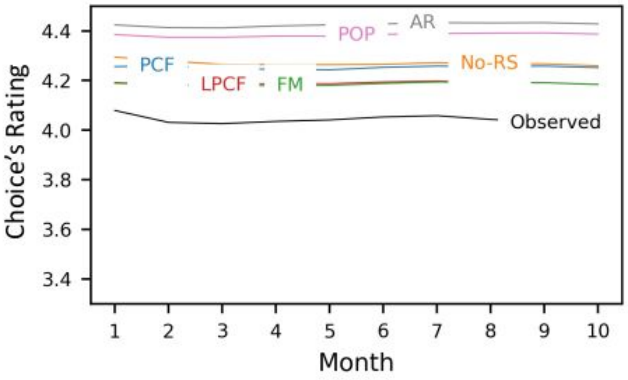
Average rating: impact of different models

N. Hazrati, F. Ricci, Information Processing and management, International journal, 2022

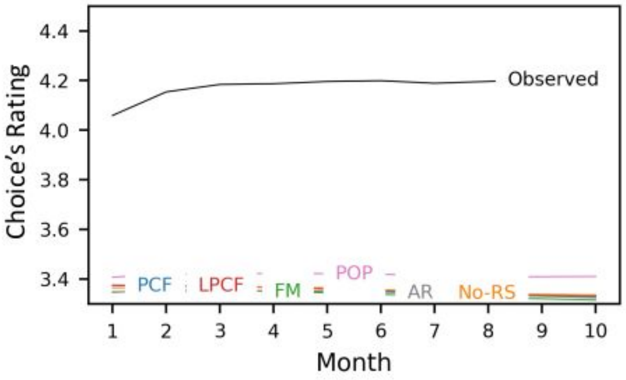
APPLICATIONS DATASET



GAMES DATASET



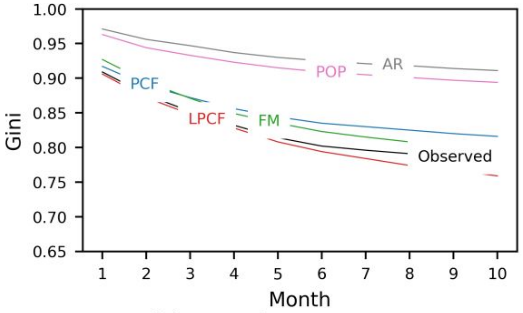
BOOKS DATASET



Aggregate diversity: impact of the awareness set size

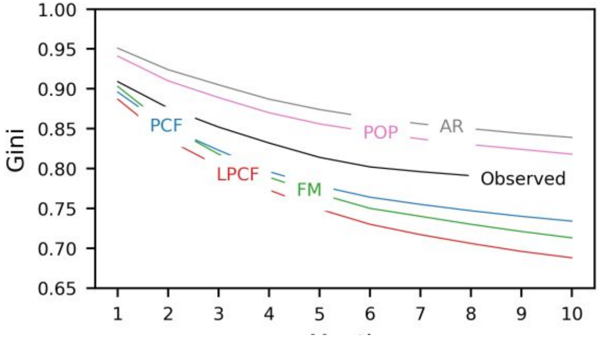
N. Hazrati, F. Ricci, Information Processing and management, International journal, 2022

Is **awareness set size** and **recommendation set size** impact on the choice diversity at collective level?

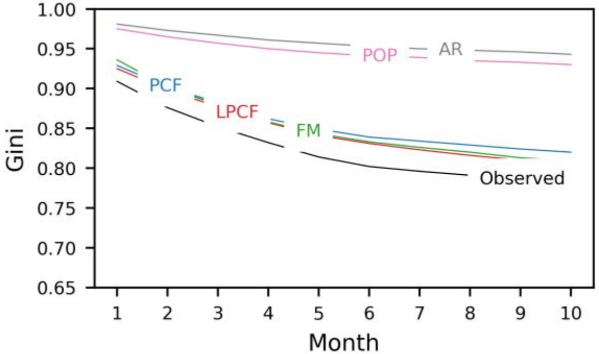


(a) Original parameter setting

Awareness set size = 2000/3000
Recomm. set size = 50



augmented **awareness set size**

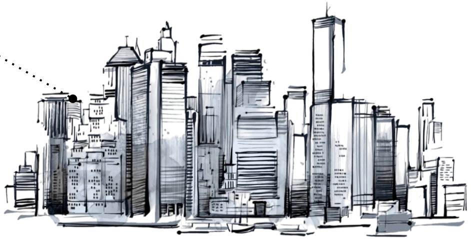


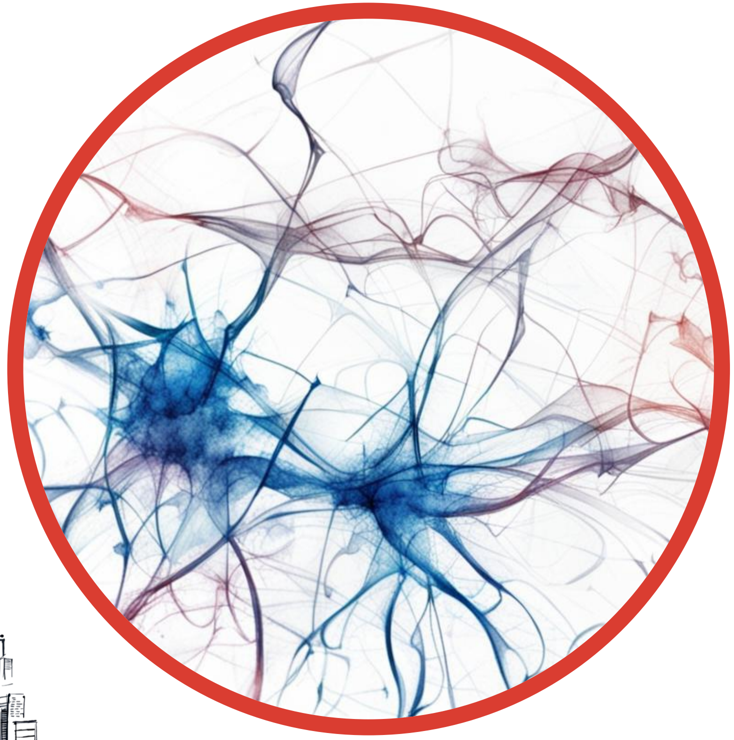
augmented **recommendation set size**

URBAN MAPPING

Recommender systems built for urban living

Helping individuals navigate daily choices such as **route directions**, **places to visit**, and **homes to rent**.





Main Platforms

mapbox
Bing Maps
waze
TOMTOM

Navigation Services

airbnb
Booking

House-renting Services

Uber
lyft

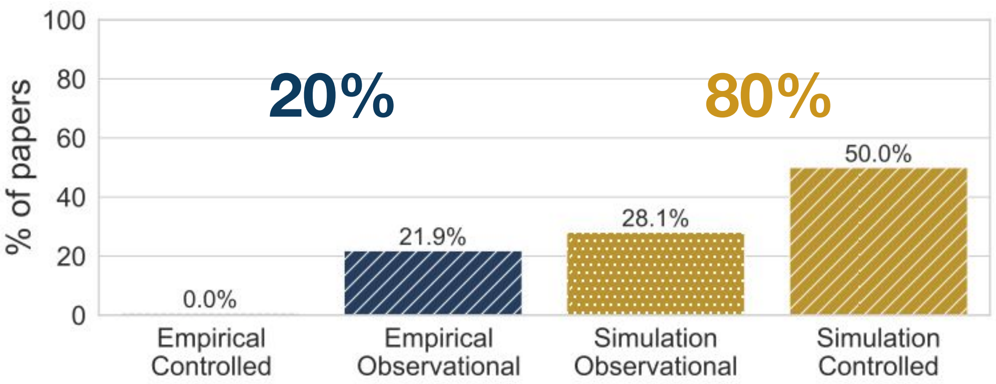
Ride-hailing

Tripadvisor
yelp

POIs Recommender

Employed Methodologies

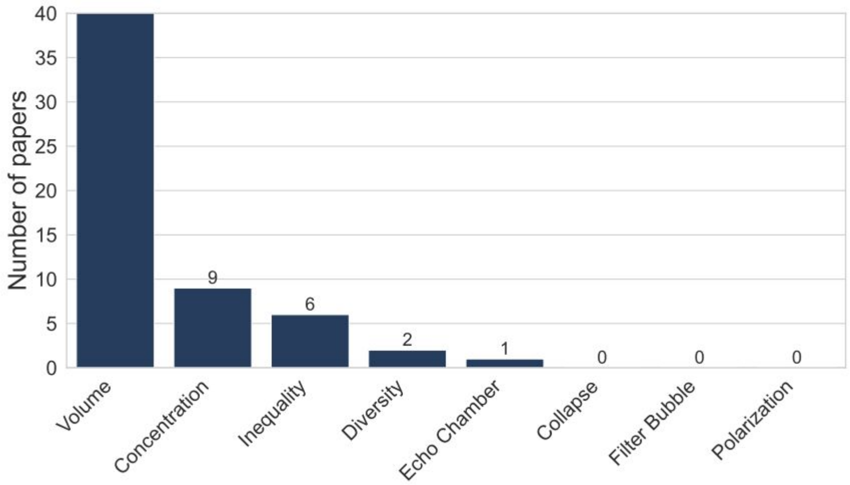
L. Pappalardo et al. A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630



- Predominance of simulation over empirical studies
- Data owned by big-tech companies

Main Outcomes

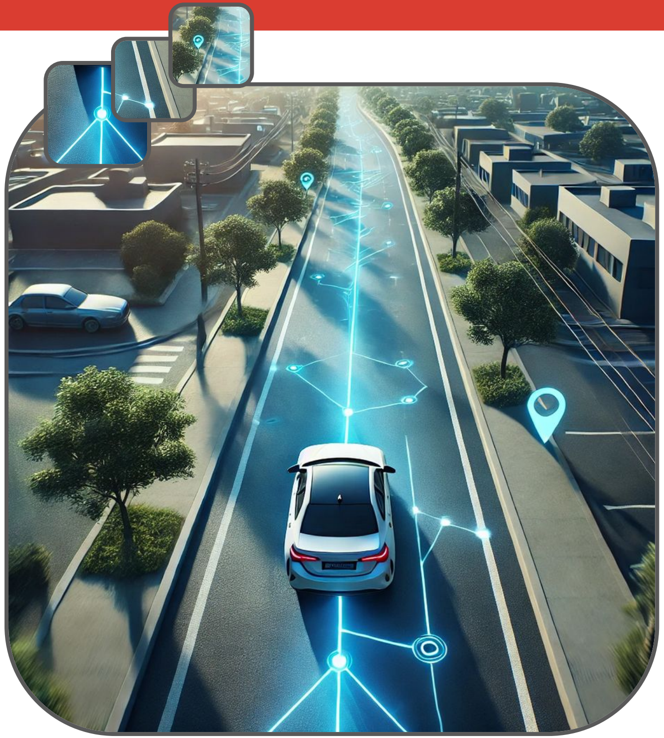
L. Pappalardo et al. A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630



- Main Outcomes:
    - Systemic-level: **volume**, **concentration**, **inequality**, **diversity**
    - Common **targets**: CO_2 emissions, travel time, and user costs (e.g., ride fare)

Navigation Services

- Several studies in this ecosystem focus on the urban **impact** of **navigation services**
- Navigation Services suggest the **fastest path** or a slight variation to reach a destination
- The **aggregation** of many **individually “optimal”** suggestions may not be **collectively optimal**



The New York Times

Navigation Apps Are Turning Quiet Neighborhoods Into Traffic Nightmares

The corner of Fort Lee Road and Broad Avenue in Leonia, N.J. With traffic apps suggesting shortcuts for commuters through the borough, officials have decided to take a stand. Bryan Anselm for The New York Times

Cronaca

Il gigante e la viabilità

I sindaci dell'Alto Adige contro Google Maps: tutta colpa dei furbetti della coda

Con i 'percorsi alternativi', traffico da bollino nero e paesi intasati, sindaci contro Google, Kompatscher: "Serve un divieto di deviazione come all'estero"

15/10/2024

Autostrada del brennero con traffico

2024

MailOnline

Residents outrage after Waze app used to avoid traffic ends up sending Los Angeles drivers down once quiet 'hidden' street

TORONTO STAR

This tiny Toronto street is choked by traffic chaos. Residents are 'fuming mad' at being trapped by daily gridlock

Since Eglinton Crosstown LRT construction began in 2011, the street has been jammed by drivers, guided there by Google Maps or Waze.

KentOnline
News you can trust

Kent residents say councils call for car sat-navs to be banned in lorries won't help

The Standard

Map apps like Waze 'turning quiet London streets into polluted rat runs'

The Telegraph

Britain's new road rage: how traffic rules are tearing our neighbourhoods apart

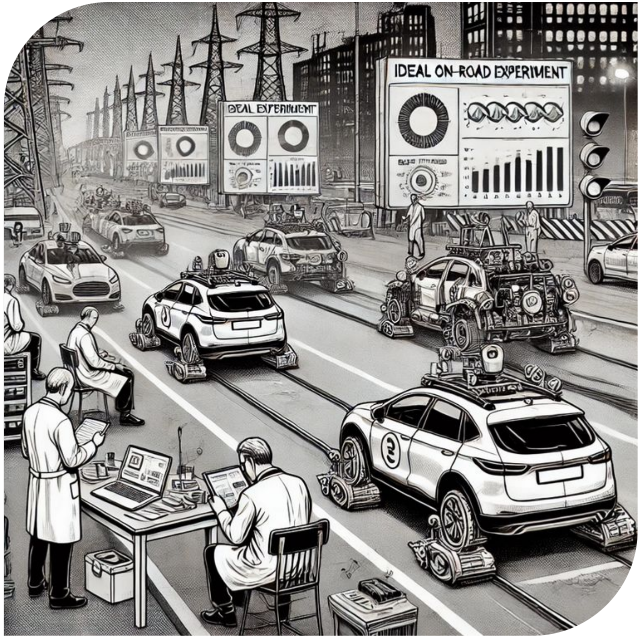
The Guardian

'Rat-running' increases on residential UK streets as experts blame satnav apps

Motoring on minor roads doubled between 2009 and 2019, regional figures reveal

- **Traffic jam by GPS: A systematic analysis of the negative social externalities of large-scale navigation technologies**, PLoS One 2024
- **In WAZE we trust? GPS-based navigation application users' behavior and patterns of dependency**, PLoS One 2022

How to study this phenomenon?



Ideal Scenario: On-Road Experiments

- assign routes to vehicles
- collect trip-related data (e.g., CO2 emissions and travel time)

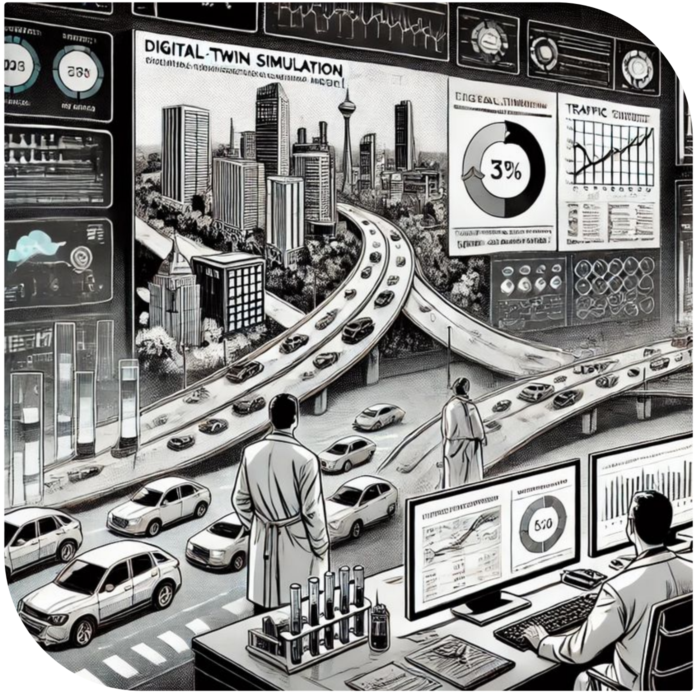
Limitations (---):

- **non-replicable**: cannot be recreated under identical initial conditions
- large-scale, real-world experiments are expensive and **logistically challenging**

How to study this phenomenon?

