Cognitive Biases in Human and Algorithmic Decision-Making

HCAI@OvGU Workshop: 05-11-2024

JOHANNES KEPLER UNIVERSITY LINZ

INSTITUTE OF COMPUTATIONAL PERCEPTION

MULTIMEDIA MINING AND SEARCH GROUP

Markus Schedl
Johannes Kepler University Linz, Austria
Linz Institute of Technology, Austria
markus.schedl@jku.at | www.mschedl.eu | www.hcai.at

JOHANNES KEPLER
UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

INSTITUTE OF COMPUTATIONAL PERCEPTION

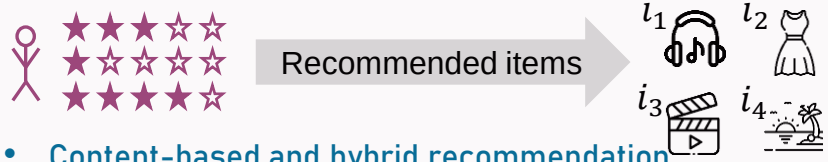
MULTIMEDIA MINING AND SEARCH GROUP

HUMAN-CENTERED ARTIFICIAL INTELLIGENCE GROUP
LIT AI LAB

MARKUS SCHEDL
markus.schedl@jku.at
https://hcai.at

Selected Research Areas

Recommender Systems



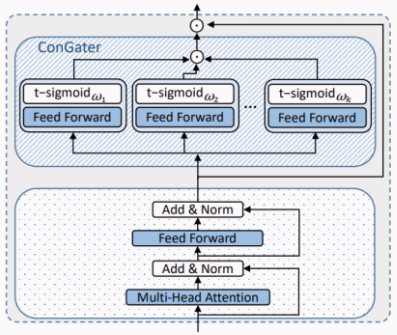
- Content-based and hybrid recommendation
- Psychology-informed recommender systems
- Domain-specific recommenders (music, jobs, etc.)
- Fairness and privacy in recommender systems
- Multiobjective and multistakeholder recommendation

Information Retrieval

- Music information retrieval
- Cross-modal retrieval
- Parameter-efficient retrieval models
- Unbiased retrieval

Natural Language Processing

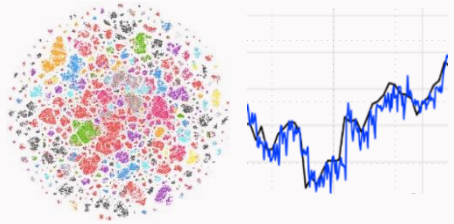
- Emotion recognition
- Representation disentanglement in LLMs
- Debiasing LLMs



Human-Computer Interaction

- Intelligent user interfaces for music/media discovery
- Perception of biases in algorithmic decision-making

Data Science



- Time series analysis
- Pattern recognition in user-generated data

Ongoing/Recent Projects

FWF Der Wissenschaftsfonds.

- Intent-aware Music Recommender Systems (Stand-alone Project)
- Human-centered Artificial Intelligence (Doc.funds.connect Project)
- Humans and Recommender Systems: Towards a Mutual Understanding (Stand-alone Project)
- Bilateral AI (Cluster of Excellence)

JKU LINZ INSTITUTE OF TECHNOLOGY

- Fair Representation Learning with Fine-grained Adversarial Regulation of Bias Flow
- Mitigating Gender Bias in Job Recommender Systems: A Machine Learning-Law Synergy

FFG

- Fairness-aware Algorithmic Decision Support Systems
- Theory-inspired Recommender Systems

Selected Collaborations

European Commission JOINT RESEARCH CENTRE
UNIVERSITY OF CAMBRIDGE
Imperial College London
Politecnico di Bari
universität innsbruck
amazon
SIEMENS energy
jobiqo
deezer
TU WIEN
TU Graz

What Are Cognitive Biases?

- *Psychology:* systematic perceptual deviations of the individual from rationality and objectivity pertaining to cognition, judgment, or decision-making, which often happens unconsciously

- *Sociology:* collective prejudices of a society that favor one group's values, norms, and traditions over others

- Overarching *research questions*:
  - In which parts of the (algorithmic) decision-making pipeline can we *observe cognitive biases (CoBis)*, e.g., user-item interactions, side information, training data, ranking algorithm and model, presentation of results?
  - Can we *leverage* (positive) and *mitigate* (negative) cognitive biases in algorithmic decision-making and the human in the loop?

