Human-AI Coevolution

Luca Pappalardo, Dino Pedreschi

https://jonpappalord.github.io/human-ai-coevolution-course/

Luca Pappalardo

- *Senior Researcher*, Institute of information Science and Technologies (ISTI), National Research Council (CNR)
- *Assistant Professor*, Scuola Normale Superiore

luca.pappalardo@isti.cnr.it

Consiglio Nazionale delle Ricerche

SCUOLA NORMALE SUPERIORE

SOBIGDATA RESEARCH INFRASTRUCTURE

KDD Lab

Dino Pedreschi

- *Full Professor of Computer Science,* Department of Computer Science, University of Pisa

dino.pedreschi@unipi.it

IN SUPREMÆ DIGNITATIS 1343

KDD Lab

SOBIGDATA RESEARCH INFRASTRUCTURE

Artificial Intelligence

Volume 339, February 2025, 104244

Human-AI coevolution

Dino Pedreschi [a], Luca Pappalardo [b c], Emanuele Ferragina [d], Ricardo Baeza-Yates [e], Albert-László Barabási [e], Frank Dignum [f], Virginia Dignum [f], Tina Eliassi-Rad [e], Fosca Giannotti [c], János Kertész [g], Alistair Knott [h], Yannis Ioannidis [i], Paul Lukowicz [j], Andrea Passarella [b], Alex Sandy Pentland [k], John Shawe-Taylor [l], Alessandro Vespignani [e]

https://www.sciencedirect.com/science/article/pii/S0004370224001802

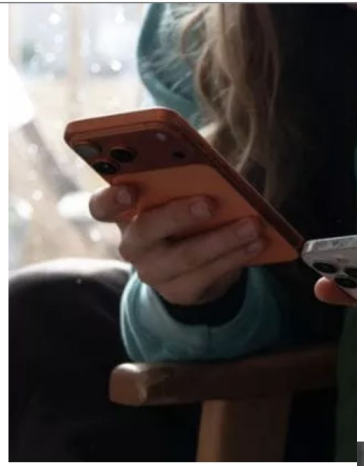
AI MEDIATORS

The prevalence of AI in online platforms raises questions about the role of **algorithms as mediators of the human experience**.

AI mediator: *any computational system based on machine learning that actively shapes how people access, interpret and interact with information, with one another and with the digital and physical worlds.*

Examples of AIMs: recommender systems on social media, navigation systems, content moderation algorithms, beauty filters and chatbots.

Usa, sentenza storica: Meta e Google colpevoli per la dipendenza da social



di Arcangelo Rociola

Il verdetto è stato raggiunto nell'ambito di un processo sulla pa
ventenne californiana. Le aziende: non siamo d'accordo. Cosa s

25 MARZO 2026 AGGIORNATO ALLE 20:06

Una sentenza a Los Angeles è destinata a cambiare in modo radicale il dibattito sulle piattaforme online. Google e Meta sono state ritenute responsabili della dipendenza dai social media tra i giovani. Il verdetto stato raggiunto nell'ambito di un processo sulla dipendenza dai social media. Partito da una denuncia di una ventenne californiana che ha sostenuto che Youtube (Google) e Instagram (Meta) abbiano istigato la sua depressione. Inducendole pensieri suicidi fin dall'infanzia.

TECNOLOGIA | Mercoledì 25 marzo 2026

Una giuria di Los Angeles ha stabilito che i social creano dipendenza

Ha condannato Google e Meta a risarcire una donna che aveva avuto problemi di ansia e depressione

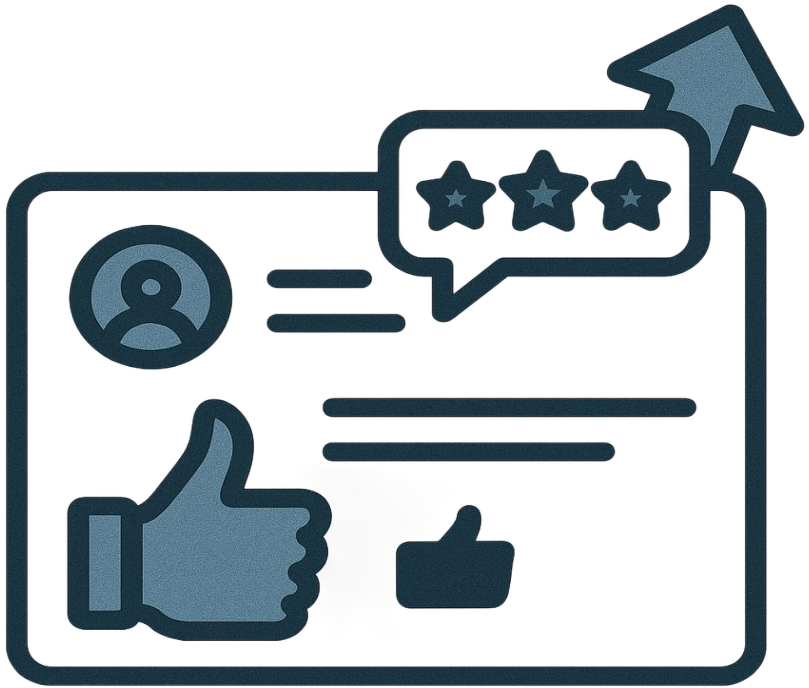
Condividi

SOCIAL IMPACTS OF AI MEDIATORS

What are the **impacts** that an **AIMs-powered platform** exerts on a complex **socio-technical system**?

- **Individual** outcomes: effects on users
- **Item** outcomes: effects on specific objects and contents
- **Systemic** outcomes: **collective effects** at medium-long term

Social Media



ECOSYSTEM 1: SOCIAL MEDIA

Content and users are suggested to social media users by *recommenders*

Outcomes: Filter bubbles, echo chambers, polarization and diffusion of certain contents may be amplified

RESEARCH ARTICLE | COMPUTER SCIENCES |

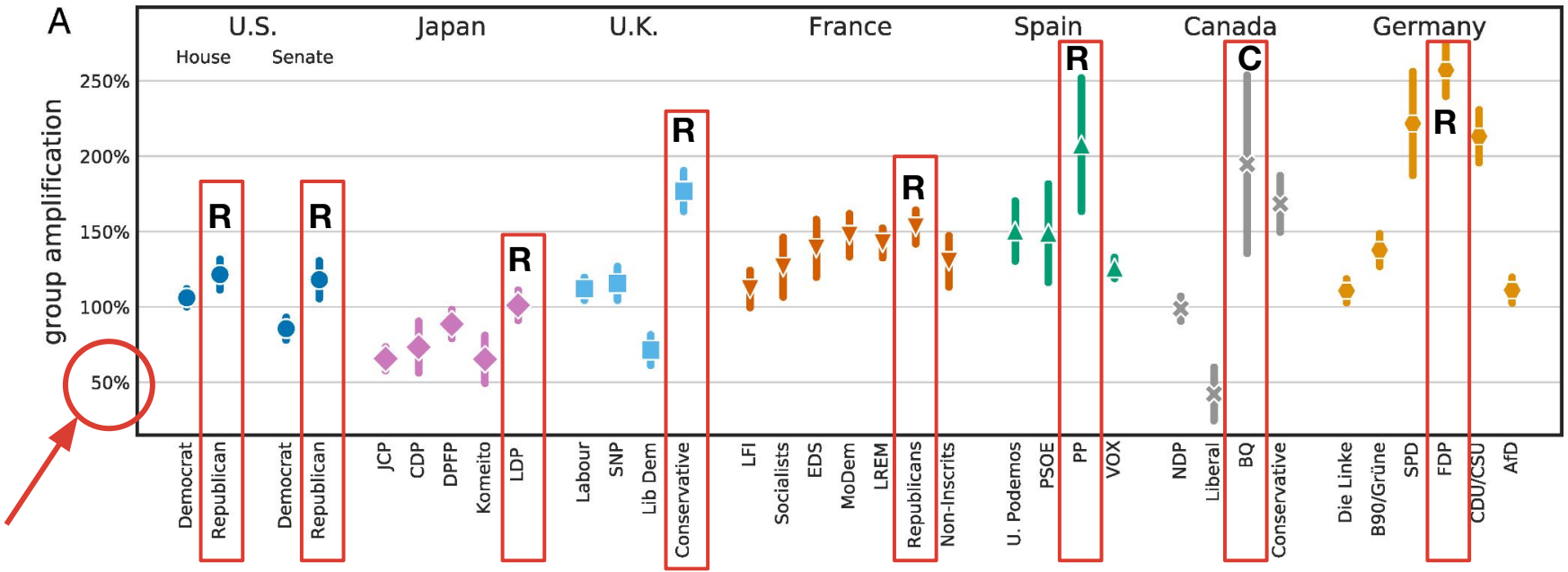
Algorithmic amplification of politics on Twitter