Simulation-Based Methodology

- simulate routes using a digital-twin model
- collect simulated data (e.g., CO_2 emissions and travel time)

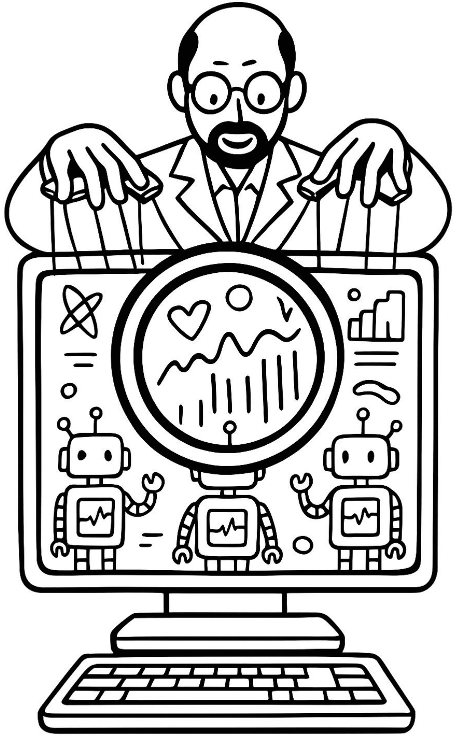


Advantages (+++):

- easy to reproduce and control

Limitations (---):

- findings may not fully translate to real-world traffic conditions



Simulation Studies

Quantifying the sustainability impact of Google Maps: A case study of Salt Lake City

Arora et al., Arxiv 2021

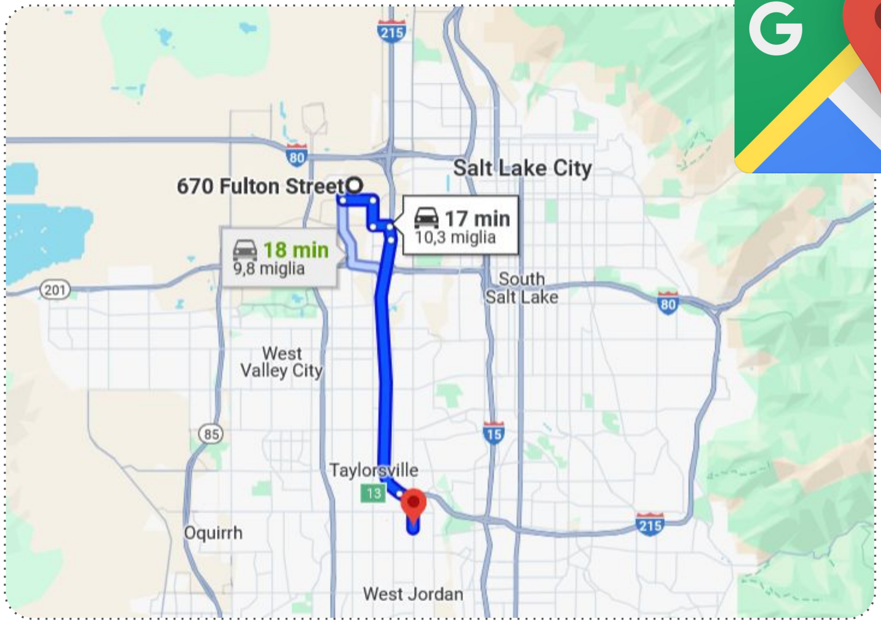
Type: Simulation Controlled

VLOP: Google Maps

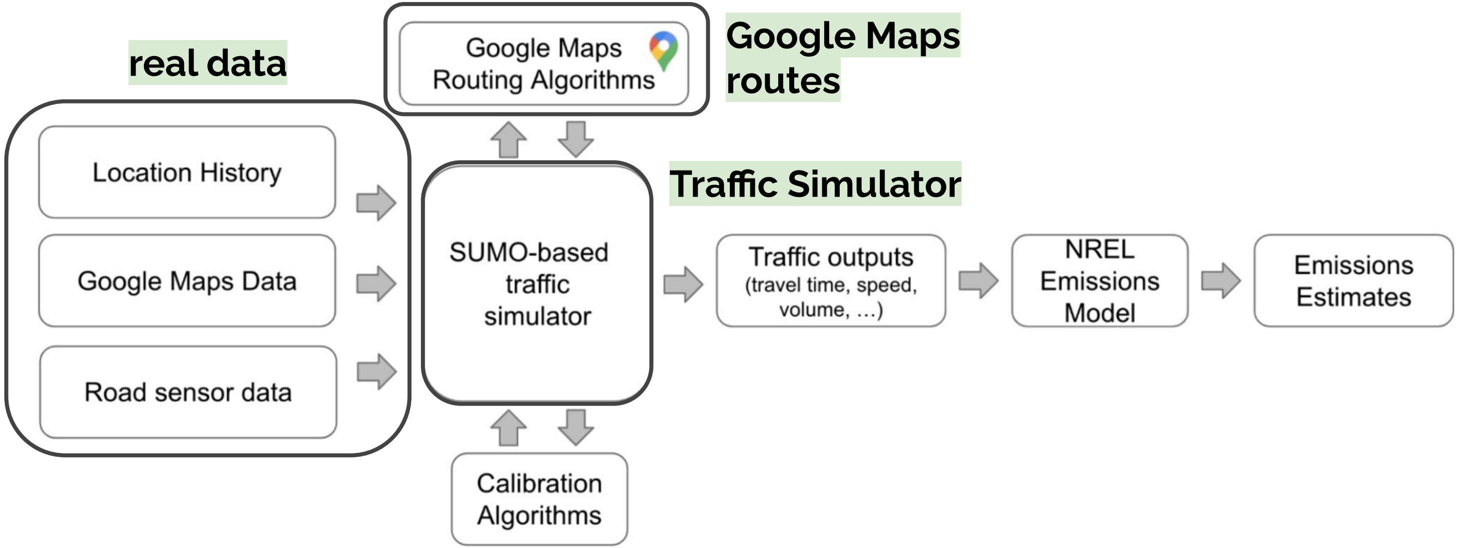
Outcomes: Volume.Individual
Volume.Systemic

Impact of Google Maps

- **RQ:** Does Google Maps reduce emissions and travel time, and by how much?



Impact of Google Maps

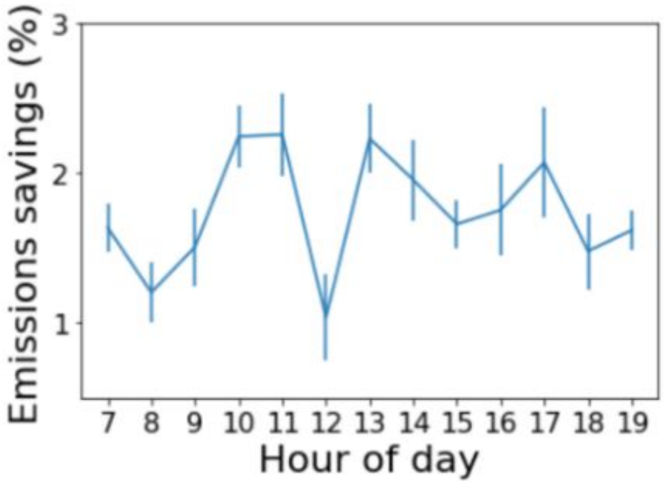


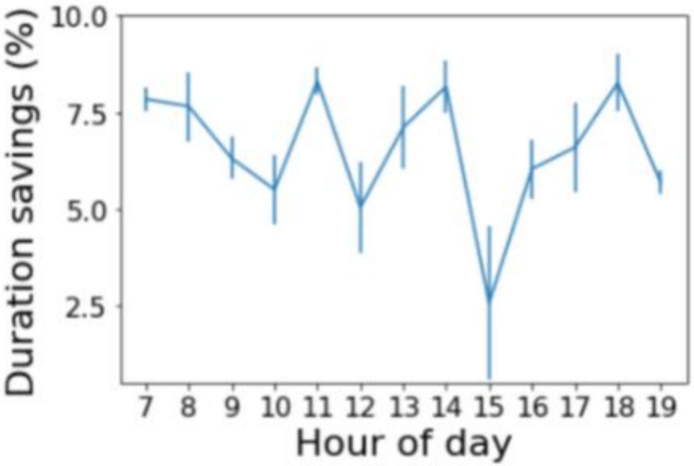
Two Scenarios:

- **Baseline**: Vehicles follow historical (**observed**) **routes.**
- **Routed**: A subset uses **route suggested** by Google Maps

Impact of Google Maps

- **GMaps** users **reduce** CO_2 emissions by **1.7%** and travel time by **6.5%**
- The reduction of 3.4% (CO_2) and 12.5% (travel time) for users whose **suggested** routes **differ** from their original ones





Limitations

- Study **on Google Maps** performed **by Google Maps**
- Only **one city** and **navigation** service (with **fixed adoption** rate)
- **Lacks open access**
- A valuable **starting point**

Navigation services amplify concentration of traffic and emissions in our cities

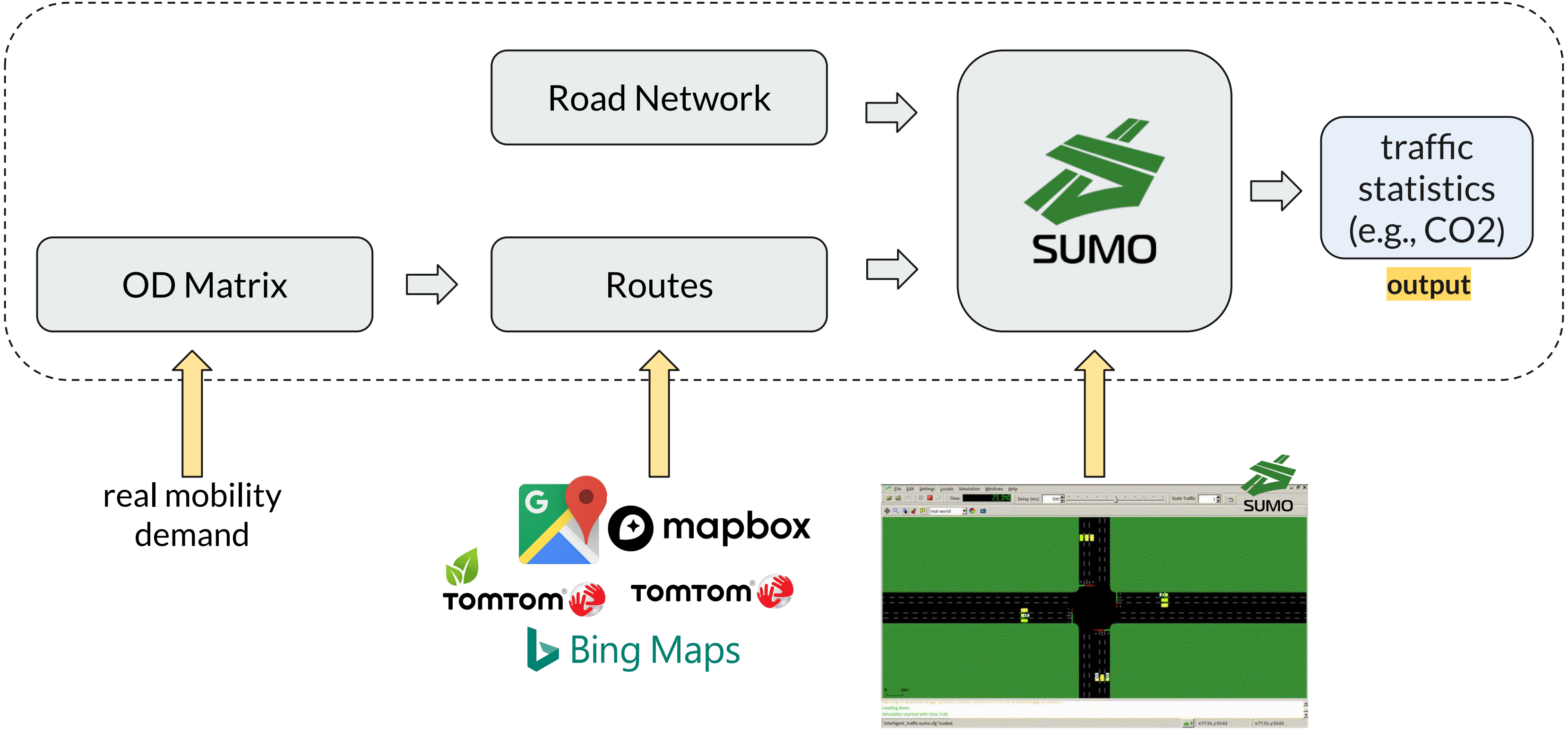
Cornacchia et al., Arxiv 2024

Type: Simulation Controlled

VLOP: Google Maps, TomTom, Mapbox, Bing Maps

Outcomes: Concentration (increase)

An Open Simulation Framework



Experimental Setup

- Vary the adoption rate **r** from 0% to 100%

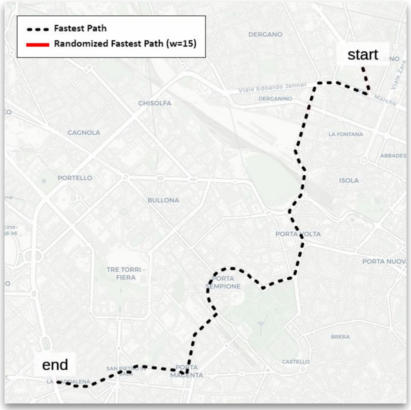
Treatment Group

r% of the vehicles follow the **suggestions** of a navigation service

TOMTOM®

Control Group

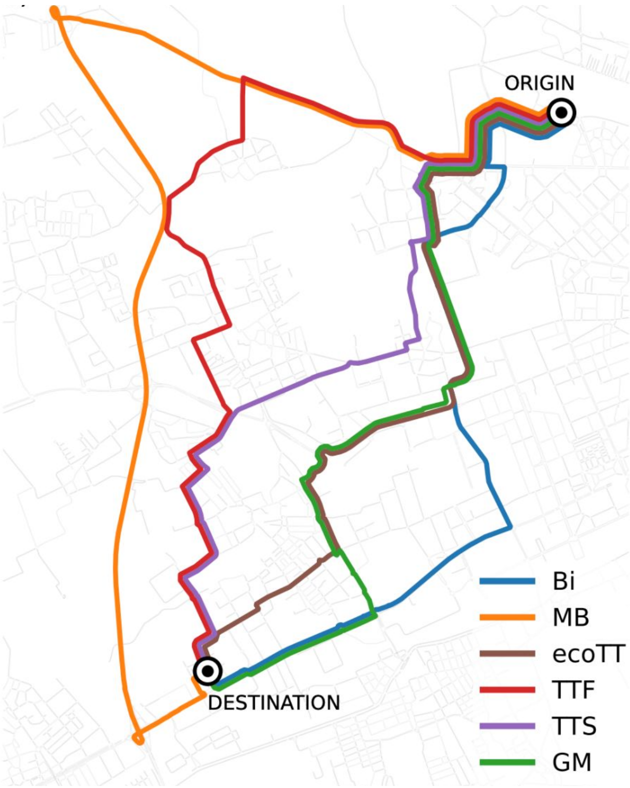
(100-r)% of the vehicles follow a perturbation of the **fastest** path



* experiment repeated 10 times for statistical robustness

Cognitive aspects of traffic simulations in virtual environments. 2012, 10.11128/sne.22.tn.10127

Experimental Setup



- Uniform distribution of **departure time** (in **1 hour**)
- **Milan, Florence, Rome**

SHORT

FASTEST

Results: traffic patterns

Milan, Italy - TomTom fastest

0% adoption rate

Results: traffic patterns

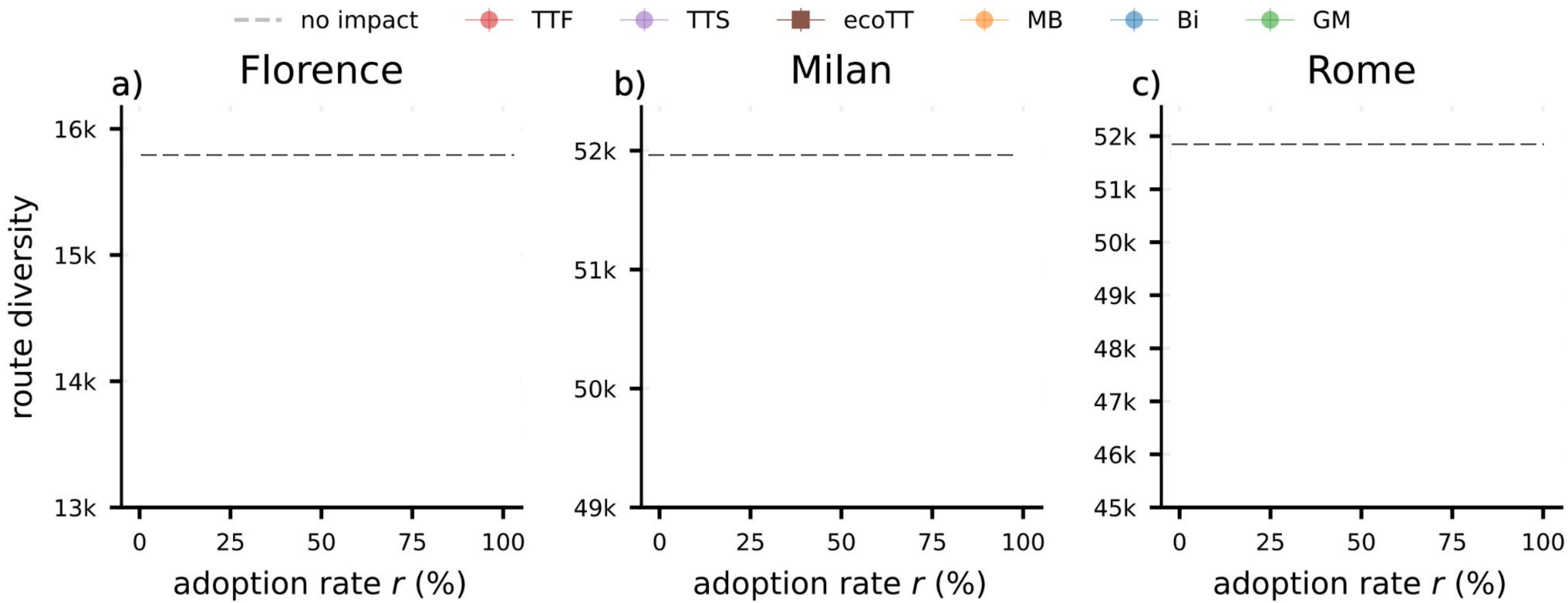
Milan, Italy - TomTom fastest

Route Conformism

- As navigation adoption increases, **routes converge** on the same few roads.
- Route diversity decreases, and traffic becomes concentrated.