Which Decision-Making Systems?

- Information Retrieval (IR) / Search
  - User → Query → Algorithm/Model → Retrieved Documents → Presentation (UI)
  - Potentially, CoBis reflected in all(?) of the above
- Recommender Systems (RecSys, RSs)
  - Interactions → User Profile → Algorithm/Model → Recommended Items → Presentation (UI)
  - Potentially, CoBis reflected in all(?) of the above
- Large Language Models (LLMs)
  - Generative models (e.g., ChatGPT, Gemini, Claude) → CoBis in prompts and responses
  - Word/sentence embeddings (e.g., BERT, RoBERTa) → When used in ranking tasks (IR/RS), CoBis in retrieval/recommendation lists

Which Decision-Making Systems?

- Information Retrieval (IR) / Search
  - User → Query → Algorithm/Model → Retrieved Documents → Presentation (UI)
  - Potentially, CoBis reflected in all(?) of the above
- Recommender Systems (RecSys, RSs)
  - Interactions → User Profile → Algorithm/Model → Recommended Items → Presentation (UI)
  - Potentially, CoBis reflected in all(?) of the above
- Large Language Models (LLMs)
  - Generative models (e.g., ChatGPT, Gemini, Claude) → CoBis in prompts and responses
  - Word/sentence embeddings (e.g., BERT, RoBERTa) → When used in ranking tasks (IR/RS), CoBis in retrieval/recommendation lists