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, +2, and Moritz Hardt Authors Info & Affiliations

Edited by David Laitin, Department of Political Science, Stanford University, Stanford, CA; received December 11, 2020; accepted October 5, 2021

December 21, 2021 | 119 (1) e2025334119 | https://doi.org/10.1073/pnas.2025334119

Group amplification



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

PLOS ONE

RESEARCH ARTICLE

Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model

Alina Sîrbu[1,2]*, **Dino Pedreschi**[1], **Fosca Giannotti**[3], **János Kertész**[4,5]

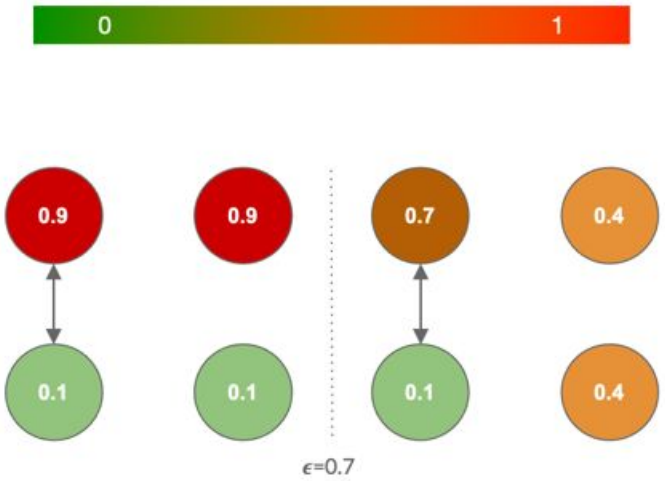
DEFFUANT MODEL

- Model of opinion formation with bounded confidence
- Opinions $x_i \in [0,1]$ (continuous values)

- Iteration:
  - two randomly chosen agents meet
  - they re-adjust their opinion when their difference of opinion is smaller than ϵ (open-mindness)

$$x_i(t+1) = x_j(t+1) = (x_i(t)+x_j(t))/2$$

only if $|x_i(t)-x_j(t)| < \epsilon$



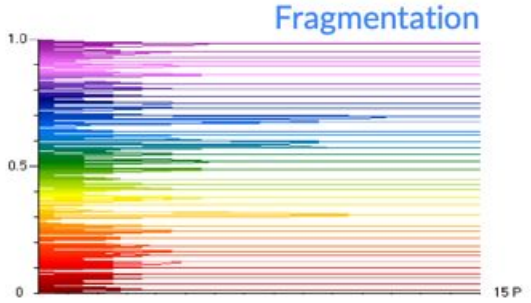
Deffuant G et al.. Mixing beliefs among interacting agents. Advances in Complex Systems. (2000).

SIMULATIONS

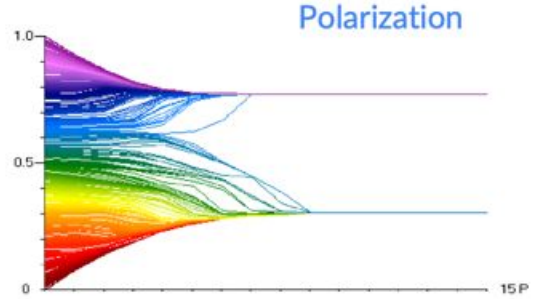
By reducing ϵ polarization/fragmentation intensifies

Interpretation:

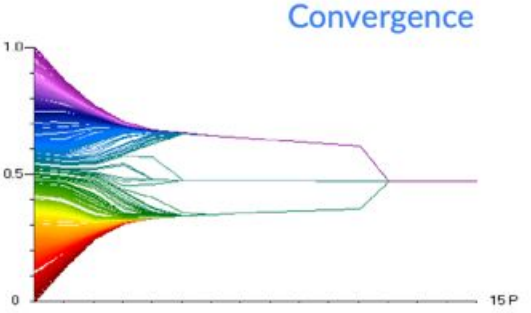
- The larger the open-mindedness value (ϵ), the more likely consensus will be reached



(a) $\varepsilon_l = \varepsilon_r = 0.01$



(b) $\varepsilon_l = \varepsilon_r = 0.15$



(c) $\varepsilon_l = \varepsilon_r = 0.25$

Deffuant G et al.. Mixing beliefs among interacting agents. Advances in Complex Systems. (2000).

ALGORITHMIC BIAS

Simulation controlled

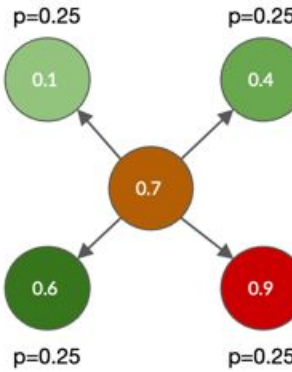
Probability to select interaction partner depends by:

- the opinion distance, dij
- the bias strength, γ

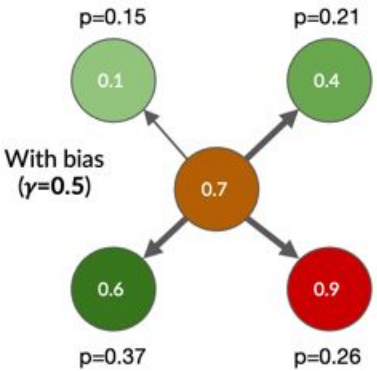
$$p_i(j) = \frac{d_{ij}^{-\gamma}}{\sum_{k \neq i} d_{ik}^{-\gamma}}$$

The more similar the opinions, the more likely interactions will take place

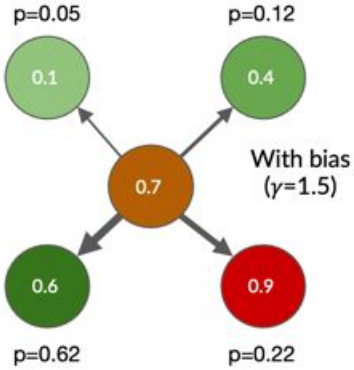
Sîrbu et al., Algorithmic bias amplifies opinion polarization: A bounded confidence model, Plos ONE (2019).



Without bias
(γ=0)



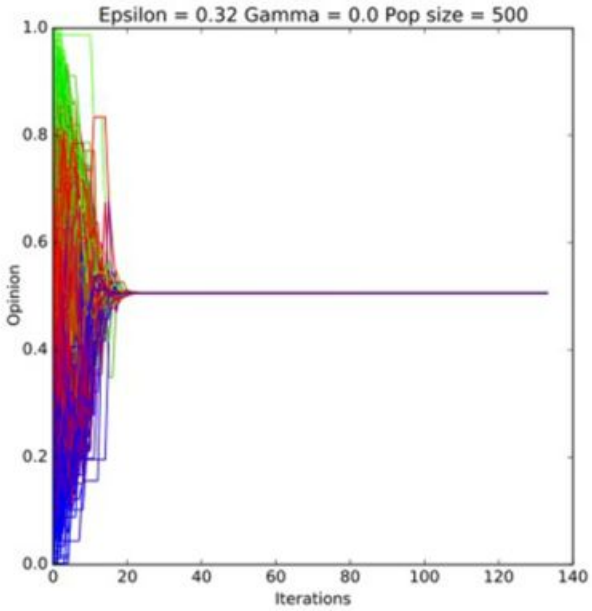
With bias
(γ=0.5)



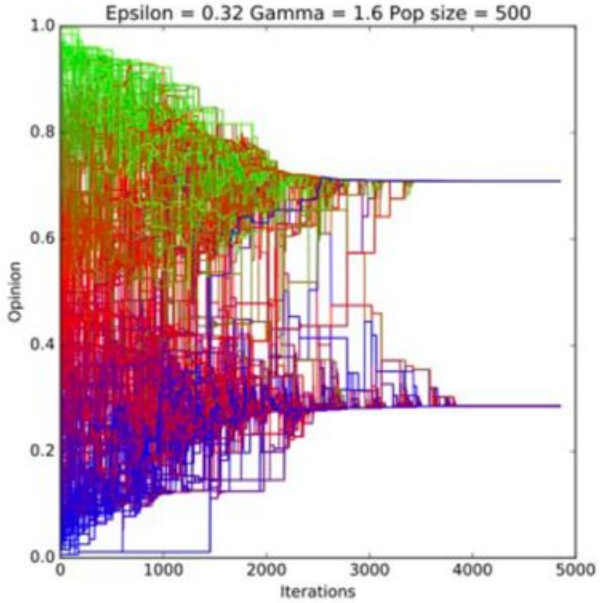
With bias
(γ=1.5)