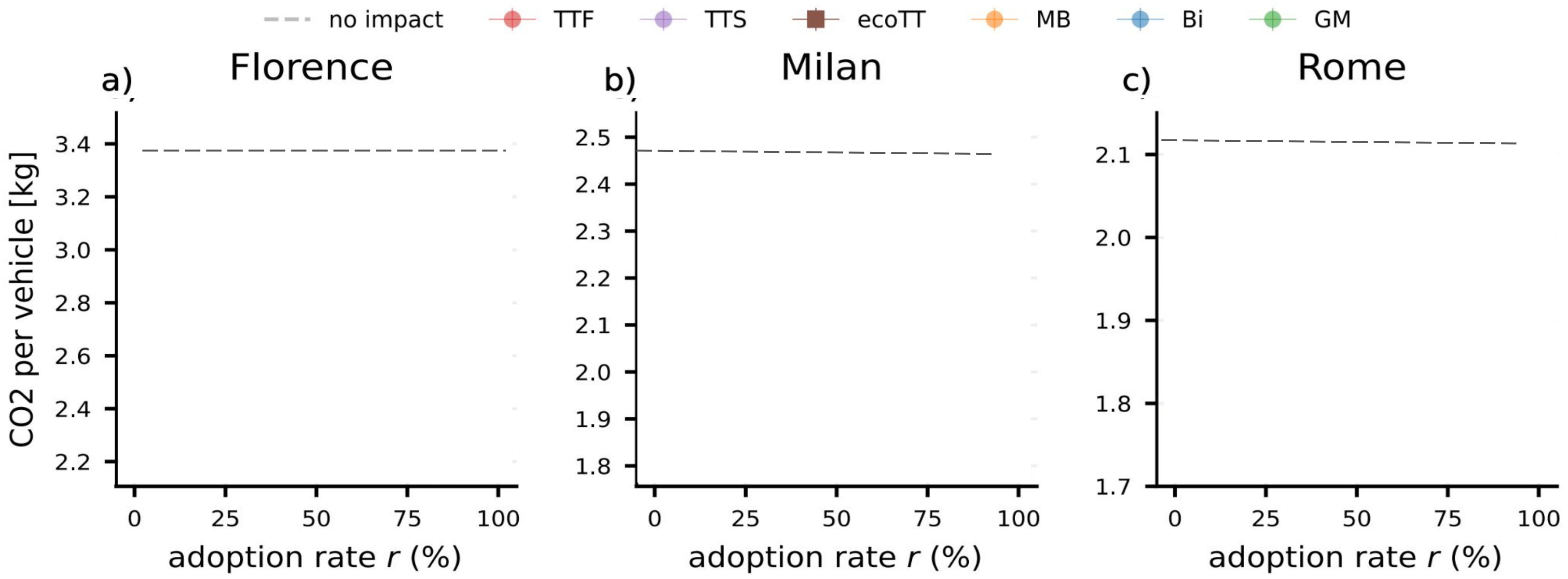
Results: route diversity

* Results are consistent with traffic loads



- Low adoption rate (0-20%): route diversity slightly increases (<1%)
- High adoption rate: strong diversity reduction (up to 15%)

Results: CO2 emissions



- Low traffic loads: services are beneficial, reducing CO2 emissions
- High traffic loads: with high adoption rates, the benefits plateau or event revert

In summary

Navigation services **amplify concentration** of traffic

Navigation services may:

- exacerbate exposure **inequality**
- **interfere** with existing **policies**
- **impact** the **economic** and social fabric of neighbourhoods

0% adoption rate

Beautiful…but at What Cost? An Examination of Externalities in Geographic Vehicle Routing

Johnson et al., ACM on Interactive, Mobile, Wearable, and Ubiquitous (2017)

Type: Simulation Observational

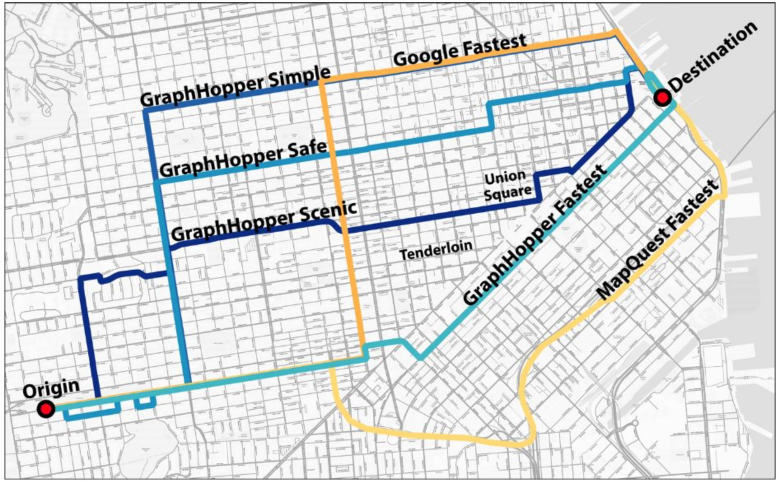
VLOP: Routing Criteria

Outcomes: Concentration

Impact of Routing Criteria

They analyze different routing criteria:

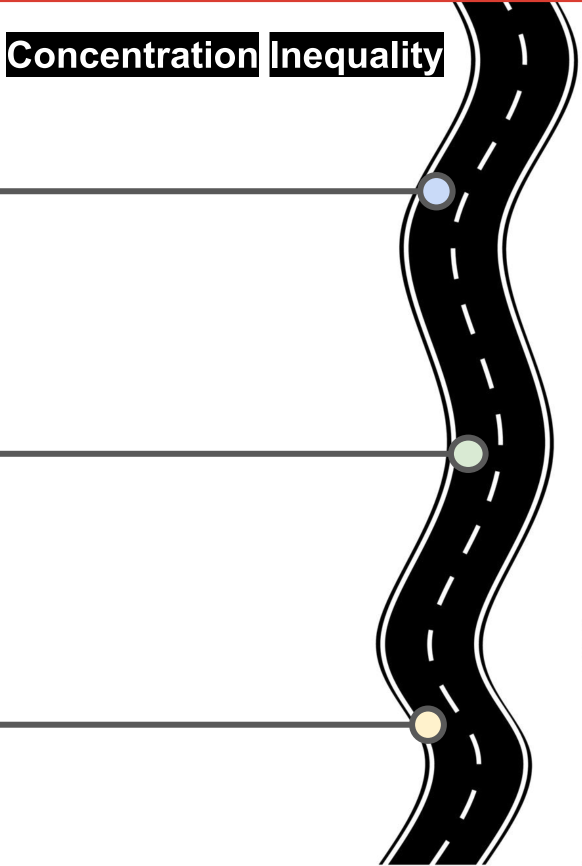
1. **Scenic Routing –**
   favors visually pleasant routes
2. **Safety Routing –**
   avoids high-crime or accident-prone areas
3. **Simplicity Routing –**
   minimizes route complexity



Experiments in four cities:

San Francisco • New York City • London • Manila

Impact of Routing Criteria



Scenic Routing

- Produces **complex** routes
- **Diverts traffic** to **parks**, **tourist** spots, and **slower** roads

Safety Routing

- Produces **complex** routes
- Shifts flow away from unsafe zones

Simplicity Routing

- Channels **traffic** onto **highways**
- Does **not** strongly **favor** or avoid any **regions**

Impact of Routing Criteria

In conclusion:

- **Routing** choices have **consequences**: Optimizing for beauty, safety, or simplicity **reshapes traffic patterns** in ways that may harm communities or reduce safety

- Routing designers must consider **social** and **geographic** impacts of each strategy

The Urban Impact of AI: Modeling Feedback Loops in Next-Venue Recommendation

Mauro et al., Arxiv 2025

Type: Simulation - Observational

VLOP: -

Outcomes: Diversity
Inequality

POI - Feedback Loop

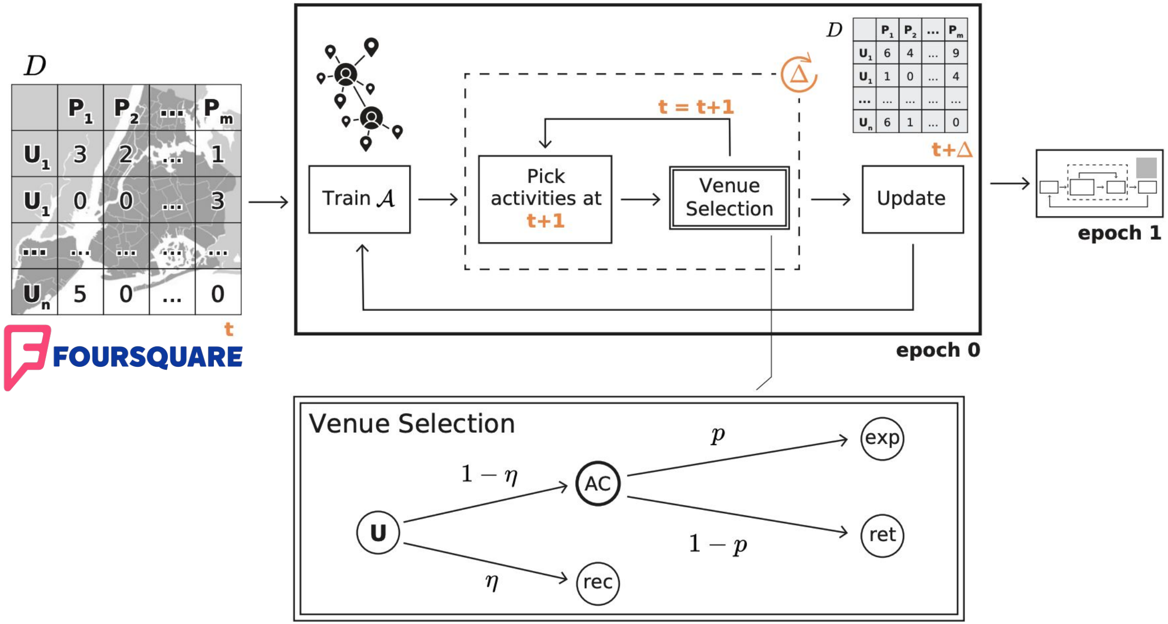
Next-venue recommenders (e.g., Google Maps, Yelp) guide **urban mobility** decisions.

Yet, their **systemic impact** on cities is **poorly understood**

POI - Feedback Loop

Modeled human–AI feedback loops:

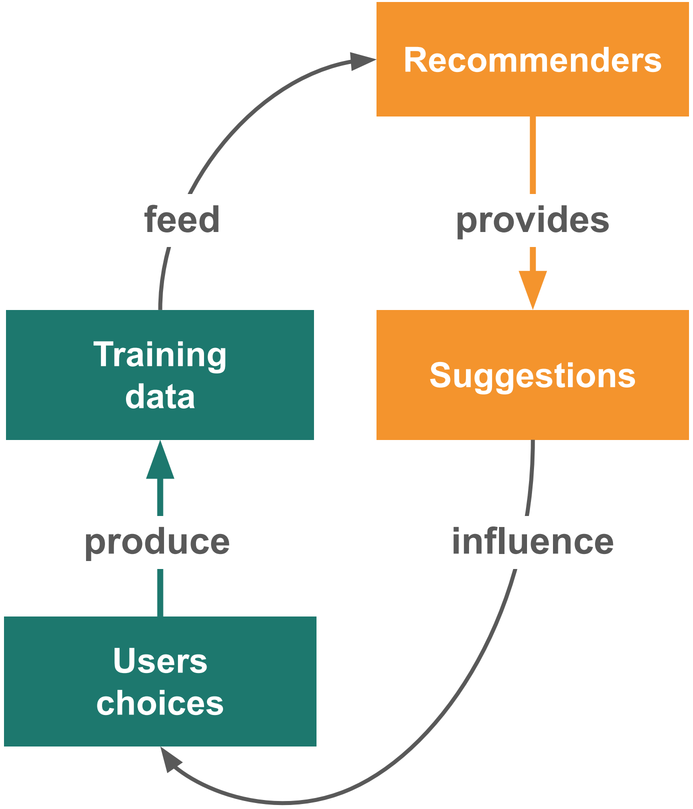
*recommendations **influence** movement → data **retrains** system → affects **future** mobility*



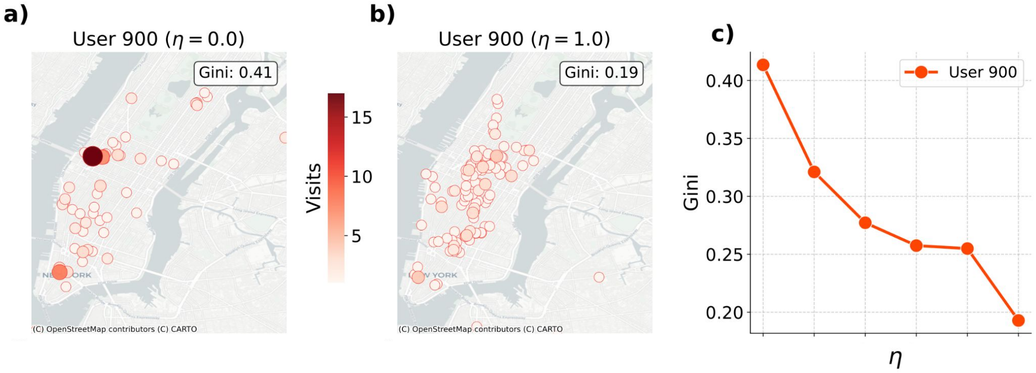
POI - Feedback Loop

Modeled human–AI feedback loops:

- *recommendations **influence** movement*
- *movements **produce** data*
- *data **retrains** system*
- *affects **future** recommendations*



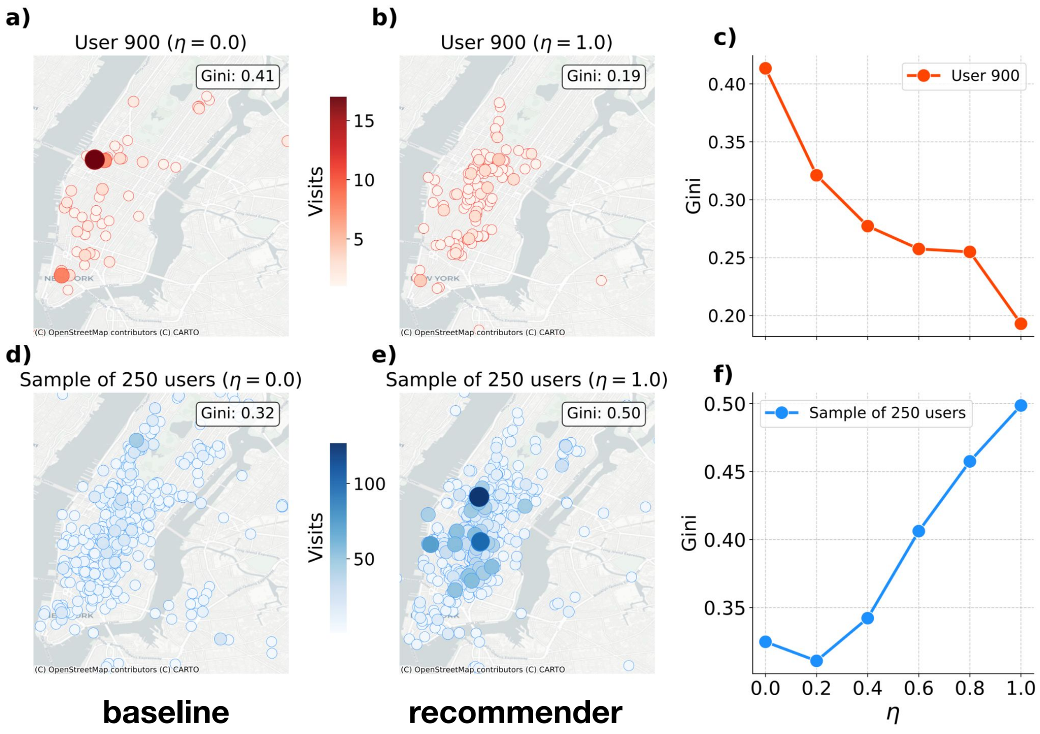
POI - Feedback Loop



baseline **recommender**

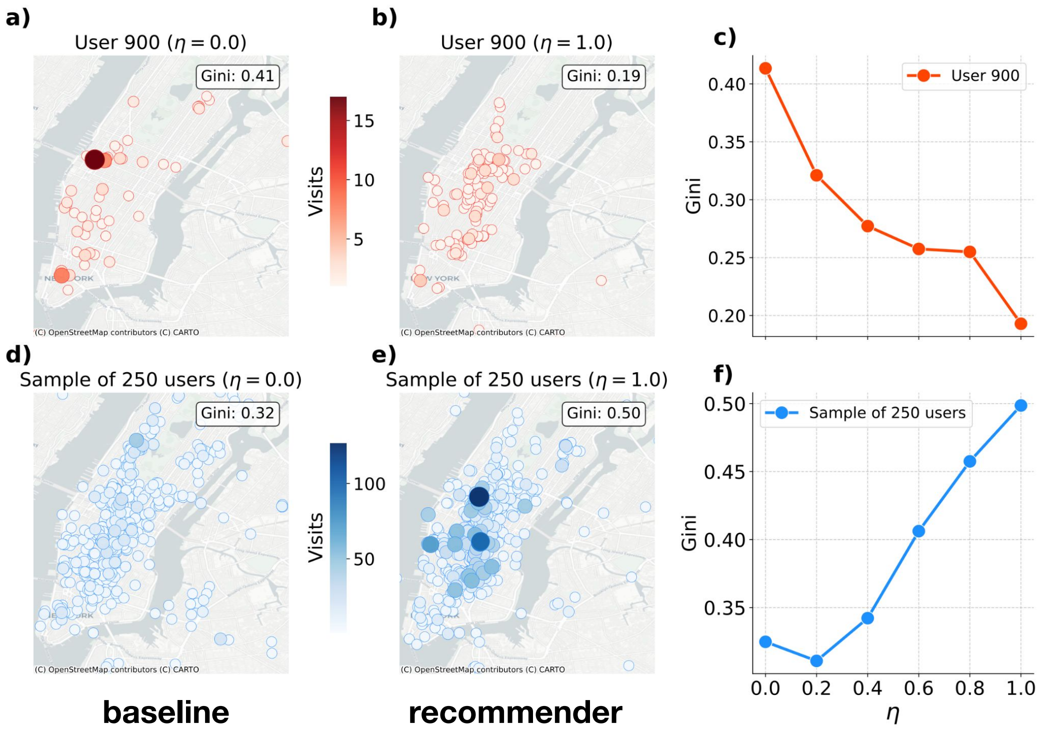
- **Individual-level:** diversity **increases** as people **explore more** venues

POI - Feedback Loop



- **Individual-level:** diversity **increases** as people **explore more** venues
- **Collective-level: inequality increases** as visits **concentrate** on **few** popular places
- **Rich-get-richer** dynamics emerge

POI - Feedback Loop



- **Individual-level:** diversity **increases** as people **explore more** venues
- **Collective-level: inequality increases** as visits **concentrate** on **few** popular places
- **Rich-get-richer** dynamics emerge

Recommenders promote **personal variety** but cause **collective centralization**

Discussion

How can we **mitigate** the effect of **navigation services**?