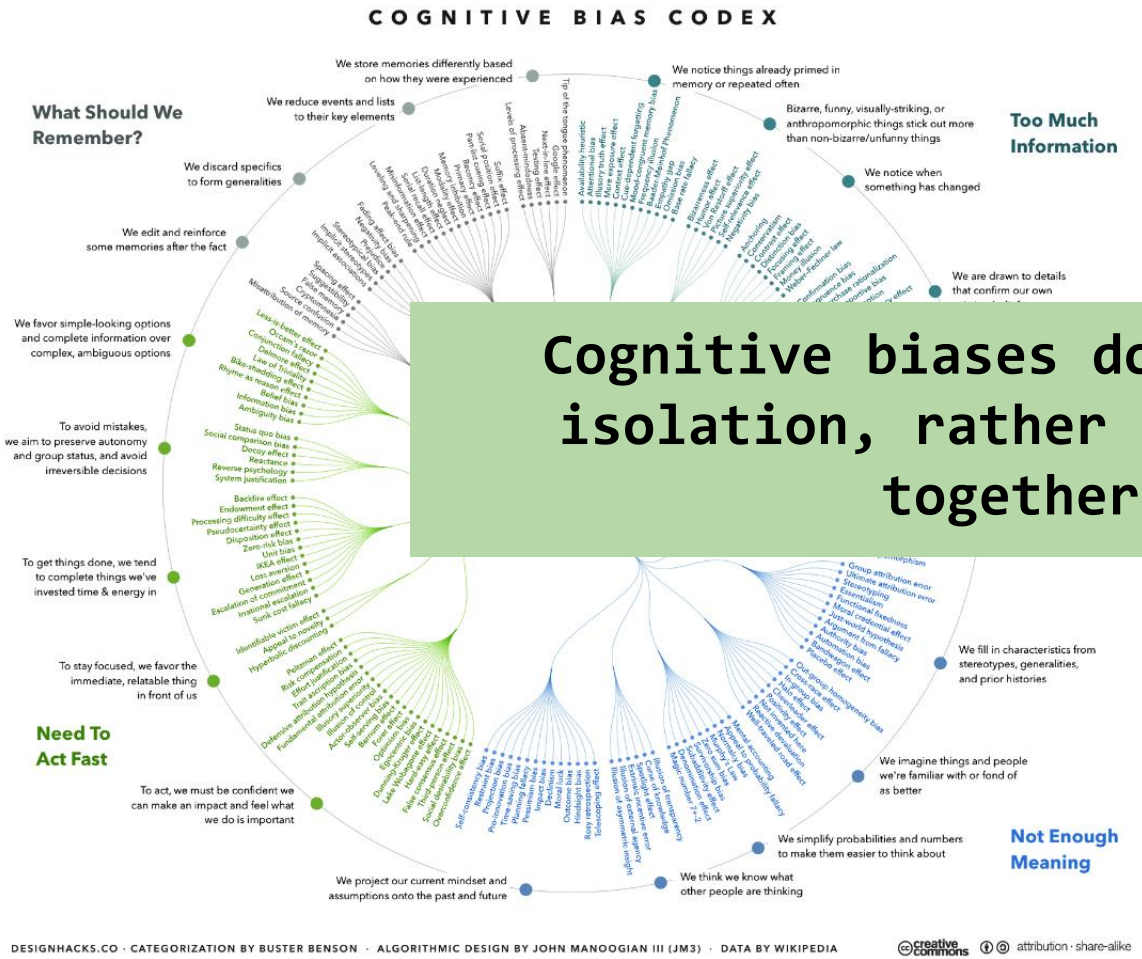
Cognitive Biases: Examples

- Feature-Positive Effect
- IKEA Effect
- (Cultural) Homophily
- Conformity Bias
- Declinism
- Primacy/Recency Effects, Position Bias
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

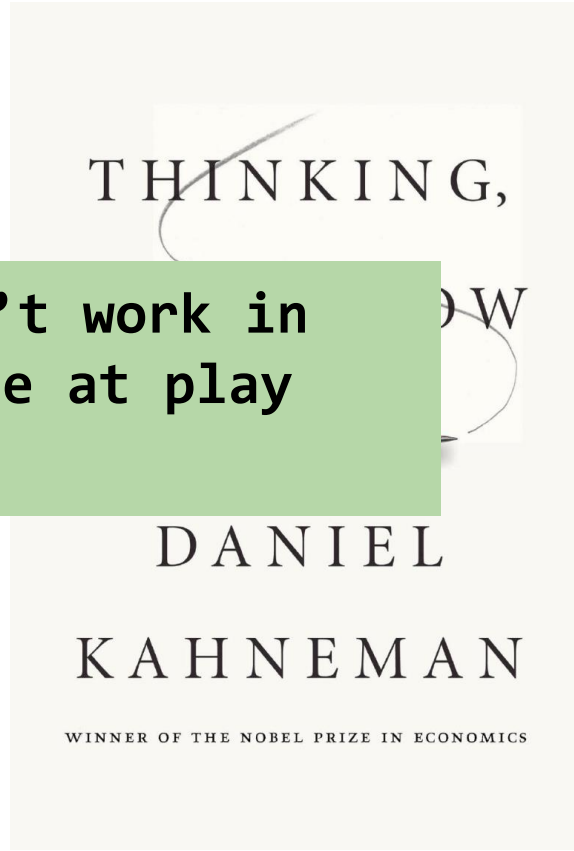
Cognitive Biases: Examples

- Feature-Positive Effect
- IKEA Effect
- (Cultural) Homophily
- Conformity Bias
- Declinism
- Primacy/Recency Effects, Position Bias
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

Cognitive Biases: Examples



Cognitive biases don’t work in isolation, rather are at play together!



https://commons.wikimedia.org/wiki/File:Cognitive_bias_codex_en.svg

https://www.visualcapitalist.com/50-cognitive-biases-in-the-modern-world

JOHANNES KEPLER UNIVERSITY LINZ

Cognitive Biases

- **Feature-Positive Effect**
- IKEA Effect
- (Cultural) Homophily
- Conformity Bias
- Declinism
- Primacy/Recency Effects, Position Bias
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

Feature-Positive Effect

Example:

- What do these lists have in common? 936, 193, 496, 829, 930, 559, 976, 139
- And these lists? 125, 922, 834, 998, 147, 980, 237, 710

Definition/Meaning:

- Humans are better at realizing (and put more emphasis on) the presence of a stimulus rather than its absence

Possible manifestations and uses in the context of ML systems:

- Possible important role in *fairness/non-discrimination*: e.g., users of an LLM might not realize that an answer is biased, e.g., some cultural group, gender, etc. is ignored
- *Explainability* through counterfactuals: e.g., which (maybe better-suited) items would have been recommended to a user if they had different traits?
- Which aspects of the data did the ML system consider during training/inference? (And, more importantly, which ones did it *not* consider?)