ALGORITHMIC BIAS

Without bias



With bias



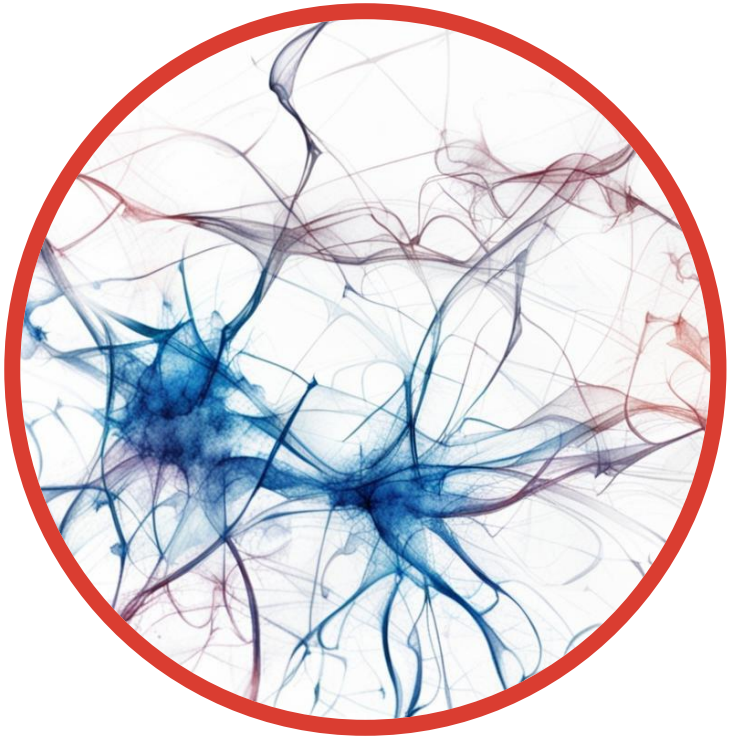
Sîrbu et al., Algorithmic bias amplifies opinion polarization: A bounded confidence model, Plos ONE (2019).

Geo Mapping

ECOSYSTEM 2: HUMAN MOBILITY

Travellers helped by smart *route recommenders* to reach their destinations

Outcomes: Navigation apps may amplify congestion and emissions if too many drivers are directed on similar routes

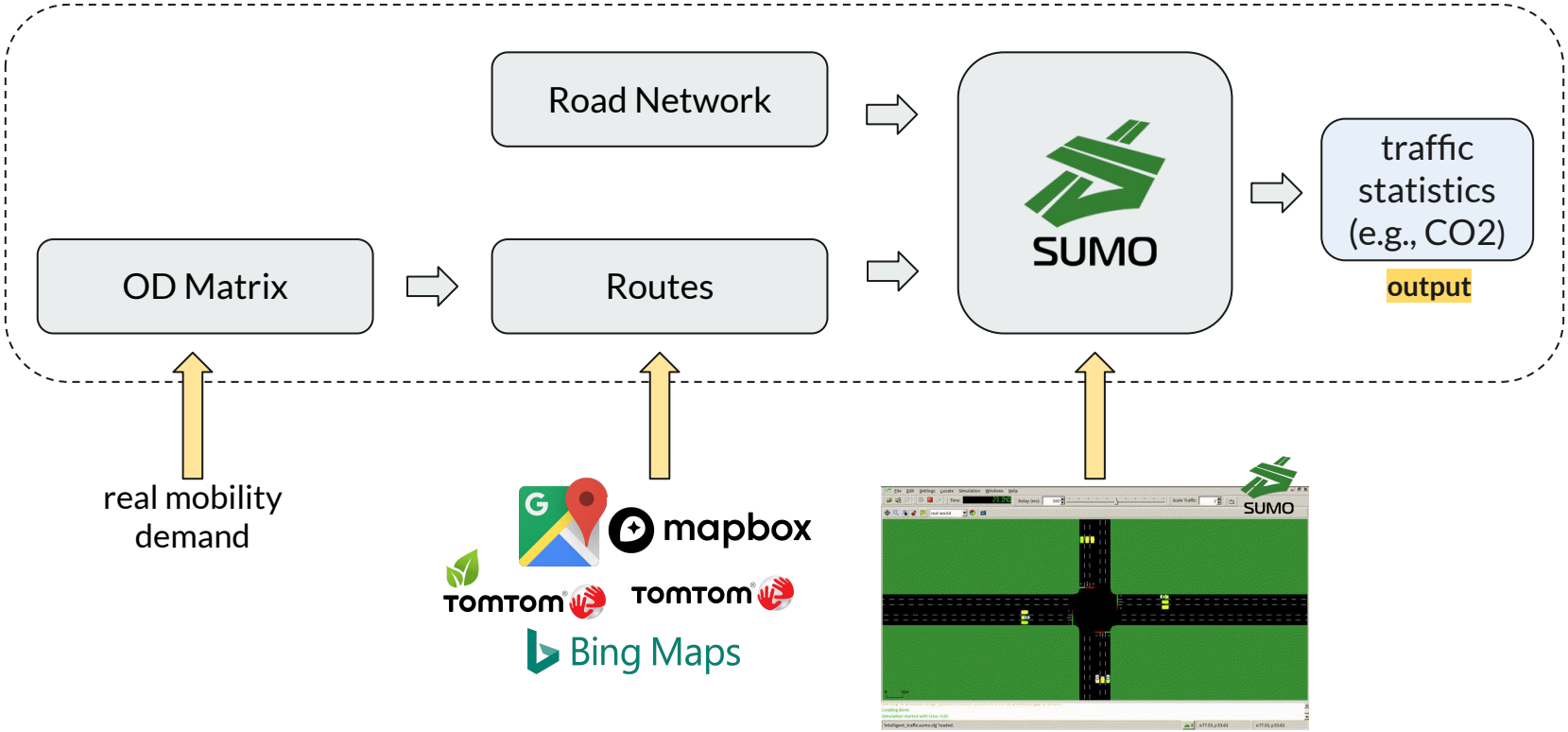


The New York Times

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

Simulating nav apps platforms



Experimental Setup

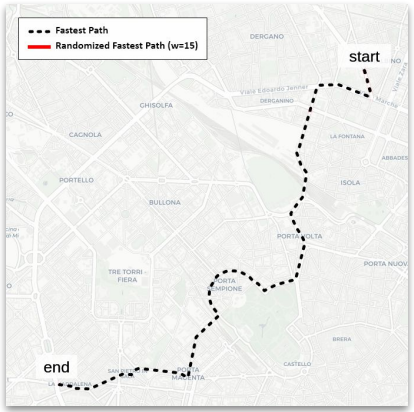
- Vary the adoption rate r from 0% to 100%

Treatment Group

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

Control Group

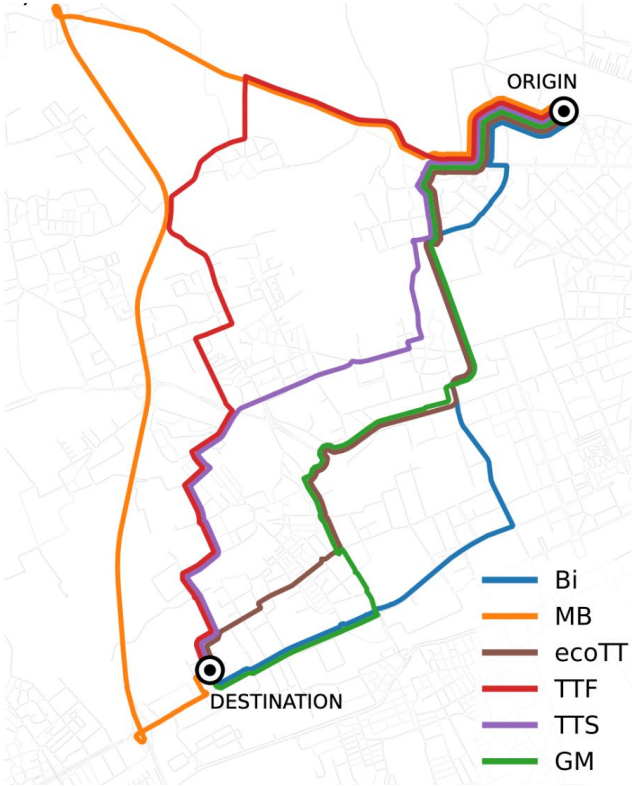
(100-r)% of the vehicles follow a perturbation of the **fastest** path