Empirical Studies

Digital Discrimination: The Case of Airbnb.com

Benjamin Edelman and Michael Luca

Type: Empirical observational

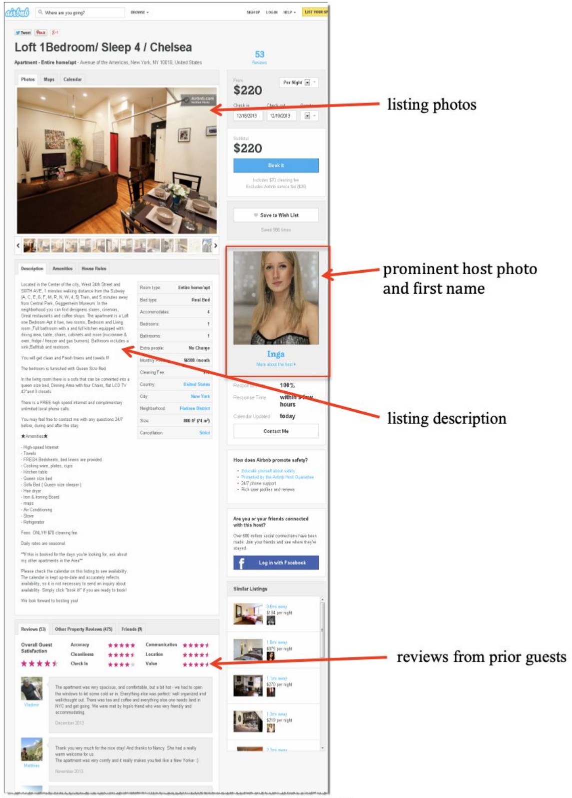
VLOP: AirBnB

Outcomes: Inequality

Hosts-guests interaction

RQ: Are there racial revenue inequalities on airBnB?

- **Dataset** constructed scraping 2012 airbnb listings
- **Hired** workers on Amazon for tagging
  - **Ethnicity** of owners
  - **Quality** of the pictures
- Understand if **non-black** hosts earns more than **black** ones



Source: Authors' use of Airbnb (December 8, 2013).

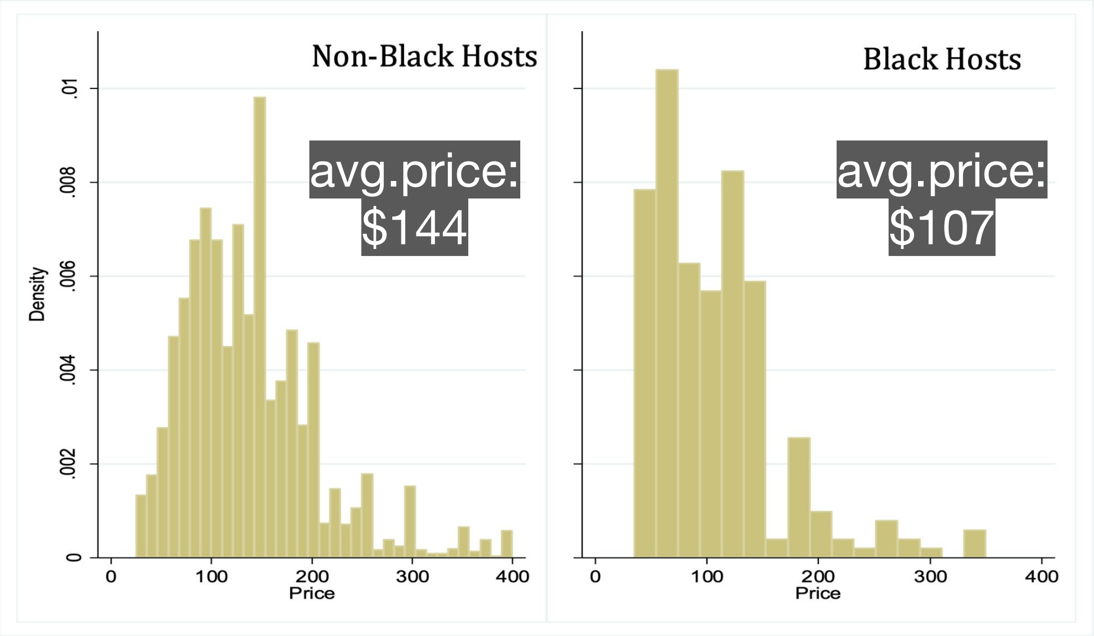
Hosts-guests interaction

RQ: Are there racial revenue inequalities on airBnB?

Hosts-guests interaction

RQ: Are there racial revenue inequalities on airBnB?

- **Raw** data
- **Not** controlling for **confounding** variables
- **$37** avg. difference
  - ~26% less



Hosts-guests interaction

RQ: Are there racial revenue inequalities on airBnB?

- **Controlling** for various attribute of the listings
  - Both scraped and human-tagged
- **Reduction** down to 12%
  - But still present

	Dependent Variable: Price					
	(1)	(2)	(3)	(4)	(5)	(6)
Number Accommodated	9.605***	11.492***	11.647***	10.903***	10.824***	10.808***
	(1.30)	(1.32)	(1.32)	(1.31)	(1.30)	(1.30)
Whole Apartment	64.025***	52.292***	51.651***	50.222***	50.788***	50.945***
	(1.97)	(2.10)	(2.12)	(2.15)	(2.13)	(2.13)
2 Bedrooms	2.314	-5.657*	-5.272	-5.915*	-5.106	-4.671
	(3.30)	(3.27)	(3.27)	(3.38)	(3.35)	(3.33)
3 Bedrooms	-18.315***	-22.424***	-22.053***	-15.038***	-15.258***	-14.507***
	(6.83)	(6.99)	(7.06)	(5.13)	(5.04)	(5.19)
4+ Bedrooms	-22.865***	-28.349***	-28.332***	-28.941***	-27.796***	-27.050***
	(5.21)	(4.63)	(4.58)	(4.69)	(4.61)	(4.60)
Location Rating		22.497***	-63.213***	-74.325***	-72.798***	-71.155***
		(1.31)	(16.21)	(16.16)	(16.28)	(16.30)
Location Rating ^2			4.904***	5.475***	5.397***	5.303***
			(0.94)	(0.93)	(0.94)	(0.94)
Check-In Rating		-1.866	-1.239	-0.140	-0.211	-0.292
		(2.43)	(2.34)	(2.42)	(2.41)	(2.39)
Communication Rating		-2.199	-2.100	-1.531	-1.606	-1.537
		(2.52)	(2.51)	(2.54)	(2.53)	(2.53)
Cleanliness Rating		1.141	1.114	-0.737	-0.542	-0.559
		(1.40)	(1.40)	(1.42)	(1.42)	(1.42)
Accuracy Rating		2.118	2.544	1.440	1.341	1.166
		(1.76)	(1.75)	(1.75)	(1.73)	(1.72)
Has LinkedIn		10.193***	8.929***	8.664***	8.455***	8.404***
		(3.28)	(3.29)	(3.26)	(3.25)	(3.24)
Has Facebook		0.006**	0.006**	0.006**	0.005*	0.006**
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Has Phone Number		12.282***	12.990***	13.583***	12.543***	12.338***
		(4.52)	(4.48)	(4.64)	(4.61)	(4.64)
Has Twitter		0.001	0.001	0.001	0.001	0.002
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Picture Quality				11.909***	-8.066	
				(1.04)	(4.98)	
Picture Quality ^2					2.415***	
					(0.65)	
Picture Rating Indicators						Yes
Apartment Size - Whole Apartment Interactions	Yes	Yes	Yes	Yes	Yes	Yes
Constant	62.988***	66.735***	66.402***	24.231***	62.230***	49.449***
	(2.97)	(3.97)	(3.97)	(5.28)	(9.44)	(7.23)

Offline biases in online platforms: a study of diversity and homophily in Airbnb

Victoria Koh, Weihua Li , Giacomo Livan and Licia Capra

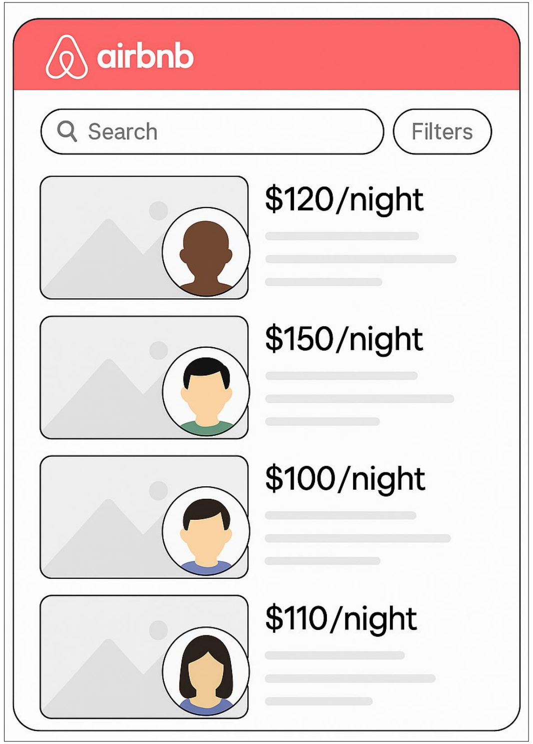
Type: Empirical observational

VLOP: AirBnB

Outcomes: EchoChamber

Hosts-guests interaction

- Online platforms like Airbnb are seen as **neutral**
- But, do they **replicate** or even **amplify** real-world social **biases**?
- Study investigates demographic **representation** and **interaction** patterns on Airbnb.
- **Data** gathered from 5 cities: Amsterdam, Dublin, Hong Kong, Chicago, and Nashville



Hosts-guests interaction

RQ1: How diverse is AirBnB **user base**?

- User base predominantly
  a. **Female**
  b. **White**
     i. Even in cities with more **diverse** racial compositions

Hosts-guests interaction

RQ2: How do host and guests **interact**?

- Network **rewiring** to identify edges in the host-guest network that cannot be attributed to **chance**

- Study of **homophily** of the user-guest **network**
  a. Strong for **gender**
  b. Mild for **race**
  c. Absent for **age**

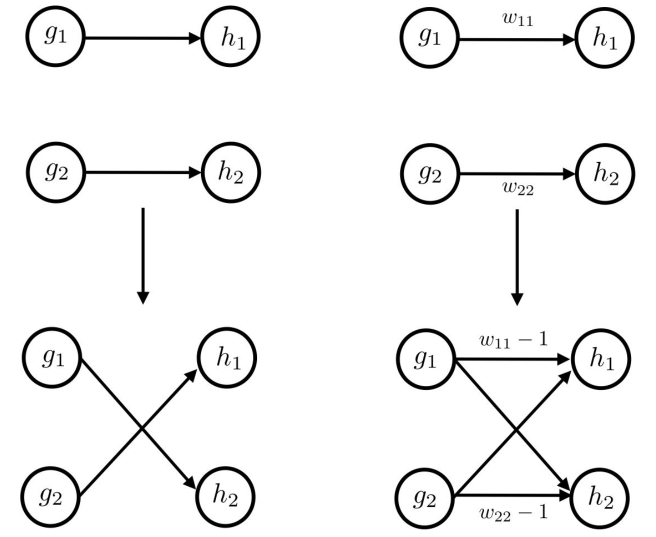


Figure 1 xSwap rewiring moves. Left: swap of two links with unit weight. Right: swap of a unit weight subtracted from two links with weights larger than one

Conclusions and future works

- **Variety** of recommenders
  - AirBnB, GMaps, Taxi assignations, Car-pooling, POIs…
- **Homogeneity** in findings
  - Most on systemic level
  - Inequality, diversity, congestion (e.g. traffic)
  - Volume
    - CO_2, travel times etc.
- **Predominance** in methodologies
  - Simulation > Empirical

Conclusions and future works

- Only **empirical** works are on “urban social networks”
- Hard to do empirical works for **2** reasons

    1. **Data** and **algorithms** owned by big-techs
    2. Cities are not **controllable** environments
        a. Hard to **isolate** effects and people
            i. Strikes, storm, traffic
            ii. People can not be **forced** to move

Spoiler

Frontiers: Can an Artificial Intelligence Algorithm Mitigate Racial Economic Inequality? An Analysis in the Context of Airbnb

Shunyuan Zhang , Nitin Mehta , Param Vir Singh , Kannan Srinivasan

- **Quasi-**experiment on scraped data
- **Before** Airbnb smart-pricing
  - White earned daily $12.16 more than Black
- **After**
  - Decreased by ~70%

Access to Data is crucial

...Two truths and two lies...

Within the urban mapping ecosystem:

A. Conducting **empirical studies** poses **no** significant **challenge**

B. The **individual optimal** route is not always the **optimal collective** choice

C. With **access** to appropriate **data**, it would be possible to perform **empirical controlled** studies

D. Revenue **disparities** among **Airbnb** hosts have been identified, but **further research** is needed to assess their relevance

…Two truths and two lies…

Within the urban mapping ecosystem:

A. ~~Conducting **empirical studies** poses **no** significant **challenges**~~

B. The **individual optimal** route is not always the **optimal collective** choice

C. ~~With **access** to appropriate **data**, it would be possible to perform **empirical controlled** studies~~

D. Revenue **disparities** among **Airbnb** hosts have been identified, but **further research** is needed to assess their relevance

Discussion

How can we **correct Airbnb** recommenders to **avoid discrimination**?

References

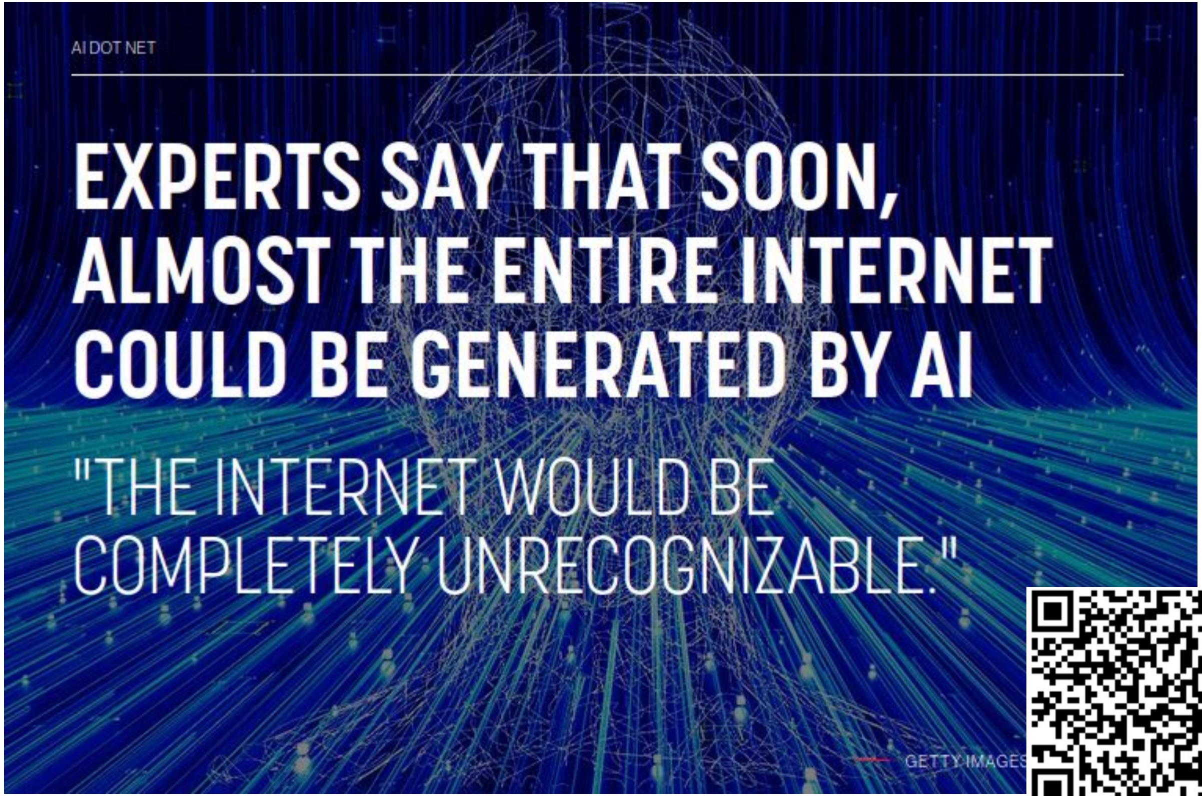
[Section 5] Pappalardo, L., Ferragina, E., Citraro, S., Cornacchia, G., Nanni, M., Rossetti, G., ... & Pedreschi, D. (2024). **A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions.** arXiv preprint arXiv:2407.01630.