[Allison et al., 1988] The Feature-positive Effect, Attitude Strength, and Degree of Perceived Consensus, Personality and Social Psychology Bulletin 14, 1988, 231–241

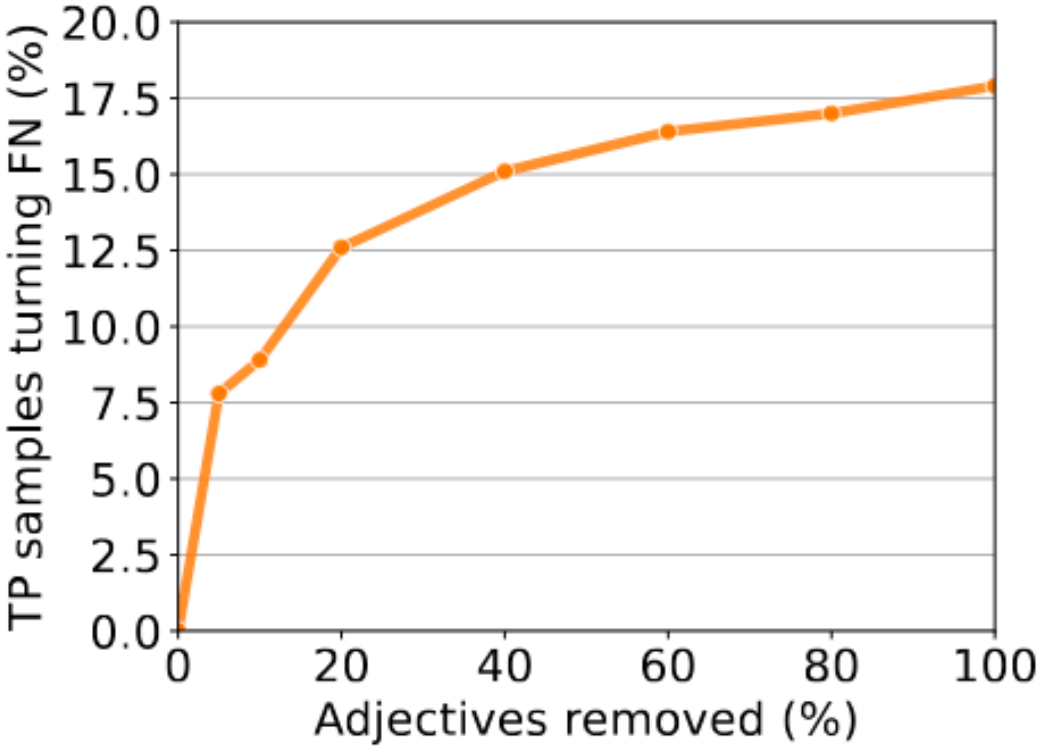
Feature-Positive Effect in Job/Candidate RecSys

- **Recruitment-related RS:** Training process may focus on what is present in job ads, overlooking what is missing
- Content-based/Text-based RecSys (matching CVs with job ads)
- Distil-RoBERTa cross-encoder model
- Employed GPT-4o to generate 2,100 CVs (350 CVs per job)
  - 6 job categories (dentist, nurse, photographer, software engineer, accountant, and teacher)
  - 1,358 samples of job advertisement from UK job board
- Trained with pairs of CV and job ad in a binary classification setup
  - For each positive sample we used 4 negative samples
  - 80% : 20% split
- Evaluate on 272 job ads and 336 unique applicants
  - Consider as positive prediction if job title in CV and job ad matches
  - 13,607 true positive (TP) and 1,625 false negative (FN) predictions

Schedl, M., Lesota, O., Masoudian, S. **The Importance of Cognitive Biases in the Recommendation Ecosystem: Evidence of Feature-Positive Effect, Ikea Effect, and Cultural Homophily**, Proceedings of the 11th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS @ RecSys 2024), Bari, Italy, October 2024. https://ceur-ws.org/Vol-3815/paper10.pdf

Feature-Positive Effect: Experiment 1

- Simulate FPE in candidate recommendation: Adjust what content the RecSys “sees” and does not “see” during training
- **Method:** Removing *adjectives* (randomly) from job ads and analyze the changes in the decisions of the candidate RecSys
  - TP : if **p** *(job ad)* then **q** *(candidate)*
  - FN : if **p** then not **q**
- **Results:** The more adjectives removed the more positive samples became negative, even though they should objectively not change result (e.g., “a passionate dentist”)
- **Conclusion:** Presence (or absence) of adjectives plays significant role in decision making of model



Feature-Positive Effect: Experiment 2

- Can FN samples become TP by leveraging adjectives that are missing in them?
- **Method**:
  - Group job ads into low-recall and high-recall group
  - Create a set of unique *adjectives* A
    - Present in high recall but missing in low recall group
    - Set A is considered as **unique information missing in the FN samples** (responsible for low recall)
  - Randomly replace adjectives from FN samples with those from A and re-evaluate the model

Group	Adjectives
Low Recall	small, referral, sexual, steady, ...
High Recall	new, full, other, good, professional, ...
Unique set	technical, annual, innovative, complex, ...