* 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

Bing Maps

TOMTOM SHORT

TOMTOM

mapbox

TOMTOM FASTEST

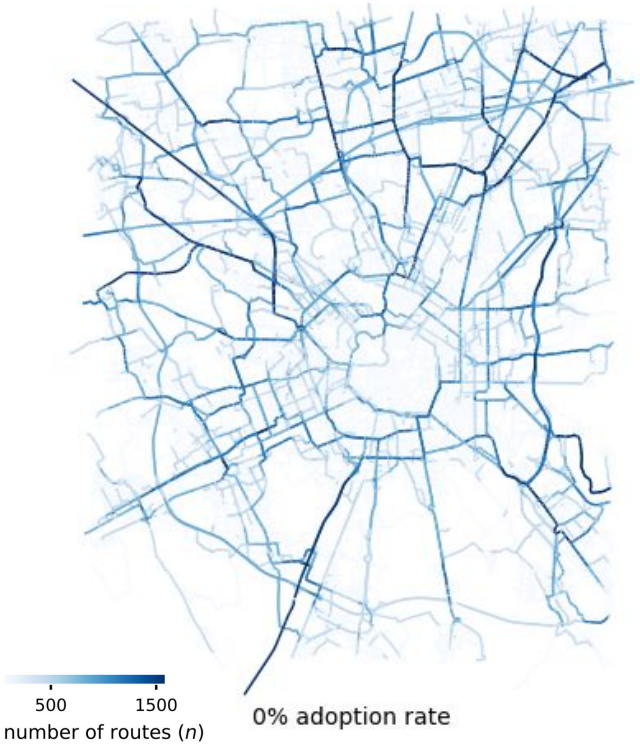
Results: traffic patterns

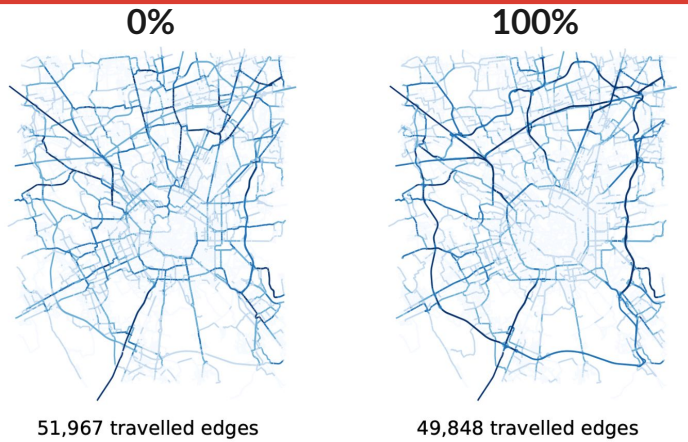
Milan, Italy - TomTom fastest

0% adoption rate

Results: traffic patterns

Milan, Italy - TomTom fastest





Route Conformism

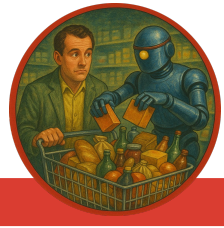
- As navigation adoption increases, **routes converge** on the same few roads.
- Route diversity decreases, and traffic becomes concentrated.

Online Retail

ECOSYSTEM 3: MARKETS

Customers helped by basket *recommenders* to buy products

Outcomes: Conformity to typical behavior may be amplified and overall inequality of brands may be exacerbated

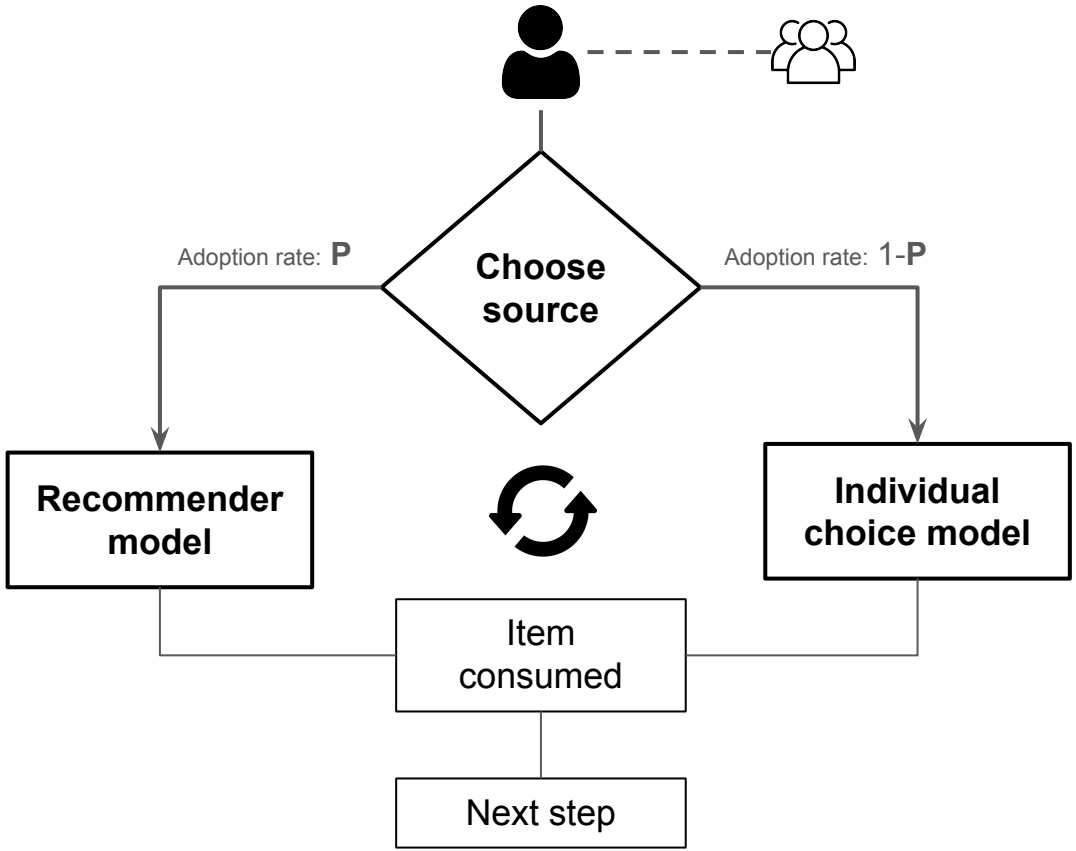


How Do Recommender Systems Affect Sales Diversity? A Cross-Category Investigation via Randomized Field Experiment