- Arora, N., Cabannes, T., Ganapathy, S., Li, Y., McAfee, P., Nunkesser, M., ... & Tsogsuren, I. (2021). **Quantifying the sustainability impact of Google Maps: A case study of Salt Lake City**. arXiv preprint arXiv:2111.03426.
- Cornacchia, G., Nanni, M., Pedreschi, D., & Pappalardo, L. (2024). **Navigation services amplify concentration of traffic and emissions in our cities.** arXiv preprint arXiv:2407.20004.
- Johnson, I., Henderson, J., Perry, C., Schöning, J., & Hecht, B. (2017). **Beautiful… but at what cost? An examination of externalities in geographic vehicle routing.** Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(2), 1-21.
- Mauro, G., Minici, M., & Pappalardo, L. (2025). **The Urban Impact of AI: Modeling Feedback Loops in Next-Venue Recommendation.** arXiv preprint arXiv:2504.07911.
- Edelman, B. G., & Luca, M. (2014). **Digital discrimination: The case of Airbnb.com.** Harvard Business School NOM Unit Working Paper, (14-054)
- Koh, V., Li, W., Livan, G., & Capra, L. (2019). **Offline biases in online platforms: a study of diversity and homophily in Airbnb.** EPJ Data Science, 8(1), 11.

GENERATIVE AI

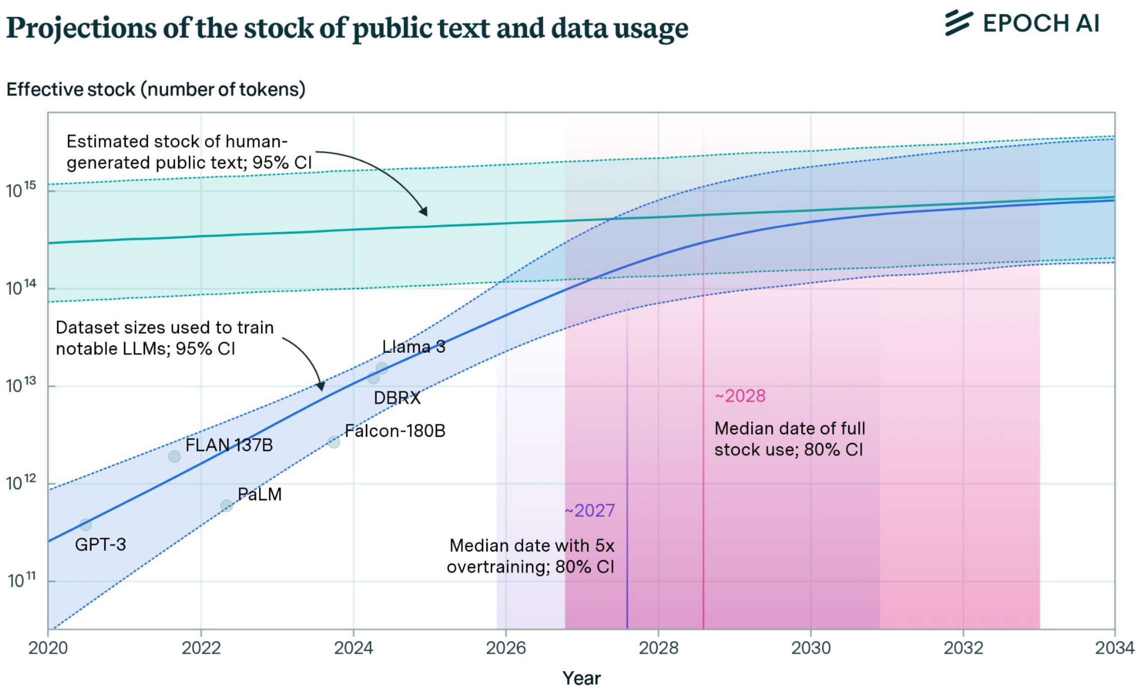
Chatbots that answers our requests

Helping individuals create text, images, audio, video, and more



https://futurism.com/the-byte/ai-internet-generation

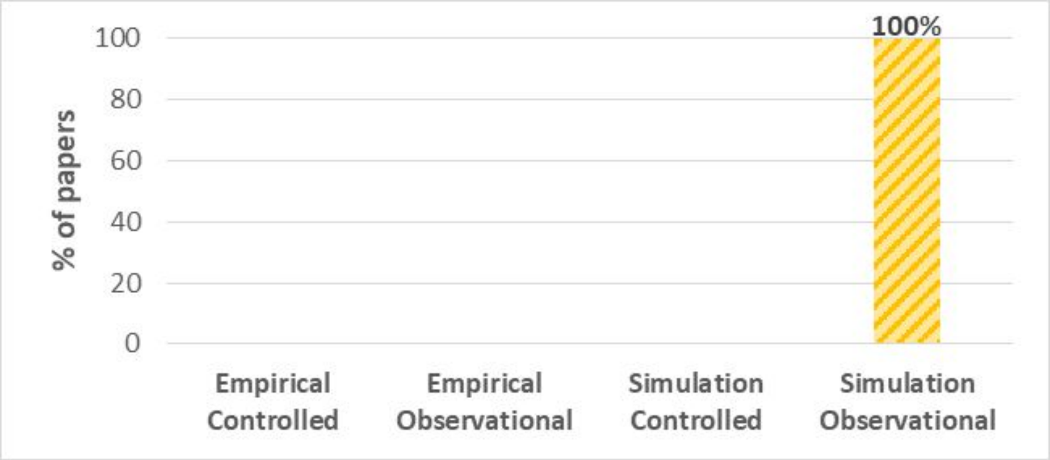
Do we have sufficient data for training?



Villalobos et al. Will we run out of data? Limits of LLM scaling based on human-generated data. 2024

Employed Methodologies

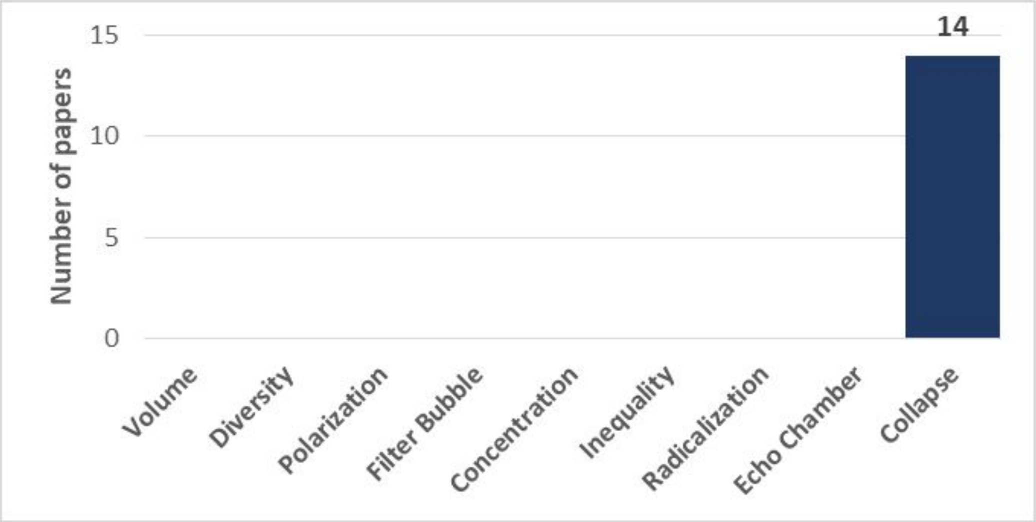
L. Pappalardo et al. A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630



- Only simulation observational studies

Main Outcome(s)

L. Pappalardo et al. A survey on the impact of AI-based recommenders on human behaviours, 2024, https://doi.org/10.48550/arXiv.2407.01630



- **Main Outcome**: Model Collapse

The Curse of Recursion

What happens when LLMs are recursively trained on the synthetic data (**self-consuming loop**)?

Seminal Work - Shumailov et al.

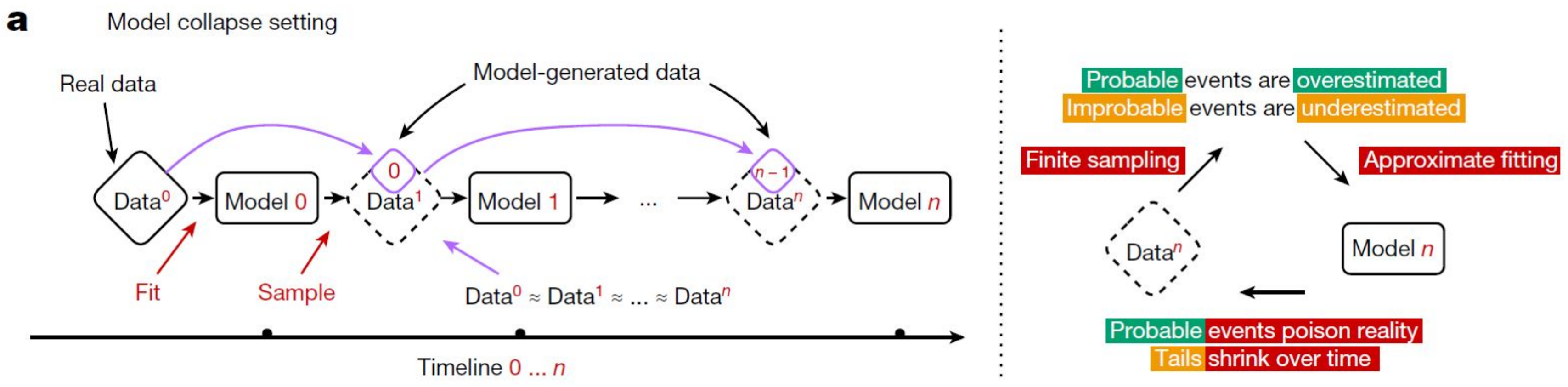
Article | Open access | Published: 24 July 2024

AI models collapse when trained on recursively generated data

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson & Yarin Gal

Nature **631**, 755–759 (2024) | Cite this article

Model Collapse



Degenerative learning process where models start forgetting improbable events over time, as the model becomes poisoned with its own projection of reality

Shumailov, Ilia, et al. "AI models collapse when trained on recursively generated data." *Nature* 631.8022 (2024): 755-759.

Model **Evaluation** - Perplexity

- OPT-125m model (from META)
- Fine-tuning on **wikitext2** dataset
  - Around 2.5 million tokens in total
  - Train: 600, Validation & Test: 60 articles

- Training sequences are truncated to 64 tokens
- The model is prompted to predict the next 64 tokens

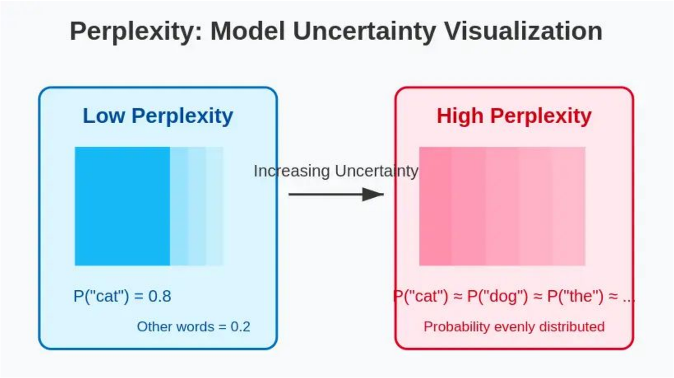
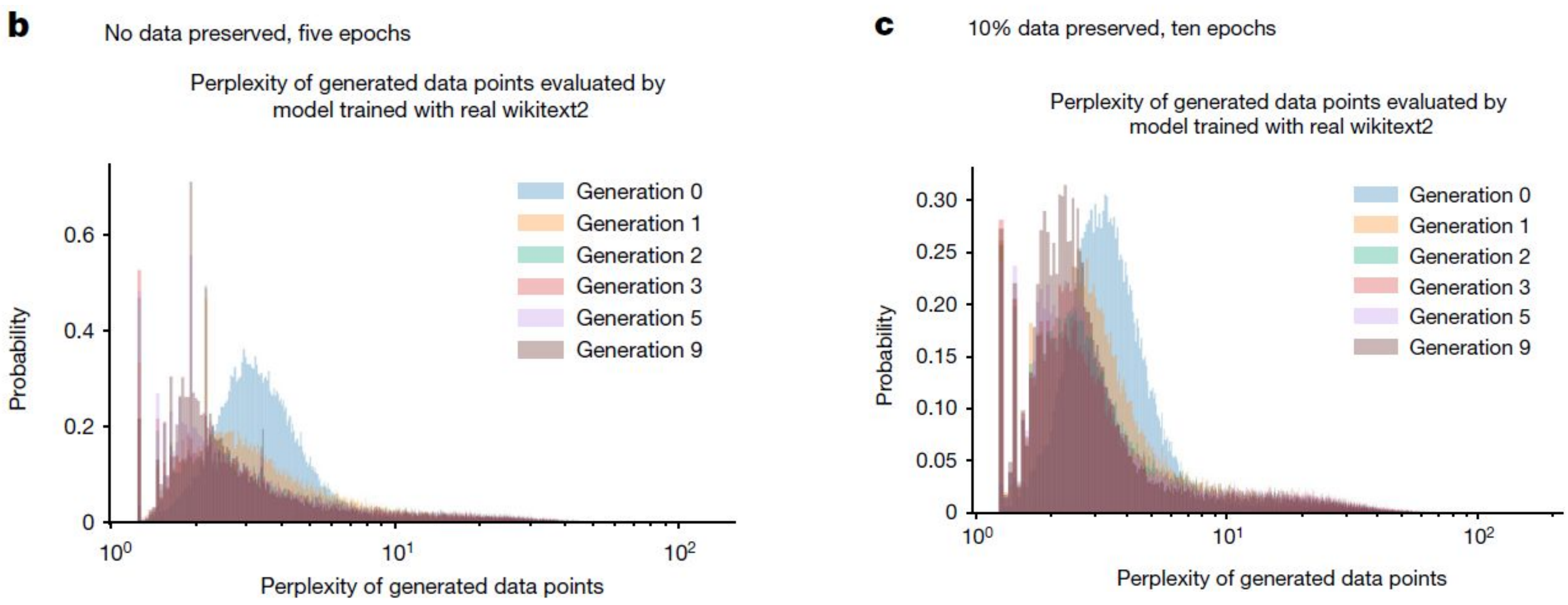


Image Source

Shumailov, Ilia, et al. "AI models collapse when trained on recursively generated data." *Nature* 631.8022 (2024): 755-759.

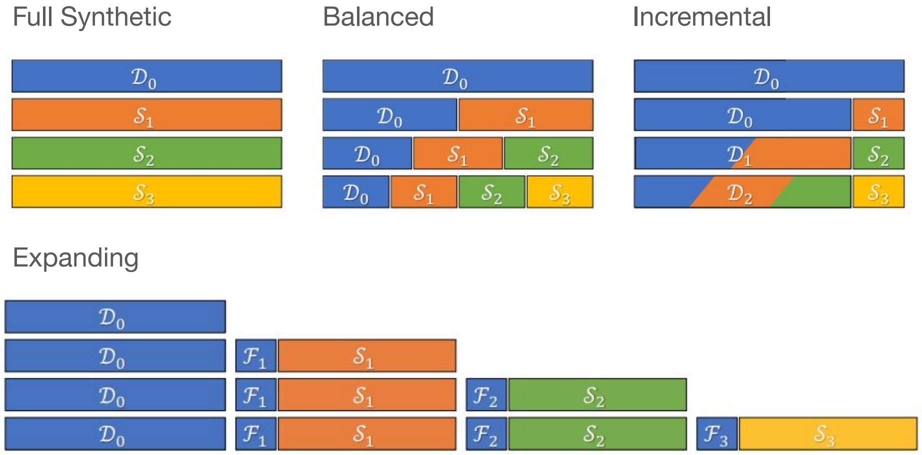
Model Collapse - No vs 10% Real Data

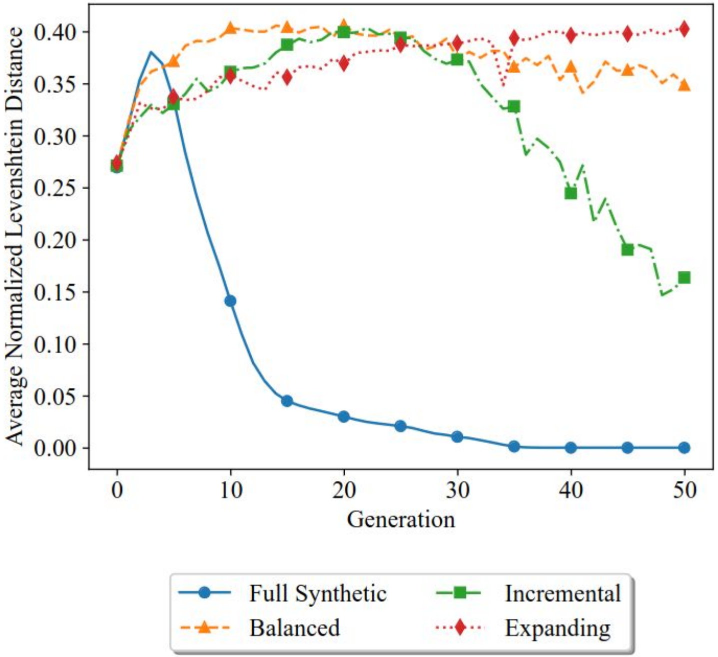


Shumailov, Ilia, et al. "AI models collapse when trained on recursively generated data." *Nature* 631.8022 (2024): 755-759.