- **Results:**
  - Average score of the CE ranking model for FN samples increased from 0.046 to 0.152
  - 52.0% improvement in FN (12.9% reclassified as TPs)
- **Conclusion:** Injecting random adjectives from high-recall group can have positive effect on decisions of candidate ranking system

Feature-Positive Effect: Experiment 2

- Can FN samples become TP by leveraging adjectives that are missing in them?
- **Method**:
  - Group job ads into low-recall and high-recall group

Group	Adjectives
Low Recall	small, referral, sexual, steady, ...

Exploitation Potential: Increasing Transparency

For *recruiters*: direct feedback on how recommended applicants change when adjusting wording of job ad

For *applicants*: identify salient words in their CVs, investigate counterfactual recommendations (e.g., altering gender or work experience)

- **Conclusion:** Injecting random adjectives from high-recall group can have positive effect on decisions of candidate ranking system

Cognitive Biases

- Feature-Positive Effect
- **IKEA Effect**
- (Cultural) Homophily
- Conformity Bias
- Declinism
- Primacy/Recency Effects, Position Bias
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

IKEA Effect

Example:

- *“The cookies I baked are much tastier than the ones I bought.”*

Definition/Meaning:

- The more effort a person invested in something, the more they will value it
- Human desire to justify their efforts

Possible manifestations and uses in the context of ML systems:

- Users of streaming platforms prefer listening to content collections they (helped) create *themselves* over collections created and shared by *others*
- Generative LLMs may give higher preference scores to content they created themselves than content provided by others

[Norton et al., 2012] The IKEA Effect: When Labor Leads to Love, Journal of Consumer Psychology 22, 2012, 453–460
[Marshet al., 2018] When and How Does Labour Lead to Love? The Ontogeny and Mechanisms of the IKEA Effect, Cognition 170, 2018, 245–253

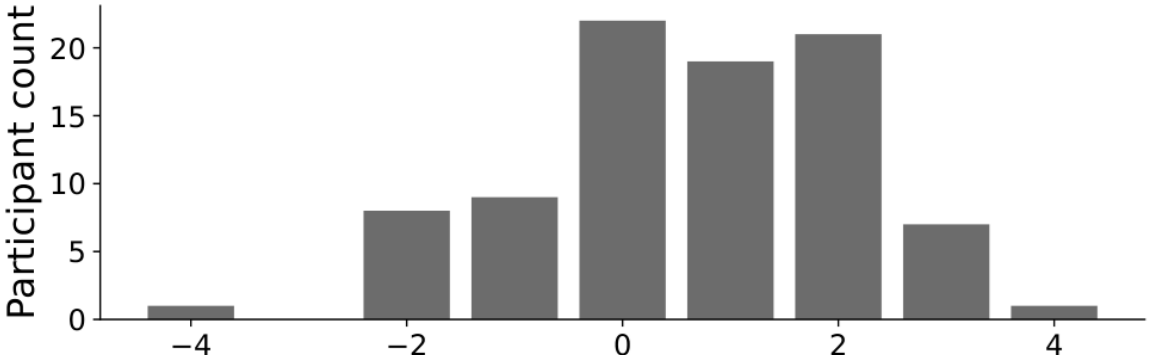
IKEA Effect in Music Playlist Generation

- **Method:**
  - User study on Prolific with 100 US users of music streaming services
  - Questionnaire with 5-point Likert scale: *Never (1) ... Very often (5)*
  - *S1: “I create or edit music collections.”*
  - *S2: “I play music collections (created by me or someone else).”*
  - *S3: “I play music collections I created or helped create myself.”*
  - *S4: “I play music collections created by someone else.”*

Schedl, M., Lesota, O., Masoudian, S. **The Importance of Cognitive Biases in the Recommendation Ecosystem: Evidence of Feature-Positive Effect, Ikea Effect, and Cultural Homophily**, Proceedings of the 11th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS @ RecSys 2024), Bari, Italy, October 2024. https://ceur-ws.org/Vol-3815/paper10.pdf

IKEA Effect in Music Playlist Generation

- **Results:**
  - Users prefer listening to their own playlists over others':
    μ (*S3*-*S4*) = 0.65 (σ = 1.52)
  - Users who invest more time creating playlists (*S1*) tend to listen more often to their own playlists (*S3*)
    Spearman's ρ (*S1*, *S3*) = 0.75
  - Users who spend more time listening to playlists in general tend to listen to the playlists they contributed to more often
    Spearman's ρ (*S2*, *S3*) = 0.66; but not to playlists created by someone else!