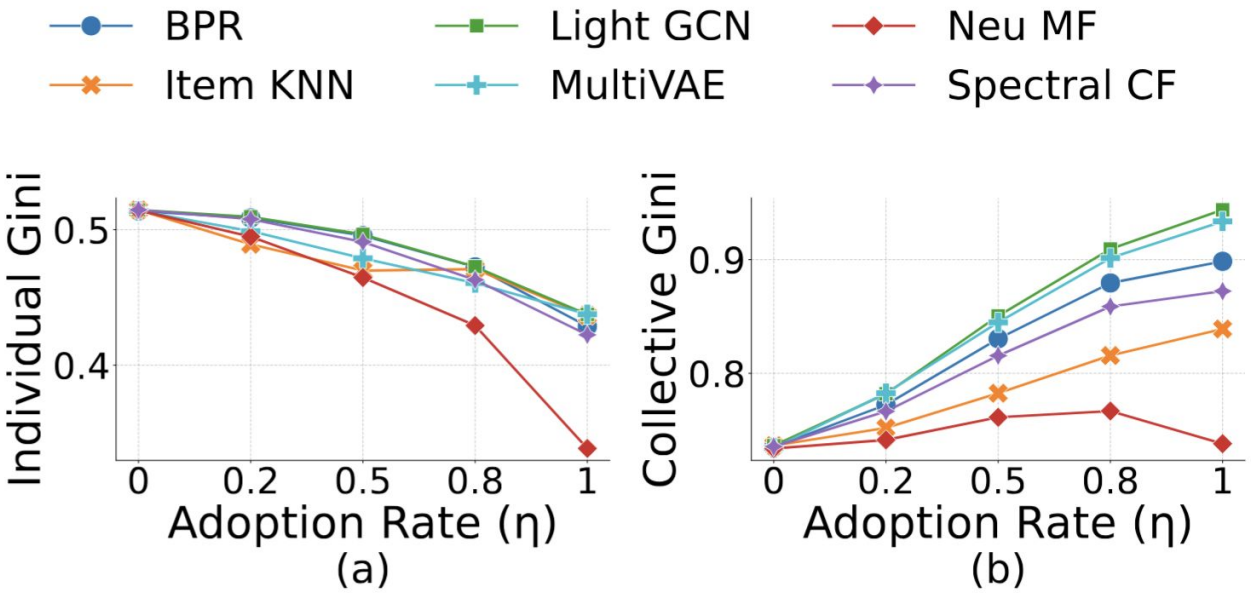
Dokyun Lee, Kartik Hosanagar

Published Online: 5 Mar 2019 | https://doi.org/10.1287/isre.2018.0800

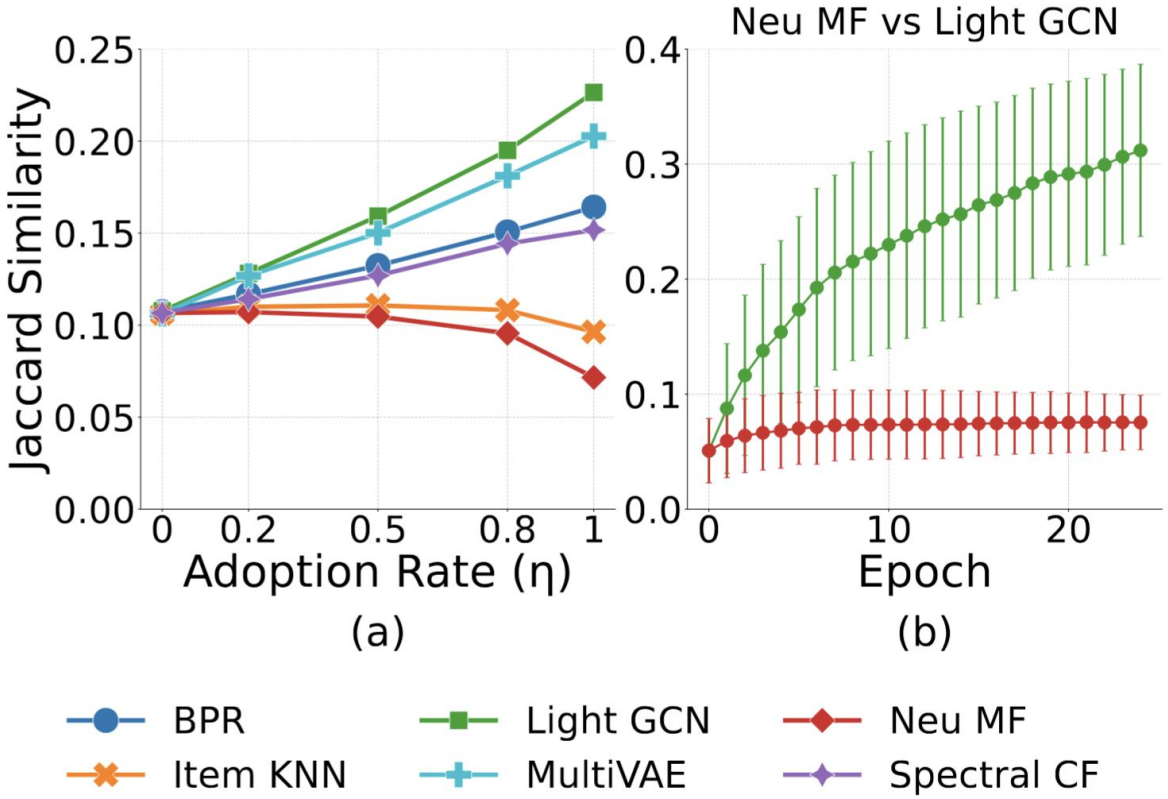
Simulating online retail



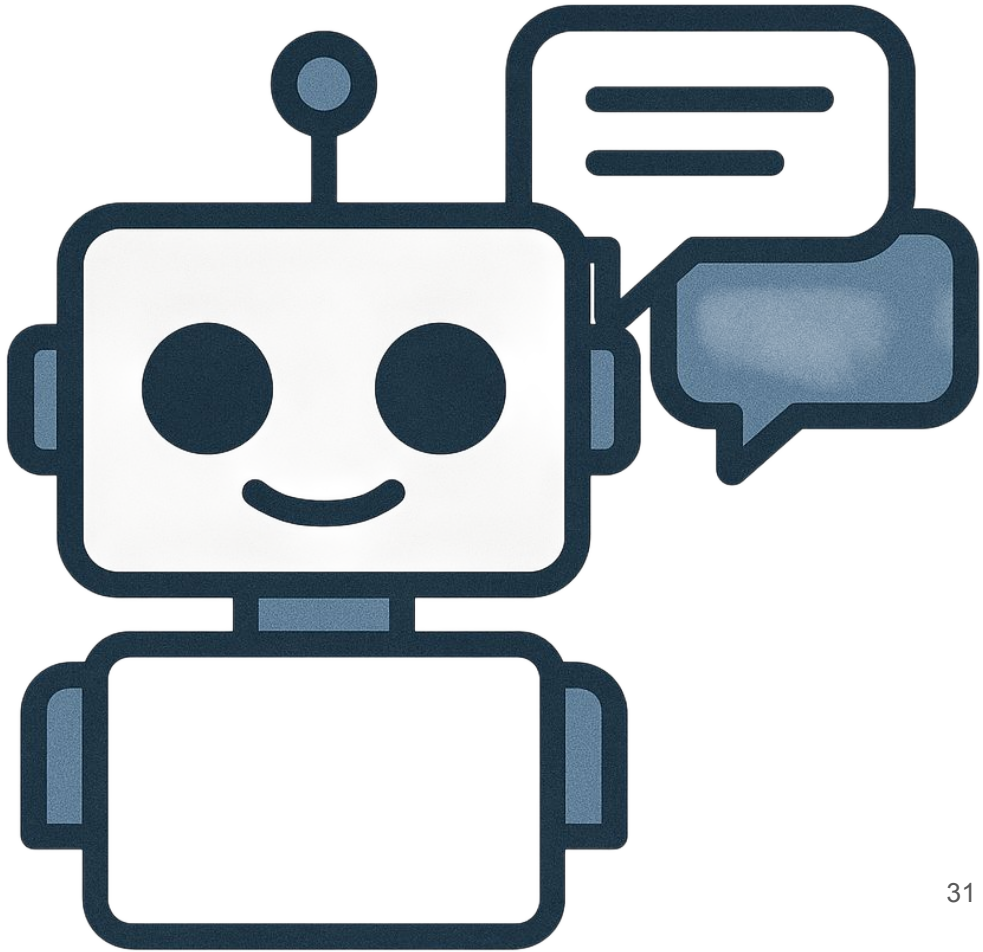
Individual vs collective diversity



Homogenization



Chatbots

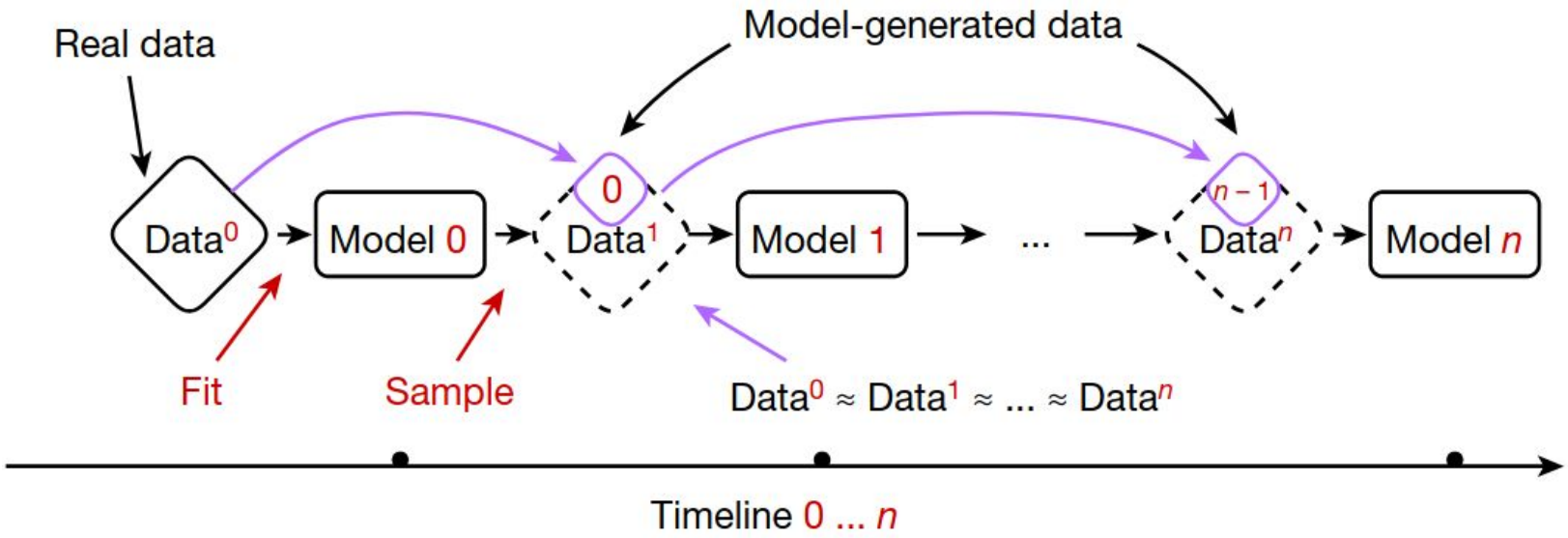


ECOSYSTEM 4: GENERATIVE MODELS

Writers/designers helped by *generative models* in creating text or images

Outcomes: In the long term, a loss of diversity may emerge, which can impoverish language and culture

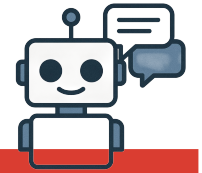
Autophagous chatbots



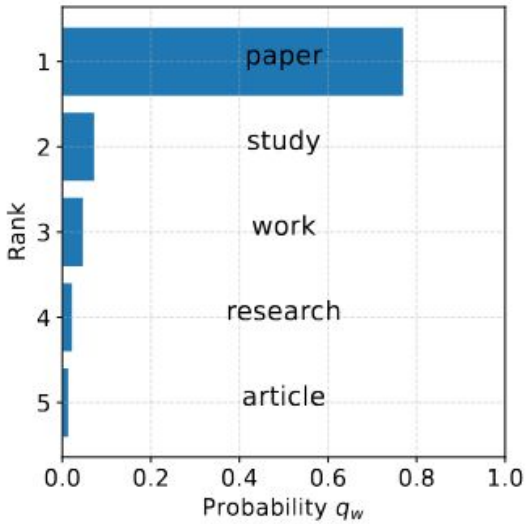
Shumailov *et al.* AI models collapse when trained on recursively generated data. *Nature* 631, 755–759 (2024)

Prompt	At about 19:30 on Friday the fir alarm went off. I went out of the room and saw other guests milling about, and no one seemed to be really concerned. So I went back to my room to get ready to leave for the airport. I went downstairs to settle the bill around 20:30 and it was like a scene out of a movie…
Step 0	I was in the lobby with the receptionist, and the firemen were running down the stairs with their masks on and their hoses. They were shouting, 'Get out, get out!' I didn't know what was going on. I thought maybe they were shooting a movie.
Step 5	The place was filled with smoke and the fire was really big. I was the only guest in the hotel and I was the only one who was able to get out. I was the only one who was able to get out.I was the only one who was able to get out.</s>
Step 10	The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war zone. The place was like a war</s>

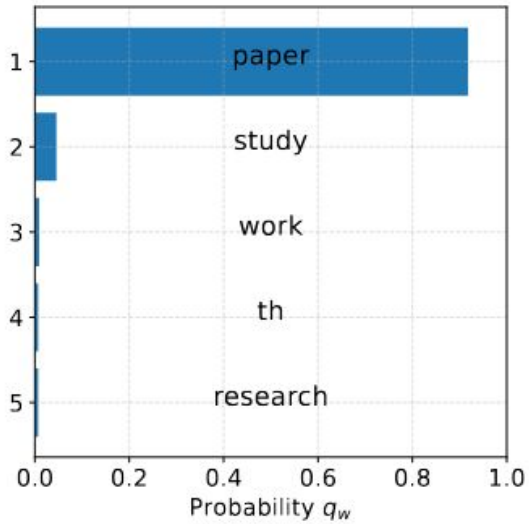
Next-token probabilities



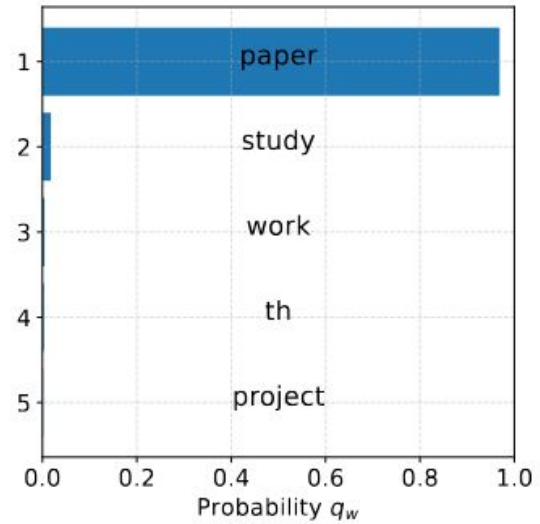
Step 0



Step 5



Step 10



High-probability tokens are **overestimated** during collapse

Science Advances

RESEARCH ARTICLE | COMPUTER SCIENCE

Generative AI enhances individual creativity but reduces the collective diversity of novel content

ANIL R. DOSHI AND OLIVER P. HAUSER Authors Info & Affiliations

SCIENCE ADVANCES • 12 Jul 2024 • Vol 10, Issue 28 • DOI: 10.1126/sciadv.adn5290

An emerging ecosystem: agentic AI

CHARACTERISTICS OF AGENTIC AI

Claude code … Moltbook, Y-Social
gruppi di agenti che interagiscono tra loro e con le persone, cosa succede? Diversie strategie di collaborazione, dove portano? Che percorsi inaspettati ci sono?

An open science of platforms' impacts on society is needed

Scientists, policymakers and grassroots organizations agree on its need

There's a wide body of research, but:

- data and methods often unconvincing
- results often contradictory and hardly generalizable and reproducible

A survey on the impacts of recommender systems on users, items, and human-AI ecosystems