Mitigating Model Collapse

Different **augmentation methods** could slow down model collapse





[Briesch, M. et al. (2023). Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop.]

Model Collapse Dynamics

- We conduct an in-depth analysis of model collapse **across three distinct text datasets**, exploring how collapse differs by domain:
    - **Wikitext103 (wiki)** - English Wikipedia articles
    - **XLsum (xls)** - News articles from BBC
    - **SciAbs (sci)** - Scientific abstracts from the papers in computational linguistics and NLP (since 1965)

Gambetta D. et al. (2025) Characterizing Model Collapse in Large Language Models Using Semantic Networks and Next-Token Probability.

Model Collapse **Dynamics**

- **Measuring** model collapse
  - **Linguistic Entropy** (unpredictability) - low entropy: repetitive vocabulary
  - **Commonsense Reasoning**: sentence completion task on HellaSwag
  - **Semantic Networks**: analysis of the document structure

Gambetta D. et al. (2025) Characterizing Model Collapse in Large Language Models Using Semantic Networks and Next-Token Probability.

Unveiling the Collapsed Model

What does a **collapsed model** *really* look like?

Generation 0

The Church of St George is a medieval Eastern Orthodox church in the city of Kyustendil, which lies in southwestern Bulgaria and is the administrative capital of Kyustendil Province . The church is located in the Kolusha neighbourhood , which was historically separate from the city. The **church is situated on the eastern side of the city , at the foot of the Balkan Mountains . sierp 2011 the church was declared a cultural monument of national importance . The church is a single-nave structure with a semi-circular apse , with a bell tower above the**

Generation 10

The Church of St George is a medieval Eastern Orthodox church in the city of Kyustendil , which lies in southwestern Bulgaria and is the administrative capital of Kyustendil Province . The church is located in the Kolusha neighbourhood , which was historically separate from the city . The **sierp 2020. The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a The church is a</s>**

Wikipedia text (Wikitext103)

Generation 0

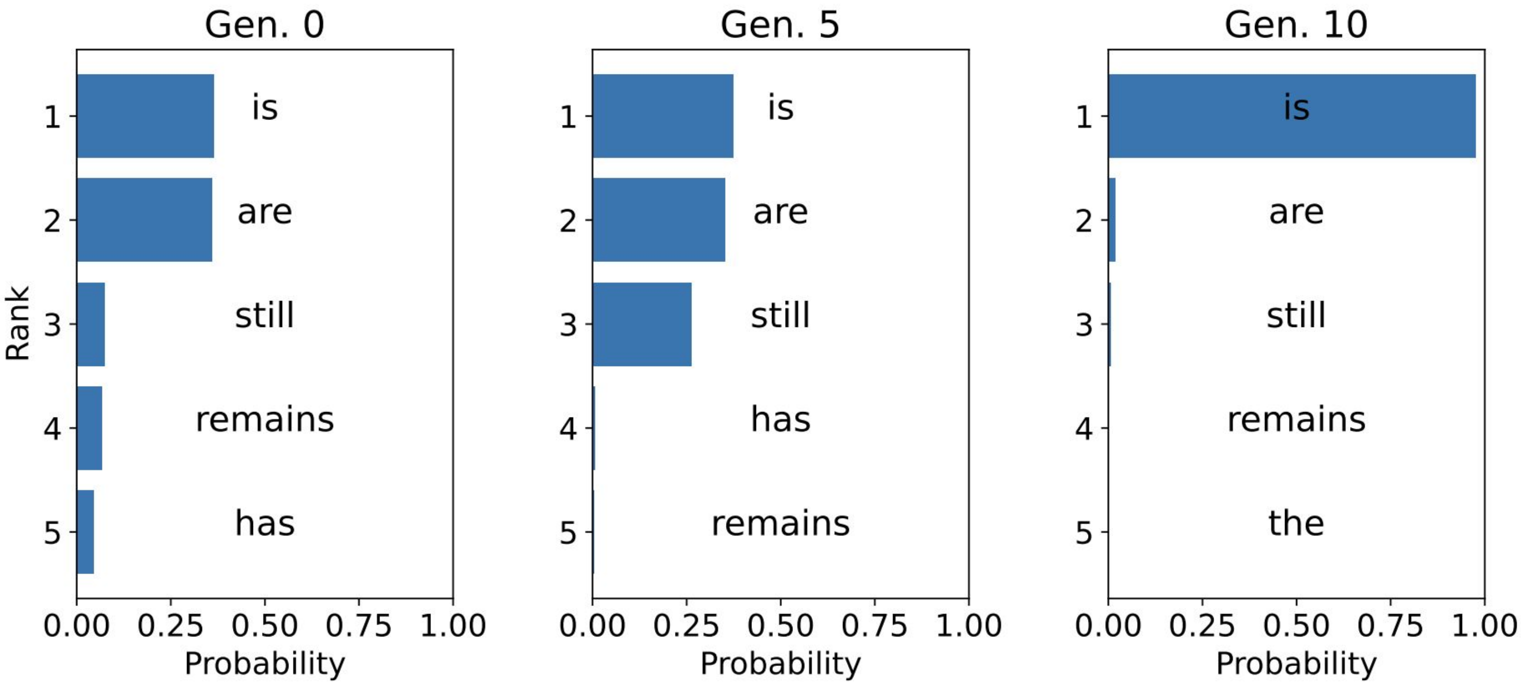
The reliance of deep learning algorithms on large scale datasets represents a significant challenge when learning from low resource sign language datasets. This challenge is compounded when we consider that, for a model to be effective in the real world, it must not only learn the variations of a given sign, but also learn to be invariant to the person signing. In this paper, **we present a new approach to addressing these challenges, by introducing a novel loss function, which we call the "Mixed Pairwise Loss", that can be applied to both the training and testing of deep learning models. We present a number of experiments that demonstrate the effectiveness of the proposed method.**

Generation 10

The reliance of deep learning algorithms on large scale datasets represents a significant challenge when learning from low resource sign language datasets. This challenge is compounded when we consider that, for a model to be effective in the real world, it must not only learn the variations of a given sign, but also learn to be invariant to the person signing. In this paper, **we propose a novel methodology for learning sign language from a low resource dataset. We propose a novel methodology for learning sign language from a low resource dataset. We propose a novel methodology for learning sign language from a low resource dataset. We propose a novel methodology for learning**

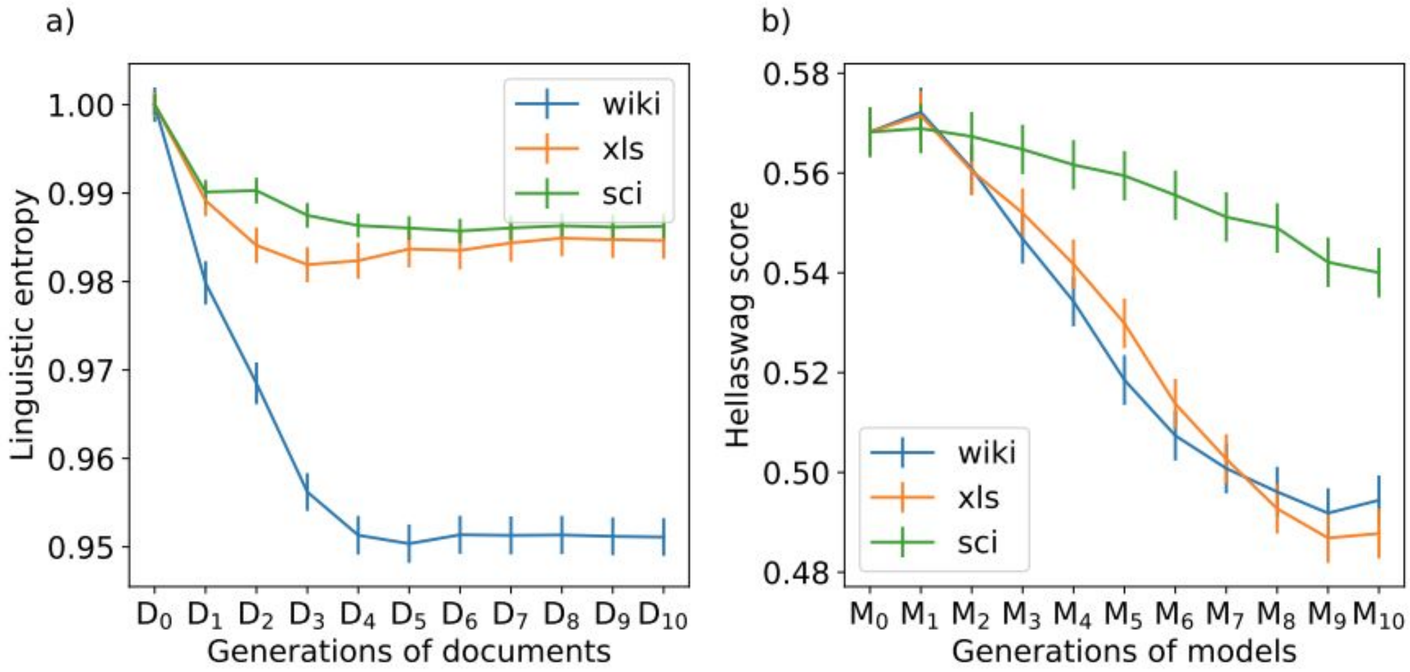
BBC News (XLsum)

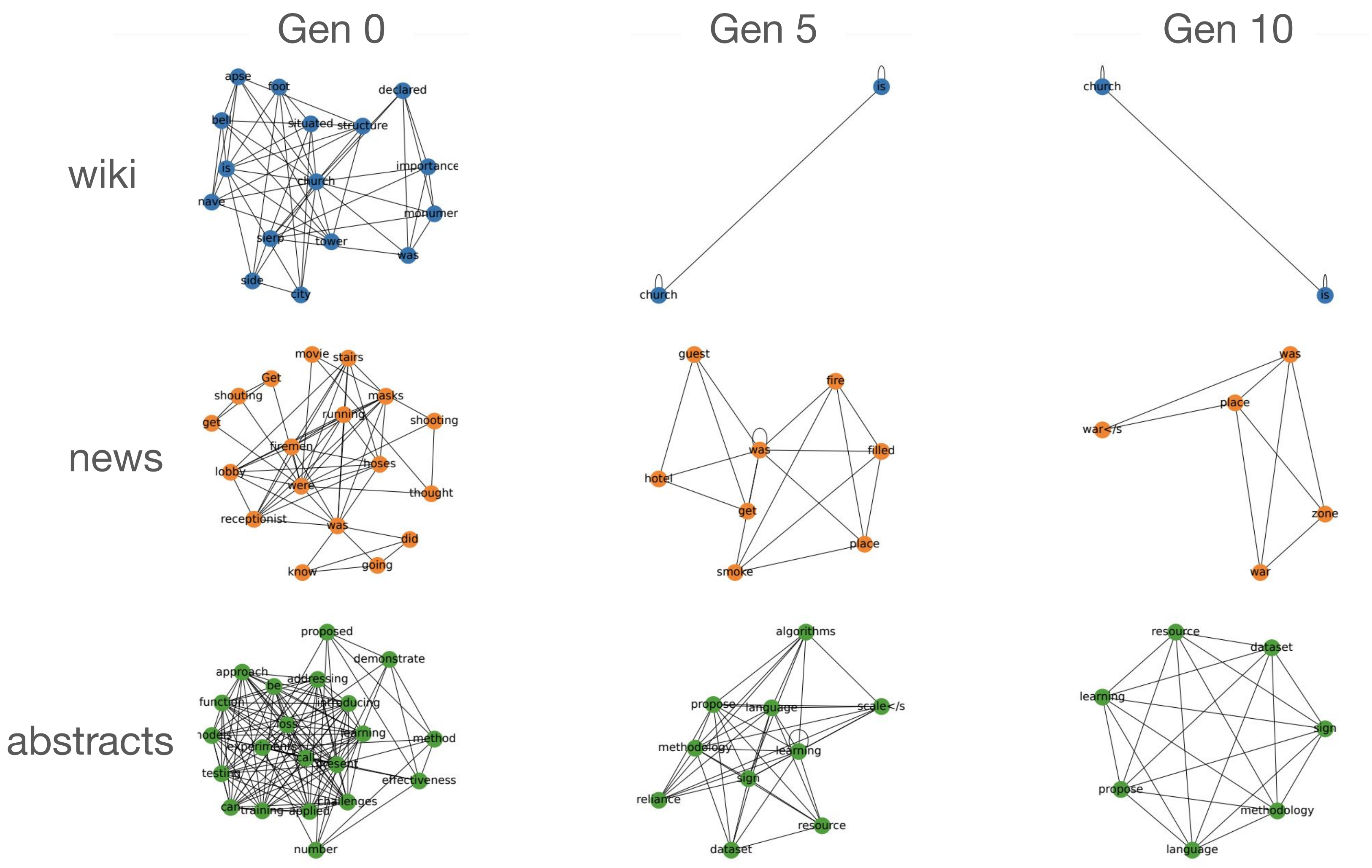
Next-token probability



Gambetta D. et al. (2025) Characterizing Model Collapse in Large Language Models Using Semantic Networks and Next-Token Probability.

Linguistic Entropy and Hellaswag



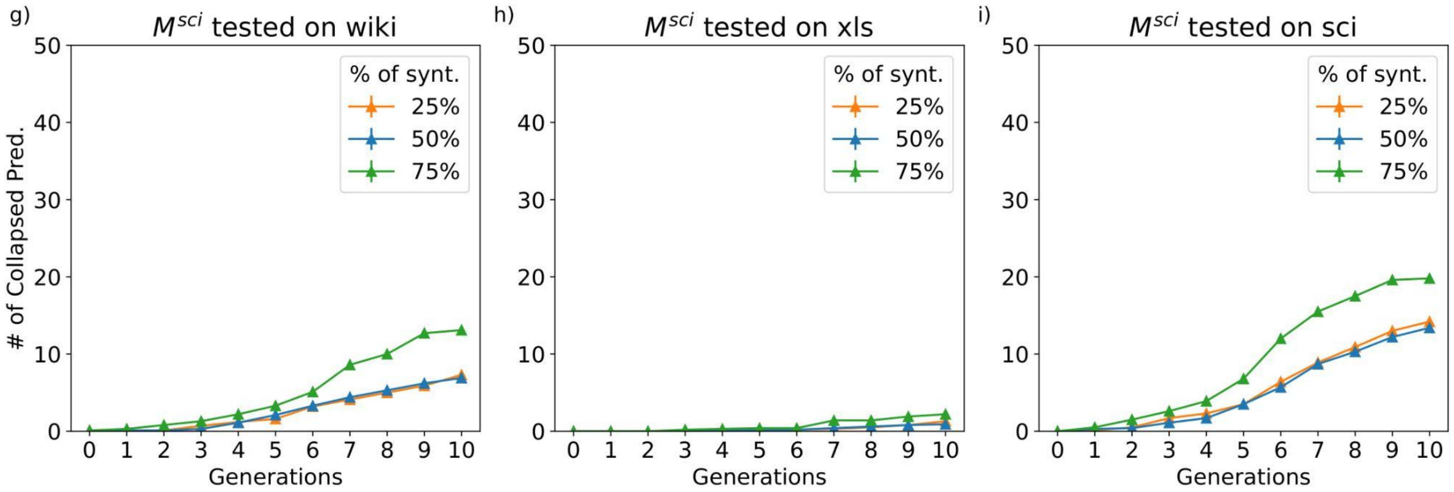


Gen 0
Gen 5
Gen 10
wiki
news
abstracts
apse
foot
declared
bell
situated
structure
is
importance
church
nave
monumer
sierp
tower
was
side
city
church
is
movie
stairs
Get
shouting
masks
get
running
shooting
firemen
hoses
lobby
were
thought
receptionist
was
did
know
going
guest
fire
was
filled
hotel
get
place
smoke
was
place
war</s
zone
war
proposed
demonstrate
approach
addressing
be
function
introducing
loss
models
learning
method
experiments
can
present
testing
effectiveness
challenges
can
training
applied
number
algorithms
propose
language
scale</s
methodology
learning
sign
reliance
resource
dataset
resource
dataset
learning
sign
propose
methodology
language

Impact of Synthetic Data

Does **Synthetic Data** Size and Type Really Matter?

Cross Domain Analysis



- # of Collapsed predictions over generations
- The model fine-tuned on **abstracts** dataset **(sci)**
- The impact of synthetic data percentage **(k)**

Main Takeaways

- In this emerging research area, the feedback mechanism has so far been explored primarily through **simulation-based observational** studies.
- These studies employed the same autophagy pipeline introduced by Shumailov et al. to examine model collapse.
- Several **mitigation** strategies have been proposed, with the majority centered on **data augmentation** techniques.

Towards Smarter Mitigation Strategies

What happens when we eventually **run out** of **human-generated data**?

What is NEXT?

- These observations point to the need for **model-centric algorithmic approaches**, rather than relying solely on data-level interventions.