Distribution of the *consumption frequency difference* between own and other playlists (responses to S3-S4). Positive values show preference towards own playlists.

- **Conclusion:** Users tend to interact more with playlists they invested effort in, which we interpret as a variant of the IKEA effect

S1: "I create or edit music collections."
S2: "I play music collections (created by me or someone else)."
S3: "I play music collections I created or helped create myself."
S4: "I play music collections created by someone else."

IKEA Effect in Music Playlist Generation

- **Results:**
  - Users prefer listening to their own playlists over others':
    μ (*S3*-*S4*) = 0.65 (σ = 1.52)

Exploitation Potential: Increasing User Experience

For instance, in sequential recommendation, items present in the user's playlists (the user put effort into picking and assigning them) can serve as *anchors* to retain user engagement within the current listening session. Using them for *explanations* could foster user trust in RecSys.

Spearman's ρ (*S2*, *S3*) = 0.66

- **Conclusion:** Users tend to interact more with playlists they invested effort in, which we interpret as a variant of the IKEA effect

S1: "I create or edit music collections."
S2: "I play music collections (created by me or someone else)."
S3: "I play music collections I created or helped create myself."
S4: "I play music collections created by someone else."

Cognitive Biases

- Feature-Positive Effect
- IKEA Effect
- **(Cultural) Homophily**
- Conformity Bias
- Declinism
- Primacy/Recency Effects, Position Bias
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

Homophily (Social/Cultural)

Example:

- *“I use to hang out with my friends because they are liberals and love reggae music.”*

Definition/Meaning:

- Humans tend to associate and form connections with others who have similar characteristics (e.g., age, culture, or religion) more often than with people who have different traits

Possible manifestations and uses in the context of ML systems:

- Users with a specific trait (e.g., country, culture, or social group) may prefer content created by producers with the same trait (e.g., domestic vs. foreign music consumption)
- Generative LLMs may produce content that is biased towards traits of its users, esp. when included in the prompt
- If queried for a particular group of people (e.g., researchers working on cognitive biases), the result of LLMs may be biased towards people with similar traits

[Mark, 2003] Culture and Competition: Homophily and Distancing Explanations for Cultural Niches, American Sociological Review 68, 2003, 319–345

Cultural Homophily in Music

Cultural homophily in music *consumption*, *recommendation*, and simulated *feedback loop*

- **Method:**
  - LFM-2b dataset (subsample: 2018-2019, 5-core-filtered)
  - ~100K songs, ~12K users, ~2.3M interactions
  - Artists' countries retrieved from MusicBrainz
  - MultVAE as base recommender
  - Feedback loop *simulation* with simple choice model (select one recommended item)

Schedl, M., Lesota, O., Masoudian, S. **The Importance of Cognitive Biases in the Recommendation Ecosystem: Evidence of Feature-Positive Effect, Ikea Effect, and Cultural Homophily**, Proceedings of the 11th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS @ RecSys 2024), Bari, Italy, October 2024. https://ceur-ws.org/Vol-3815/paper10.pdf

Homophily in Music Consumption

	base	*Con*	*Con/base*
US	**0.397**	**0.626**	1.578
UK	0.155	0.266	1.713
DE	0.068	0.169	2.481
SE	0.045	0.159	3.519
CA	0.038	0.083	2.202
FR	0.028	0.091	3.232
AU	0.023	0.077	3.289
FI	0.023	0.170	**7.536**
BR	0.022	0.141	6.288
RU	0.019	0.073	3.870

Proportions of domestic music among all available tracks ($base$), among consumed tracks by users from the country (Con), and in relation ($Con/base$)

Ex.:

~40% of all tracks on a music streaming platform have been created by US artists

~63% of tracks consumed by US users have been created by US artists

→ Significant effect for all investigated countries, but particularly for *FI* and *BR*