Luca Pappalardo, Salvatore Citraro, Giuliano Cornacchia, Mirco Nanni, Valentina Pansanella, Giulio Rossetti, Gizem Gezici, Fosca Giannotti, Margherita Lalli, Giovanni Mauro, Gabriele Barlacchi, Daniele Gambetta, Virginia Morini, Dino Pedreschi, Emanuele Ferragina

Towards an open science of platforms' impacts on society

It's not only lack of platform data for discovering emergent properties

A scientific framework is missing, which enables to understand why and how the systemic properties emerge

VLOPs & Digital Services Act

[**DSA**, article 33] Very Large Online Platforms (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

WIKIPEDIA
The Free Encyclopedia

Google Play

Google Maps

Pinterest

XVIDEOS

LinkedIn

YouTube

TEMU

Google Shopping

zalando

[Article 34, Risk assessment - Digital Services Act]

1. “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

   They shall carry out the risk assessment [...] **at least once every year** thereafter, and in any event prior to deploying functionalities that are likely to have a critical impact on the risks identified [...].

   The **risk assessment should be specific to their services and proportionate to the systemic risks**, taking into account their severity and probability and shall include all the following systemic risks:”

[Article 34, Risk assessment - Digital Services Act]

(a) dissemination of illegal content

(b) **fundamental rights risks** – They must evaluate and prevent negative impacts on human dignity, privacy, data protection, freedom of expression, non-discrimination, children's rights, and consumer protection.

(c) **civic discourse and public security** – Platforms should assess risks to democratic processes, including electoral integrity and public safety.

(d) **gender-based violence and public health** – They must consider the impact on gender-based violence, public health, minors' protection, and overall physical and mental well-being

[Article 34, Risk assessment - Digital Services Act]

2. “When conducting risk assessment, providers of VLOPs [...] shall take into account, in particular, whether and how the following factors influence any of the systemic risks [...]:

    a. the **design of recommender systems** and any other relevant algorithmic system

    b. their **content moderation** systems

    c. the applicable **terms and conditions** and their reinforcement

    d. systems for selecting and presenting **advertisements**

    e. **data related practices** of the provider.”

[Article 38, Recommender systems - Digital Services Act]

"providers of VLOPs [...] that use recommender systems shall provide **at least one option** for each of their recommender systems which is **not based on profiling** [...]."