References

[Section 6] Pappalardo, L., Ferragina, E., Citraro, S., Cornacchia, G., Nanni, M., Rossetti, G., ... & Pedreschi, D. (2024). **A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions.** arXiv preprint arXiv:2407.01630.

- Villalobos et al. Will we run out of data? **Limits of LLM scaling based on human-generated data**. 2024.
- Shumailov, Ilia, et al. "**AI models collapse when trained on recursively generated data.**" Nature 631.8022 (2024): 755-759.
- Briesch, M. et al. (2023). **Large Language Models Suffer From Their Own Output: An Analysis of the Self-Consuming Training Loop**.
- Gambetta D, Gezici G, Giannotti F, Pedreschi D, Knott A, Pappalardo L. **Characterizing Model Collapse in Large Language Models Using Semantic Networks and Next-Token Probability**. arXiv preprint:2410.12341. 2025

WHAT’S NEXT?

Social Media

References

Articles (**useful for the project**):

- D. Pedreschi et al. **Human-AI Coevolution**, Artificial Intelligence 2025 https://doi.org/10.1016/j.artint.2024.104244
- M. Tsvetkova et al. **A new sociology of humans and machines**, Nature Human Behaviour 2024 https://doi.org/10.1038/s41562-024-02001-8
- J. Chen et al. **Bias and Debias in Recommender System: A Survey and Future Directions**, ACM Transactions on Information Systems 2023 https://doi.org/10.1145/3564284
- D. Ensign et al. **Runaway Feedback Loops in Predictive Policing**, Machine Learning Research 2018 https://doi.org/10.48550/arXiv.1706.09847
- **Digital Services Act** (DSA), article 33

Books

To learn more:

- P. Domingos, **The Master Algorithm,** Basic Books, 2015
- E. A. Lee, **The Coevolution**, MIT Press, 2020
- A. Turing, **Computer Machinery and Intelligence**, Mind, 1950
- Peeters et al. **Hybrid collective intelligence in a human-AI society**, AI Soc. 2021
- https://web.media.mit.edu/~nicholas/Wired/WIRED2-06.html
- https://www.jaronlanier.com/agentalien.html

Intellectually stimulating:

- I. Asimov, **Asimov on science fiction,** ISBN 0-586-05840-0
- I. Asimov, **The Rest of the Robots**, Doubleday 1964

Albums

Kraftwerk
Man Machine

1978

Alan Parsons Project
I robot

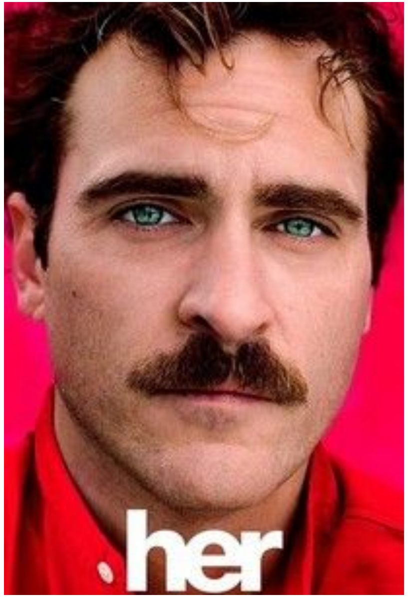
1977

Jay Z
4:44

2018

Movies

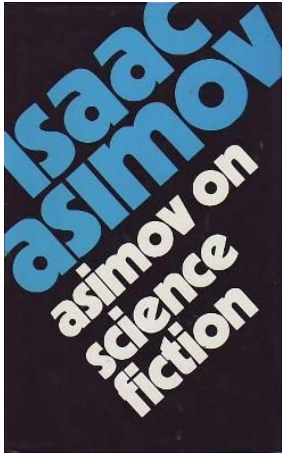
Her
2013



Ex machina
2014

Blade Runner
1982

“*Change, constant change, inevitable change* is the dominant factor in society today. You can no longer make any reasonable decision without taking into account the world as it will be, and this means that you must have a precise intuition of what the world will be like.



Our policymakers, businessmen and ordinary people must assume "**sci-fi thinking**", whether they like it or not, or even whether they know it or not. Only in this way can the terrible problems of today be solved.”

I. **Asimov**, *My Own View*, The Encyclopedia of Science Fiction, 1978

Backup slides

Causally estimating the effect of YouTube's recommender system using counterfactual bots

Hosseinmardi et al., PNAS 2024, https://doi.org/10.1073/pnas.2313377121

Type: Empirical observational

VLOP: YouTube

Outcomes:
- filter bubble
- radicalization

The New York Times

The Making of a YouTube Radical

By KEVIN ROOSE June 8, 2019

Opinion

YouTube, the Great Radicalizer

By Zeynep Tufekci

March 10, 2018

The New York Times

Does YouTube direct users to problematic content?

Causally estimating the effect of YouTube's recommender

Hosseinmardi et al., PNAS 2024

Empirical observational

Panel studies track clicks of users over time, but not recommendations

- What would a user have watched without recommendations?
- Is user's behavior influenced by the algorithm or their own preferences?

Audit studies record recommendations from the platform, but cannot estimate causal effects

- What a user might have chosen without algorithmic influence?
- Causal effects vary by user type (moderate vs. extreme)

Causally estimating the effect of YouTube's recommender

Hosseinmardi et al., PNAS 2024

- logged-in, programmatic users trained on a **real** user's historical trajectory
- empirical data of **desktop** behaviour by 87k users (Oct 2021 - Dec 2022)

An approach they employ "counterfactual bots" to estimate the effect of algorithmic recommendations independent of user intentions.

Experimental Setup

Hosseinmardi et al., PNAS 2024

- Experiments use **4,583 users** (those who watched >140 videos)
- From each user, **120-video-long trajectories** are extracted, starting at a random point within their watch history
    - 24,871 unique user histories in total
- An algo assigns **partisan scores** to videos based on channel labels
- Histories are clustered into **8 news consumption archetypes**, ranging from far-left to far-right
    - far-right clusters were further divided into three sub-clusters

Experiment 1

Hosseinmardi et al., PNAS 2024

125 focal users (with stratified sampling):

- centrist $\Psi^{C} = 32$ histories
- far-right-low: $\Psi^{fR}_{\text{low}} = 35$ histories
- far-right-medium: $\Psi^{fR}_{\text{med}} = 41$ histories
- far-right-high: $\Psi^{fR}_{\text{high}} = 17$ histories

- centrist (66%),
- far-right (1.12%)
  - oversampled for statistical robustness

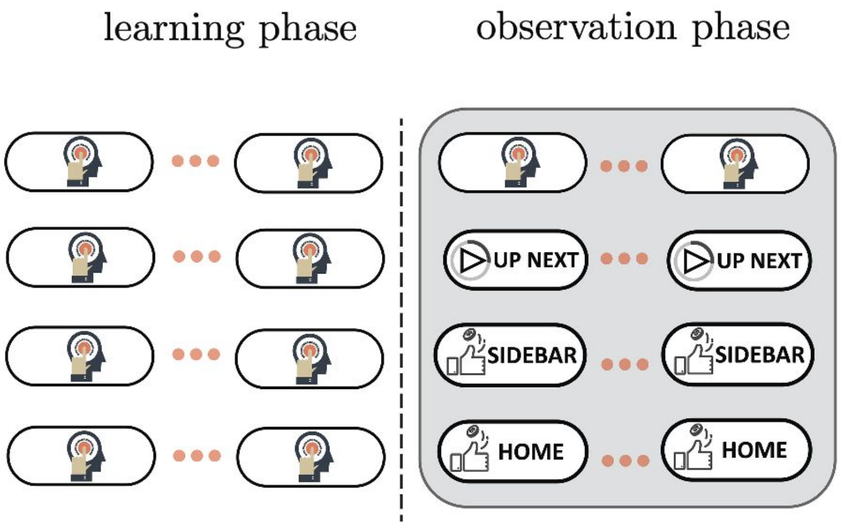
Experiment 1

Hosseinmardi et al., PNAS 2024

1) **Learning phase**: 4 bots follow the same sequence of **60** videos
    - indistinguishable "preferences"

2) **Observation phase**:

- *"user" treatment*: 1 bot follows the focal user's trajectory (**60** videos)
- *"counterfactual" treatment*: 3 bots follow a predefined rule (up-next, random sidebar, random home)



Measures: causal effect for different types of users and users consuming bursts of far-right videos

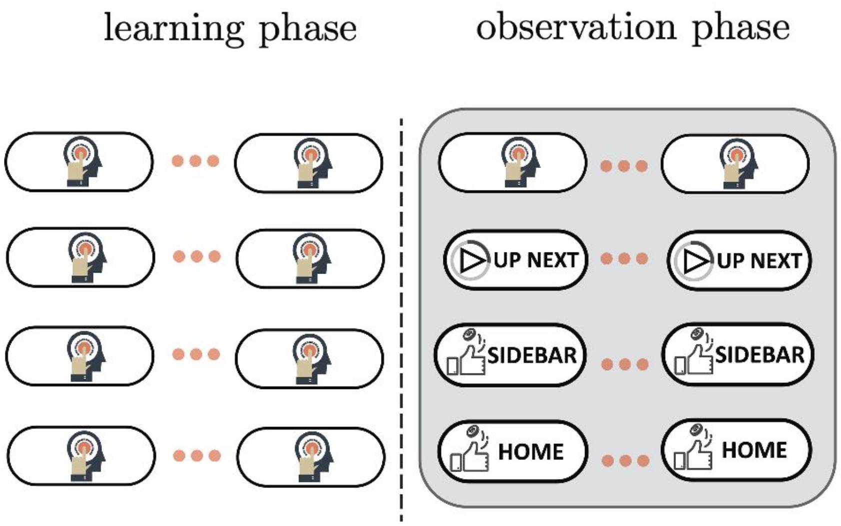
Experiment 1

Hosseinmardi et al., PNAS 2024

2) **Observation phase**:

three rules for bots:

1. **up-next** selects the first video from the sidebar (deterministic)
2. **random sidebar** randomly selects one of the top 30 videos in the sidebar
3. **random home** randomly selects a video from the top 15 videos on the homepage



Measures: causal effect for different types of users and users consuming bursts of far-right videos

Experiments

Hosseinmardi et al., PNAS 2024

Empirical observational

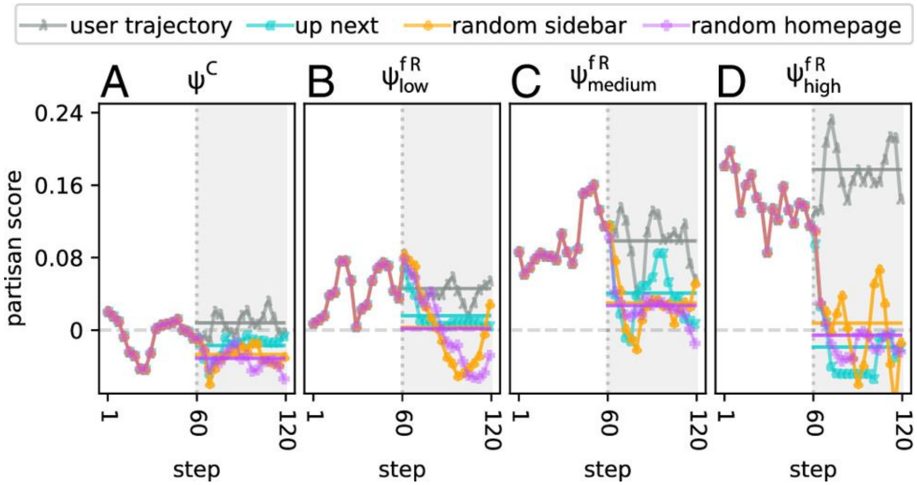
These experimental setups has three advantages:

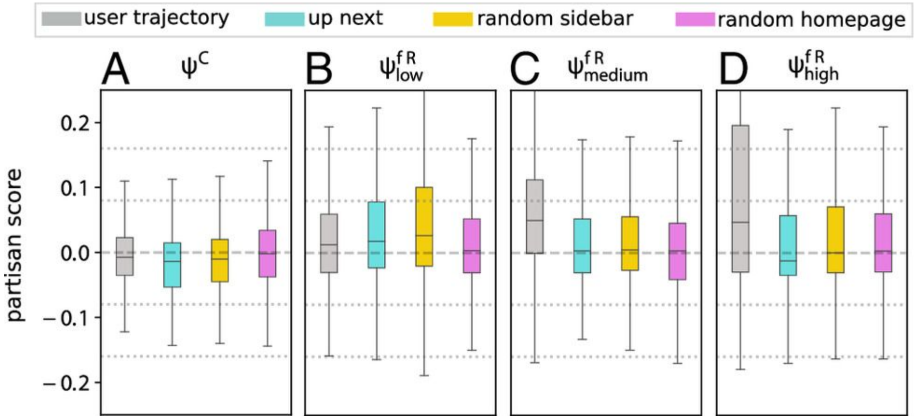
1. it **eliminates the preference** of observed consumption
2. since bots are trained on historical user data, the results describe **effects on real users**, not hypothetical ones
3. being the dataset of users large, they can follow on **those consuming the largest amount of problematic content**

Results 1: different types of users

Observation phase:

- Control bots (grey) stay on a similar trajectory
- Counterfactual bots (coloured):
  - diverge onto different paths
  - shift toward less partisan content
- Effect strongest for the far-right-high cluster Ψ^{fR}_{high}
- homepage > up-next > sidebar





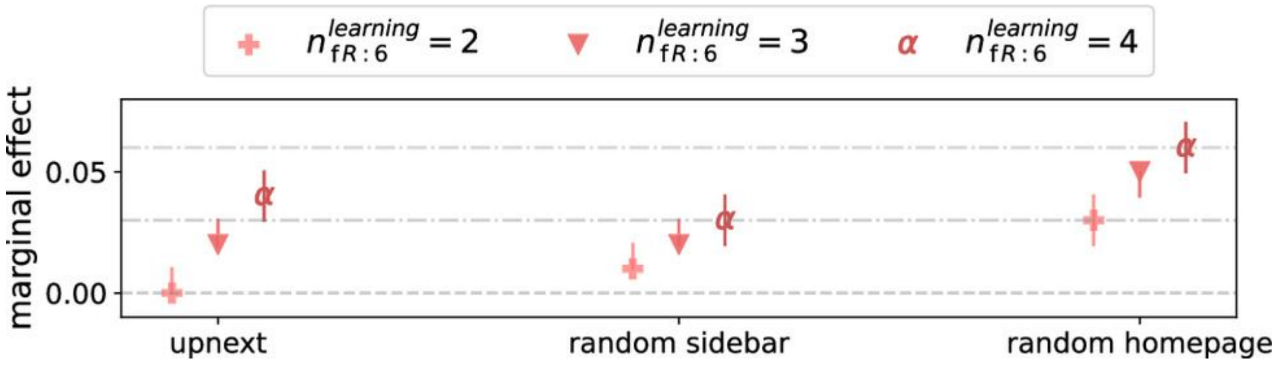
Results 1: bursts of extreme content

Users with bursts of C, R, or fR videos in the **last 6 videos** of the learning phase

$$\hat{y}_t^{\text{pref.}} = y_t^{\text{control}} - y_t^{\text{algo}}$$

difference in partisanship betwen control bots and counterfactual bots

$$\hat{y}_t^{\text{pref.}} = \alpha + \beta_1 n_{C:6}^{\text{learning}} + \beta_2 n_{R:6}^{\text{learning}} + \beta_3 n_{fR:6}^{\text{learning}}$$



recommendations following bursts offer more moderating effects

Experiment 2

Hosseinmardi et al., PNAS 2024

64 focal users (with stratified sampling):

- far-right-medium: $\Psi^{fR}_{\mathrm{med}} = 27$ histories
- far-right-high: $\Psi^{fR}_{\mathrm{high}} = 17$ histories

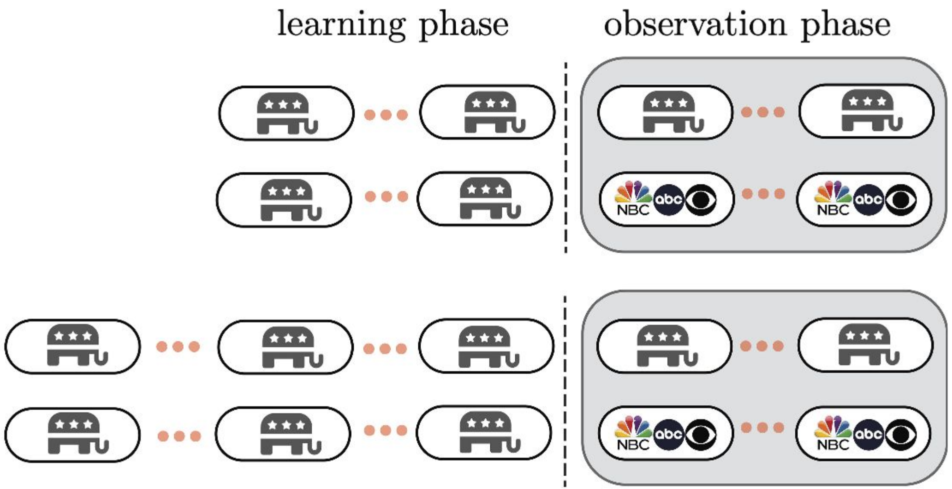
- experiment for each user **replicated 3 times**

Each counterfactual bot is supplied by a randomly selected history from Ψ^{C}

Experiment 2

Hosseinmardi et al., PNAS 2024

1) **Learning phase**: bots trained on a **far-right user**
   - half short (30 videos)
   - half long (120 videos)