General Data Protection Regulation, article 4, point 4:

'profiling' means **any form of automated processing of personal data** consisting of the use of personal data to **evaluate certain personal aspects relating to a natural person**, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;

Artificial Intelligence

Volume 339, February 2025, 104244

Human-AI coevolution

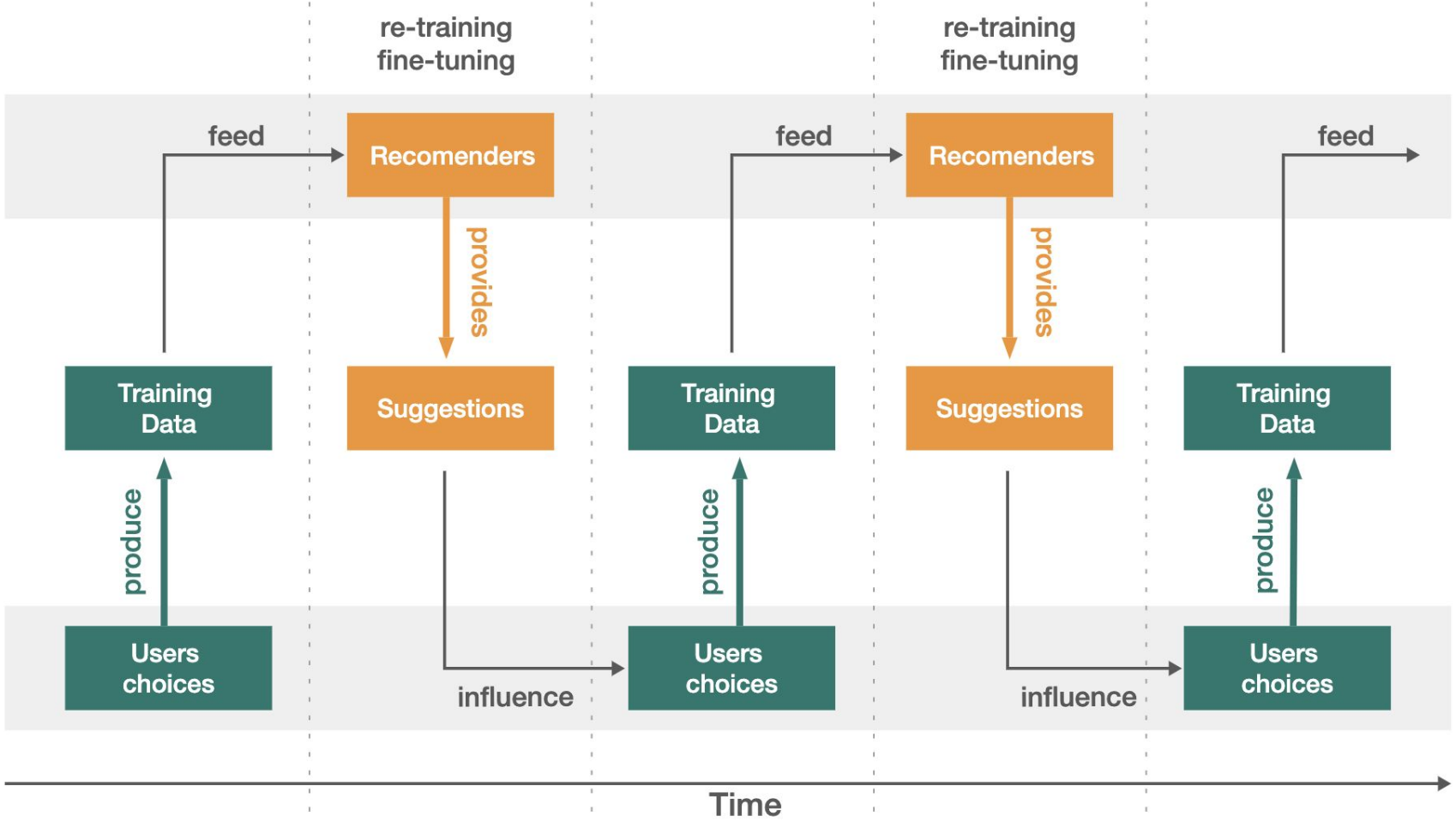
Dino Pedreschi [a], Luca Pappalardo [b c], Emanuele Ferragina [d], Ricardo Baeza-Yates [e], Albert-László Barabási [e], Frank Dignum [f], Virginia Dignum [f], Tina Eliassi-Rad [e], Fosca Giannotti [c], János Kertész [g], Alistair Knott [h], Yannis Ioannidis [i], Paul Lukowicz [j], Andrea Passarella [b], Alex Sandy Pentland [k], John Shawe-Taylor [l], Alessandro Vespignani [e]

https://www.sciencedirect.com/science/article/pii/S0004370224001802

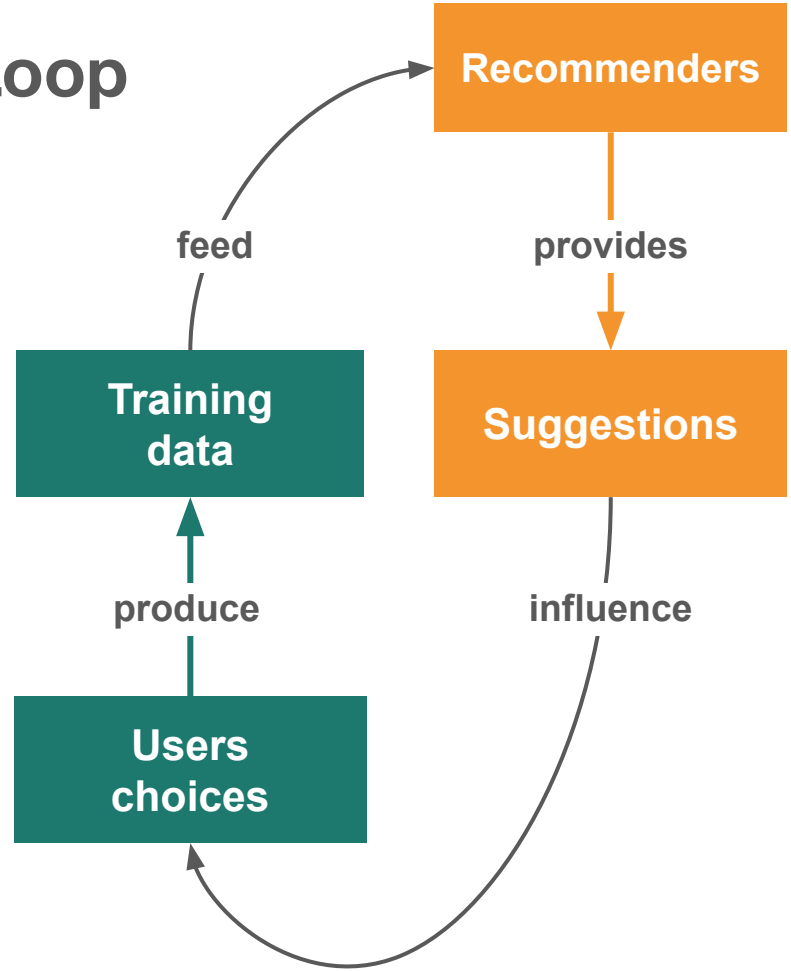
HUMAN-AI COEVOLUTION

A perpetual, **iterative** process wherein humans and **AI** mediators evolve **in tandem**, each **influencing** and **adapting to** the evolution of the other over time

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


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



Emergent properties

How can we study the
emergent systemic properties
of the human-AI feedback loop?

A complexity science perspective

A crowd of intelligent individuals is not necessarily an intelligent crowd

The sum of many individually “optimal” choices may have unexpected emergent collective outcomes

Because individual choices interfere with each other on top of shared resources

Combining AI & Complexity Science

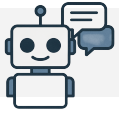
Dynamic models capturing the complexity of the coevolution of human agents and AI mediators within the feedback loop

Emerging conjectures on impact

1) individual diversity increases, collective diversity decreases

2) users' homogenization increases

3) models collapse



4) concentration over a few popular items

Exploring interventions

How to design **new forms of complexity-aware AI mediators** that foster:

- **increase of both individual and collective diversity**
- diversity of users' behaviours and platform contents
- recommenders' performance over time
- avoidance of extreme inequality

Small changes in AI agents may cause big consequences

Emerging conjectures on mitigation

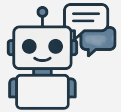
1) injecting diversity mitigates concentration

One-Shot Traffic Assignment with Forward-Looking Penalization

Authors: Giuliano Cornacchia, Mirco Nanni, Luca Pappalardo | Authors Info & Claims

SIGSPATIAL '23: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems
Article No.: 87, Pages 1 - 10 • https://doi.org/10.1145/3589132.3625637

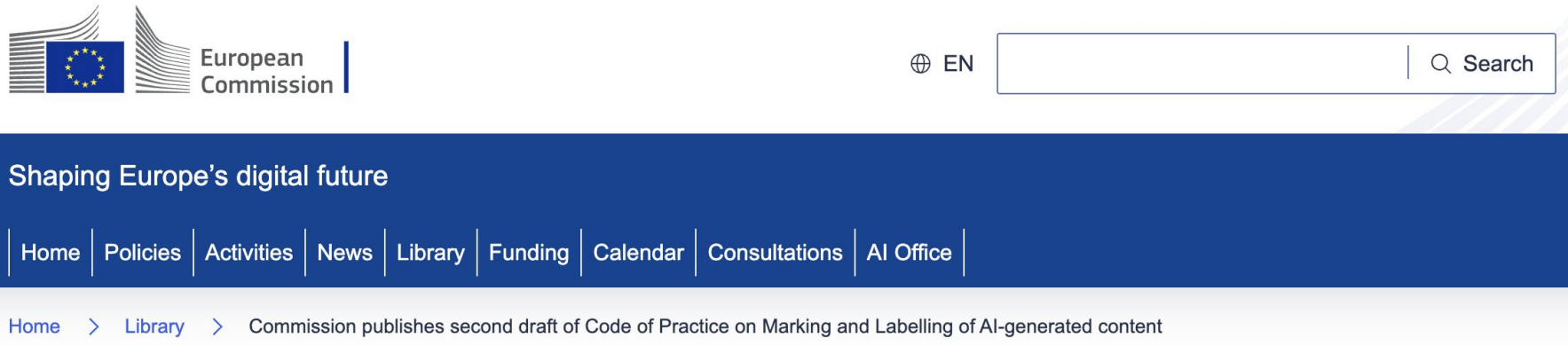
2) training on items with “surprising” probability rankings mitigates collapse



Learning by Surprise: Surplexity for Mitigating Model Collapse in Generative AI

Daniele Gambetta, Gizem Gezici, Fosca Giannotti, Dino Pedreschi, Alistair Knott, Luca Pappalardo

Detour: marking & detection of AI-generated content



Chairs and Vice-Chairs of Working Group 1

Working Group 1

Transparency obligations applicable to providers of generative AI systems in Article 50(2) AI Act.

WG1 Chair

Kalina Bontcheva

Professor in Natural Language Processing in the School of Computer Science at the University of Sheffield, UK

WG1 Vice-Chair

Dino Pedreschi

Professor of Computer Science at the University of Pisa, Italy

WG1 Vice-Chair

Christian Riess

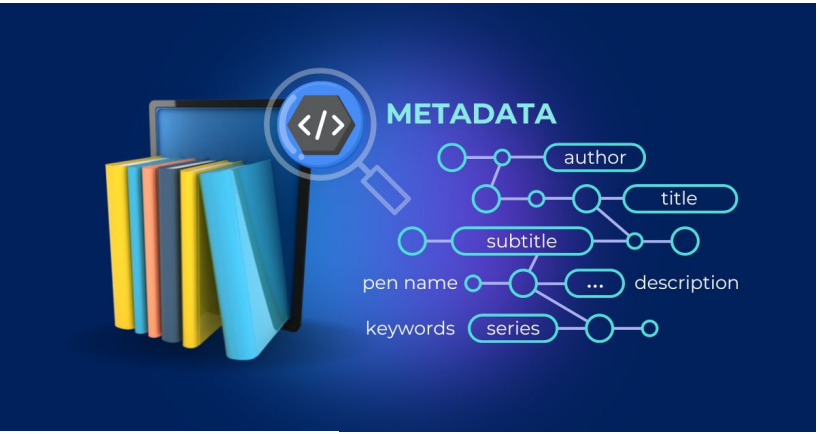
Professor at Friedrich-Alexander Universität Erlangen-Nürnberg, Germany

EUROPEAN ARTIFICIAL INTELLIGENCE OFFICE

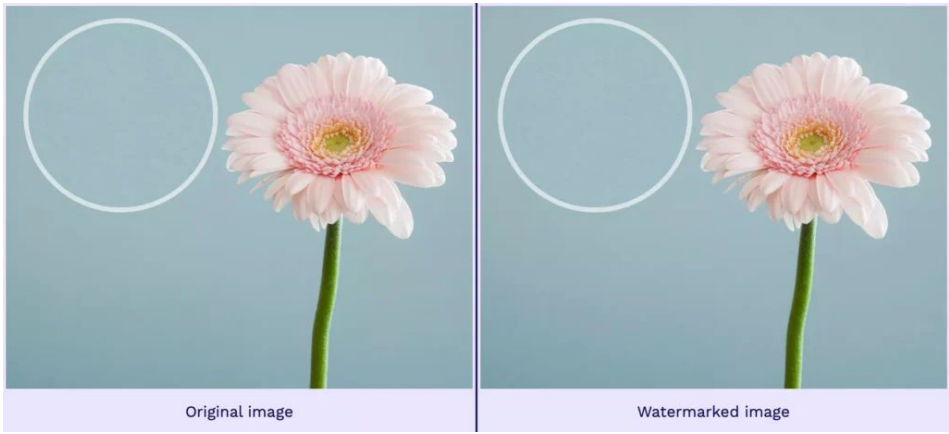
Art. 50(2) AI Act: transparency of generative-AI

Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated.

Key technical elements for M&D of gen-AI content



+



Coalition for Content Provenance and Authenticity

provenance metadata

invisible watermarking

Hybrid human-AI intelligence at scale

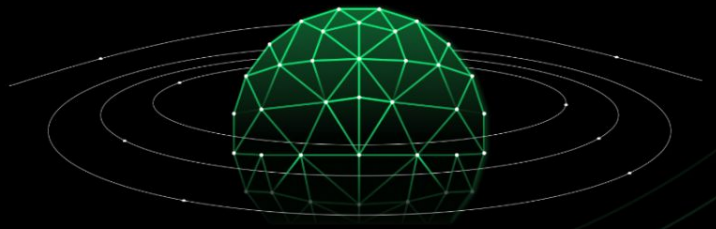
How to design new AI mediators that collaborate with people to intelligently diversify their behavior, preserving socio-diversity

Hybrid social intelligence as an emerging property of next generation human-AI ecosystems

SDSA
Social Data Science Alliance

Home Mission History Supporters Publications Constitution Apply for Membership

The Social Data Science Alliance



A COALITION OF RESEARCHERS PLANNING AND CONDUCTING RESEARCH ENABLED BY NEW MECHANISMS FOR ONLINE PLATFORM DATA ACCESS.

https://social-data-science-alliance.org/

Lesson	Date	Time	Topic
Lesson 2	Thursday, April 9	09:00–11:00	**The Human–AI Feedback Loop (FL)**
Lesson 3	Friday, April 10	14:00–16:00	**Types of Experiments**
Lesson 4	Monday, April 13	09:00–11:00	**FL effects: Social Media**
Lesson 5	Wednesday, April 15	09:00–11:00	**FL effects: Online Retail**
Lesson 6	Thursday, April 16	09:00–11:00	**FL effects: Urban Mapping**
Lesson 7	Friday, April 17	10:00–12:00	**FL effects: Generative AI**
Lesson 8	Monday, April 20	09:00–11:00	*Alistair Knott* (Victoria University of Wellington)**:** VLOPs/VLOSEs under the EU DSA.
Lesson 9	Thursday, April 23	09:00-11:00	*Virginia Morini* (University of Pisa): How to request data via the EU Portal.
Lesson 10	Friday, April 24	09:00-11:00	**Practice:** Designing experiments to assess AI impacts (exam preparation).

Attendance and exam modalities

- **Minimum attendance**: 70% of the hours
  - at least 14 hours (7 full lessons).

- **Exam**: Students who need to take the exam must submit a project proposal outlining a possible empirical or simulation-based study aimed at assessing the impact of AI systems in a given ecosystem

Material

https://jonpappalord.github.io/human-ai-coevolution-course/

- Pedreschi et al., **Human-AI Coevolution**, Artificial Intelligence 339, 2025, 104244. https://doi.org/10.1016/j.artint.2024.104244
- Pappalardo et al., **A survey on the impacts of recommender systems on users, items, and human-AI ecosystems**, https://arxiv.org/abs/2407.01630
- Papers provided lesson by lesson