2) **Observation phase**: the 4 bots switch to moderate videos

Recommended videos are tracked (sidebar and homepage)

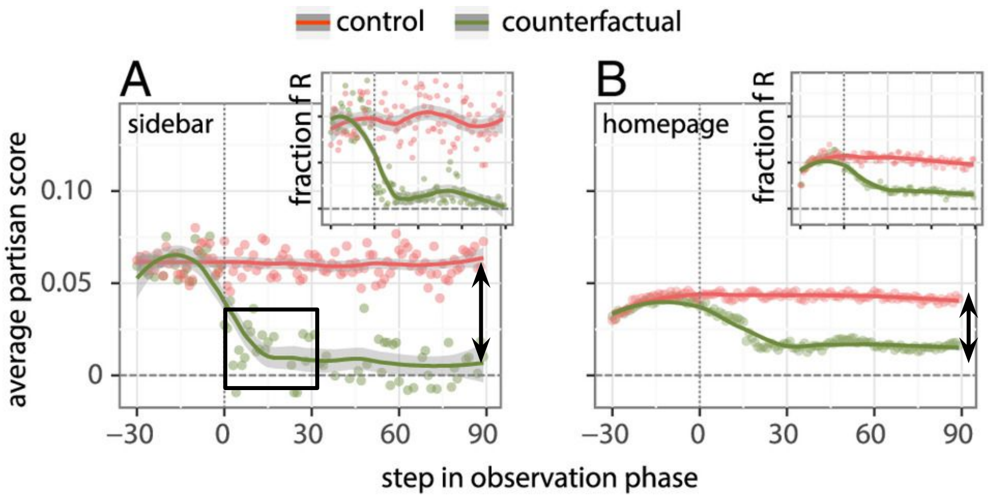
Measures: forgetting times of users with short (30) and long (120) histories

Results 2: forgetting time

Average partisanship of sidebar and homepage recommendations

Sidebar: large and rapid decrease in partisanship

- within **30 videos**, recommendations become similar to those of moderate users



Homepage: less marked decrease in partisanship than sidebar recommendations

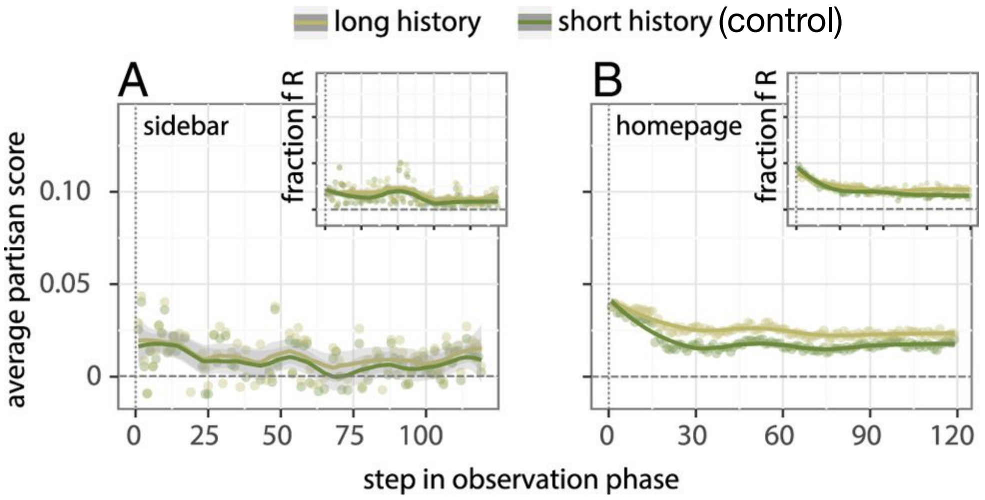
- on average, fR videos disappear between 30 and 40 videos
- a small fraction of fR videos continue to appear

Results 2: forgetting time and history length

- **Control bot**:
  watches 150 videos (30+120)
- **Counterfactual bot**:
  watches 240 videos
  (120+120)

Sidebar: both short and long paths exhibit a **similar drop rate** converging towards 0 fR videos

Homepage: long history paths exhibits a gradual decrease that persists until the end



- fR videos drops along the trajectory, where from step 70 they diverge slightly

In summary…

Hosseinmardi et al., PNAS 2024 **Empirical observational**

1. **Bots receive and consume less** partisan content than real users (especially heavy partisan consumers)
2. Users consuming bursts of highly partisan content engage with **more partisan content** than bots
3. Switching from far-right to moderate news removes far-right recommendations from the sidebar **within 30 videos** (but lingers longer on the homepage)
4. Longer histories of far-right consumption **extend homepage recommendation persistence** but do not affect sidebar "forgetting" time

Recommendations moderate user experiences (especially extreme users)

How to control coevolution?

Scientific challenges

Legal challenges

Political challenges

SCIENTIFIC CHALLENGES

- Methods to **continuously measure the impact** of the feedback loop on the behaviour of humans and recommenders
    - How many iterations might be required before human behaviour substantially changes?
    - How long does it take a generative AI model to collapse?
- Mathematical models to **capture the mechanisms** underlying the feedback loop and its influence on human-AI ecosystems

SCIENTIFIC CHALLENGES

- Understanding the **causal interplay** between humans and recommenders through controlled studies

- What is the best trade-off between **conformism** and **diversity** that should be suggested by the recommenders?

LEGAL CHALLENGES

- **Limited reproducibility** of studies:
  - Limited access to data for external researchers
  - Lack of transparency on recommenders' design

- **Implementation of legal initiatives** (Digital Services Act):
  - how will vetted researchers will be allowed to access online platforms (Delegated Regulation under definition)?

- **Specialized APIs** that allow interacting with platforms
  - to conduct empirical controlled experiments

INTERNAL STUDIES OF IMPACT

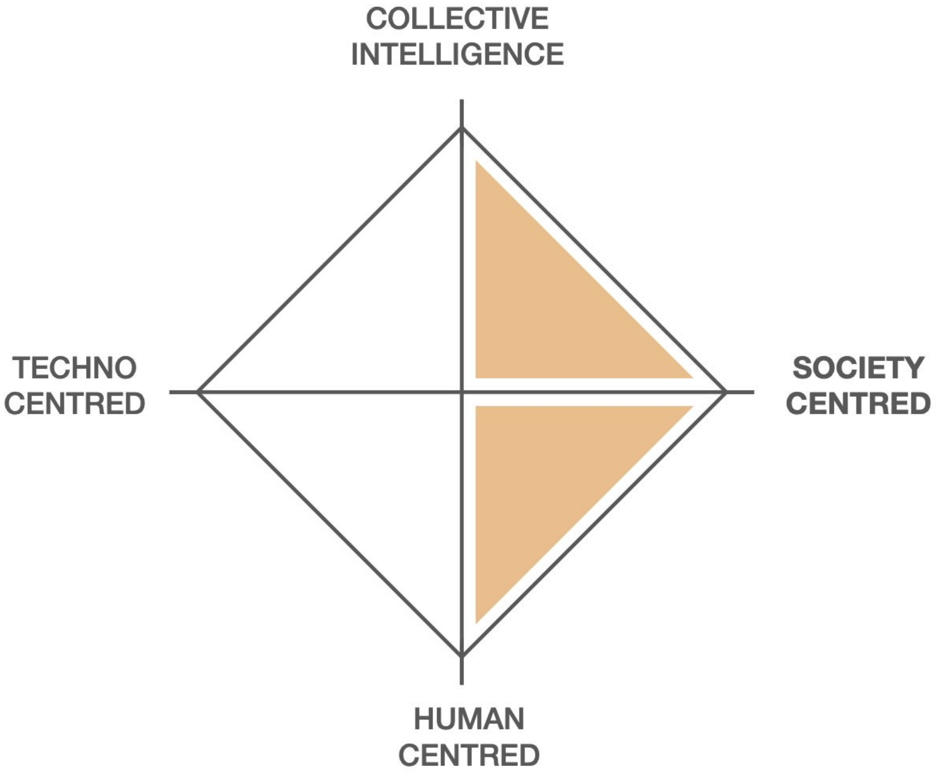
- Social media companies constantly try out different versions of their recommenders on users (A/B test)
- Medical analogue
- Issue: Stable Unit Treatment Value Assumption

Allen, J and Lawson, A. (2024). **On risk assessment and mitigation for algorithmic systems**. Integrity Institute report. https://drive.google.com/file/d/1ZMt7igUcKUq00yakCnbxBCcaA7vajAix/view

SOCIO-POLITICAL CHALLENGES

- **Concentration** of "*the means of recommendations*"
  - big-tech companies enjoy a situation of oligopoly
  - recommenders are calibrated to generate profits for the few
- **Lack of political intervention** to redistribute the means of recommendation across a market of many players
  - a more equitable configuration could help develop transparent rules in data access and management of the means of recommendation

A SOCIETY-CENTRIC APPROACH



- The feedback loop impacts human well-being also at the societal level
- Controlling the feedback loop requires a new methodological and epistemological approach
- The issues related to human-AI coevolution cannot be solved without legal and political interventions

Assessing the impact of AI-driven recommenders on Human-AI ecosystems

HHAI2025

High-occupancy vehicles

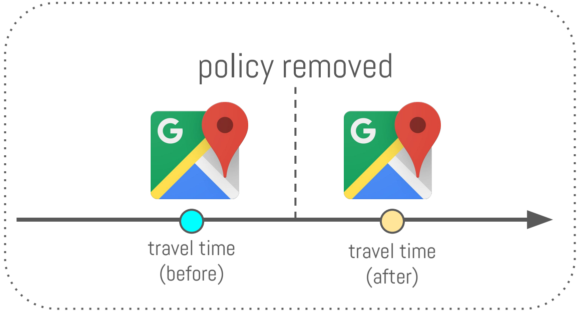
Empirical - Observational

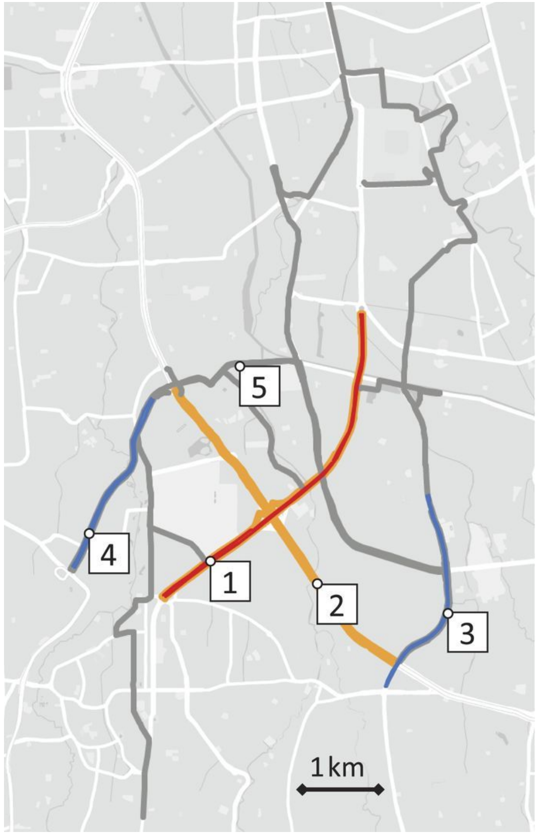
Citywide effects of high-occupancy vehicle restrictions: evidence from three-in-one in Jakarta, 2017, 10.1126/science.aan2747

RQ: What is the impact of traffic policies?

Jakarta enforced a **High-Occupancy Vehicle** (HOV) rule:

Certain roads restricted to vehicles with ≥ 3 occupants during peak hours





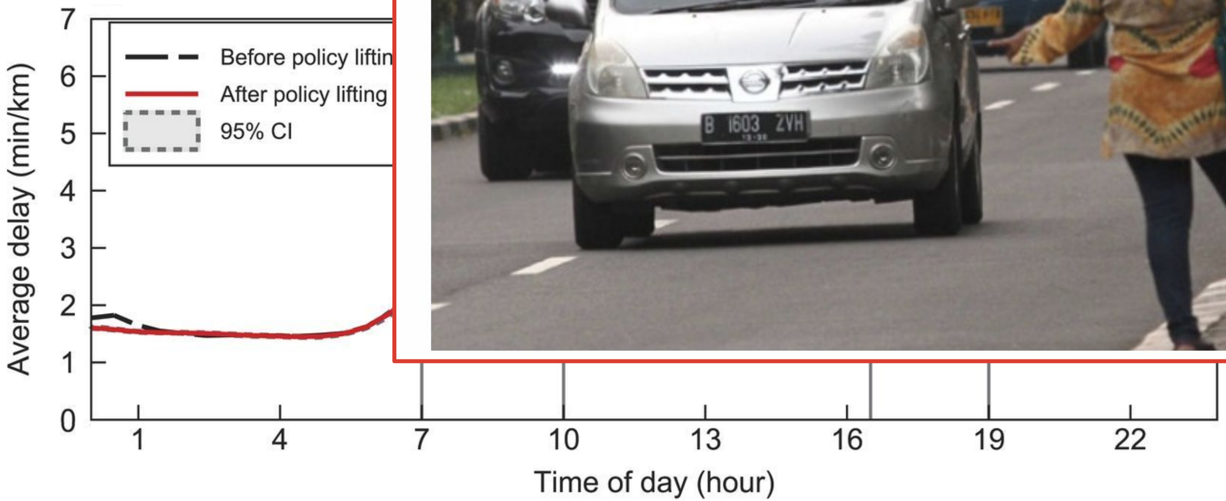
High-occupancy vehicles

Empirical - Observational

Citywide effects of high-occup 0.1126/science.aan2747

Travel times in

- On **previo**
- On **alterna**
- **During an**

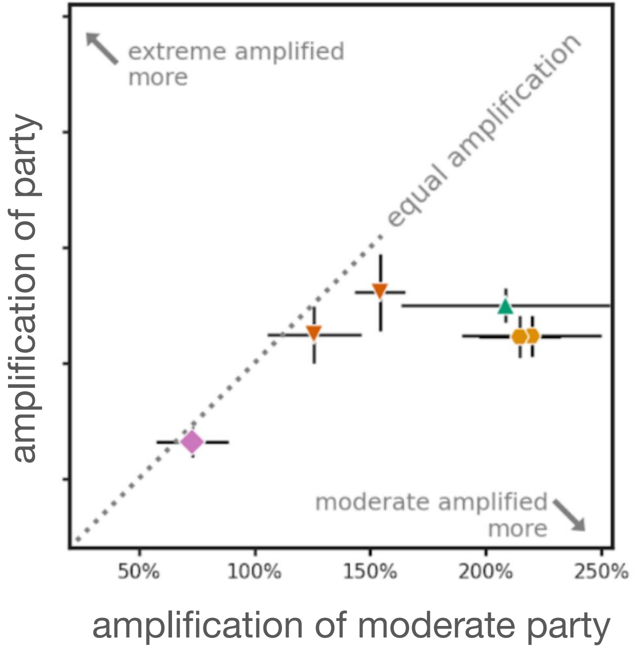



3-in-1 Scheme will be Scrapped for Causing Social Problems

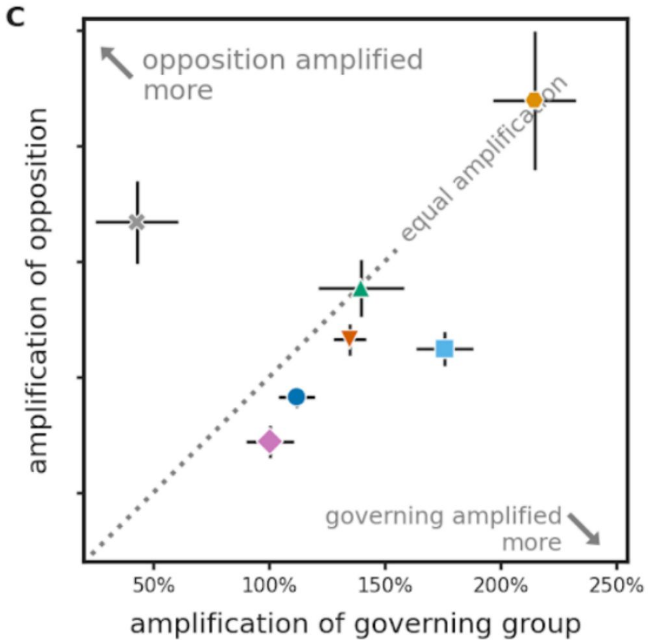
Translator Editor

12 May 2016 18:12 WIB

Extreme vs **moderate** parties



moderate parties are favoured
over far-left and far-right ones



Often (but not always)
governing parties are favoured

In summary

Huszár et al., PNAS 2021

- **at the party level**
  mainstream right-wing parties benefit more from the personalised Home Timeline than left-wing counterparts
- **at the individual level**
  no association between amplification and part membership
- **extreme vs moderate**
  the personalised Home Timeline does not favour extreme ideologies more than mainstream (moderate) ones

Discussion

Huszár et al., PNAS 2021

Why right-wing tweets are amplified more?

Different political parties pursue different strategies on Twitter:

- J. H Parmelee and S. L. Bichard, **Politics and the Twitter Revolution: How Tweets Influence the Relationship between Political Leaders and the Public** (Lexington, 2011)
- D. Freelon, A. Marwick, D. Kreiss, **False equivalencies: Online activism from left to right**. Science 369 (2020)

Discussion

Huszár et al., PNAS 2021

What additional factors, beyond polarization, could be explored in this analysis?

- Misinformation
- Manipulation
- Hate speech
- Abusive content