Homophily in Music Recommendation

Cultural homophily in music consumption, recommendation, and simulated feedback loop

- **Results:**

	base	*Con*	*Con/base*	Rec_1	$Rec_1/base$	Rec_{20}	$Rec_{20}/base$
US	**0.397**	**0.626**	1.578	**0.629**	1.587	**0.595**	1.501
UK	0.155	0.266	1.713	0.227	1.458	0.232	1.495
DE	0.068	0.169	2.481	0.176	2.590	0.166	2.439
SE	0.045	0.159	3.519	0.102	2.266	0.088	1.948
CA	0.038	0.083	2.202	0.030	0.797	0.041	1.091
FR	0.028	0.091	3.232	0.039	1.377	0.041	1.447
AU	0.023	0.077	3.289	0.017	0.728	0.026	1.103
FI	0.023	0.170	**7.536**	0.166	7.325	0.132	5.820
BR	0.022	0.141	6.288	0.187	**8.347**	0.150	**6.714**
RU	0.019	0.073	3.870	0.081	4.262	0.066	3.515

Proportions of domestic music among all available tracks ($base$), among consumed tracks by users from the country (Con), and among recommender tracks (Rec) at iteration *1* and *20* of the simulation

JOHANNES KEPLER UNIVERSITY LINZ

Homophily in Music Consumption and Rec.

Cultural homophily in music consumption, recommendation, and simulated feedback loop

- **Conclusion:**
  - Users listen more frequently to music originating from their own country than a random choice would warrant
  - Effect strength varies strongly between countries (cf. US, UK vs. FI, BR)
  - RecSys and feedback loops can have some leveraging effect for cultural homophily (e.g. SE, FI)
  - In some cases (e.g. CA, AU), RecSys even introduces a "homophobic" behavior w.r.t. domestic recommendations

	base	*Con*	*Con/base*	Rec_1	$Rec_1/base$	Rec_{20}	$Rec_{20}/base$
US	**0.397**	**0.626**	1.578	**0.629**	1.587	**0.595**	1.501
UK	0.155	0.266	1.713	0.227	1.458	0.232	1.495
DE	0.068	0.169	2.481	0.176	2.590	0.166	2.439
SE	0.045	0.159	3.519	0.102	2.266	0.088	1.948
CA	0.038	0.083	2.202	0.030	0.797	0.041	1.091
FR	0.028	0.091	3.232	0.039	1.377	0.041	1.447
AU	0.023	0.077	3.289	0.017	0.728	0.026	1.103
FI	0.023	0.170	**7.536**	0.166	7.325	0.132	5.820
BR	0.022	0.141	6.288	0.187	**8.347**	0.150	**6.714**
RU	0.019	0.073	3.870	0.081	4.262	0.066	3.515

Homophily in Music Consumption and Rec.

Cultural homophily in music consumption, recommendation, and simulated feedback loop

- **Conclusion:**
  - Users listen more frequently to music originating from their own country than a random choice would warrant

Exploitation Potential: Diversification and Calibration

Formalized *homophily models* as additional indicator of user taste could be useful for: (1) diversification of recommendations, (2) calibration between user profiles and recommendations, in terms of country, etc.

DE	0.068	0.169	2.481	0.176	2.590	0.166	2.439
SE	0.045	0.159	3.519	0.102	2.266	0.088	1.948
CA	0.038	0.083	2.202	0.030	0.797	0.041	1.091
FR	0.028	0.091	3.232	0.039	1.377	0.041	1.447
AU	0.023	0.077	3.289	0.017	0.728	0.026	1.103
FI	0.023	0.170	**7.536**	0.166	7.325	0.132	5.820
BR	0.022	0.141	6.288	0.187	**8.347**	0.150	**6.714**
RU	0.019	0.073	3.870	0.081	4.262	0.066	3.515

Cognitive Biases

- Feature-Positive Effect
- IKEA Effect
- (Cultural) Homophily
- **Conformity Bias**
- Declinism
- Primacy/Recency Effects, Position Bias
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

Conformity Bias

Example:

- In a meeting: *I am quite sure all the others are wrong, but I won't raise my voice; don't want to cause the meeting to last forever; and the others may be right anyway (Why else would they all have an opposite opinion to mine?)*

Definition/Meaning:

- Tendency of individuals to align their beliefs, behaviors, and actions with those of a group, often disregarding their own independent judgment

Possible manifestations and uses in the context of ML systems:

- Showing users (artificial or true) *averaged ratings* before asking them to provide their own ratings on an item changes their behavior towards the shown ones
- Users are more likely to click on an item if they see that *many other users* clicked on it

[Adomavicius et al., 2011] Recommender Systems, Consumer Preferences, and Anchoring Effects, Proceedings of the Workshop on Human Decision Making in Recommender Systems, 2011, pp. 35–42.
[Zheng et al., 2021] Disentangling User Interest and Conformity for Recommendation with Causal Embedding, Proceedings of The Web Conference, 2021, pp. 2980–2991.
[Ma et al., 2024] Temporal Conformity-aware Hawkes Graph Network for Recommendations, Proceedings of The Web Conference, 2024, pp. 3185–3194.

Conformity Bias

Example:

- In a meeting: *I am quite sure all the others are wrong, but I won't raise my voice; don't want to cause the meeting to last forever; and the others may be right anyway (Why*

Exploitation Potential: Influencing Rating/Consumption Behavior

- Showing them adjusted (or even fake) ratings could *trick users* into believing their preference towards an item is higher or lower than it actually is.

+ Confronting users with their change in rating behavior (given them as reference their typical rating for highly similar items) may also serve to *raise awareness* of the phenomenon.

- Users are more likely to click on an item if they see that *many other users* clicked on it

[Adomavicius et al., 2011] Recommender Systems, Consumer Preferences, and Anchoring Effects, Proceedings of the Workshop on Human Decision Making in Recommender Systems, 2011, pp. 35–42.
[Zheng et al., 2021] Disentangling User Interest and Conformity for Recommendation with Causal Embedding, Proceedings of The Web Conference, 2021, pp. 2980–2991.
[Ma et al., 2024] Temporal Conformity-aware Hawkes Graph Network for Recommendations, Proceedings of The Web Conference, 2024, pp. 3185–3194.

Cognitive Biases

- Feature-Positive Effect
- IKEA Effect
- (Cultural) Homophily
- Conformity Bias
- **Declinism**
- Primacy/Recency Effects, Position Bias
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

Declinism

Example:

- *"Music used to be much better in the 90s."*
- *"The world was a much better place when I was a teenager than today!"*

Definition/Meaning:

- The perception that the world or society is declining, i.e., things get worse over time
- Partly the result of *rosy retrospection* — humans' tendency to remember the past as more positive as it actually was

Possible manifestations and uses in the context of ML systems:

- Identifying trends, e.g., in sentiment (positive or negative) in lyrics, social media or news articles, tags, etc.; formalize them via statistical models
- Can these models be used to adjust outcomes, to counteract (or amplify) declinism?
- Is declinism reflected in interaction logs (used as training data) with news or music (spanning decades), extracted from item or user *side information*?

[Mitchell et al., 1997] Temporal Adjustments in the Evaluation of Events: The "Rosy View", Journal of Experimental Social Psychology 33, 421-448, 1997

Declinism in Music Lyrics

- **Method:**
  - 353,320 songs from LFM-2b
  - 5 genres (Pop, Rock, Rap, Country, R&B), 5 decades (1970-2020)
  - Lyrics from Genius.com
  - LIWC dictionary to describe emotions („positive emotions“)
  - Linear regression on (positive/negative) emotions over all years

- **Results:**
  - Increase of negative emotions: Rap (*m=0.0217*), R&B *(m=0.0187)*
  - Decrease of positive emotions: R&B *(m=−0.048552)*, Country *(m=−0.0217)*

- **Conclusion:**
  - Clear overall trend towards more positive and less negative emotions in the past

Parada-Cabaleiro, E., Mayerl, M., Brandl, S., Skowron, M., Schedl, M., Lex, E., Zangerle, E. **Song Lyrics Have Become Simpler and More Repetitive over the Last Five Decades**, Nature Scientific Reports, 14, 5531, 2024. https://www.nature.com/articles/s41598-024-55742-x

Declinism

- **Method:**
  - 353,320 songs from LFM-2b
  - 5 genres (Pop, Rock, Rap, Country, R&B), 5 decades (1970-2020)
  - Lyrics from Genius.com
  - LIWC dictionary to describe emotions („positive emotions“)

Exploitation Potential: Adjust Level of Positiveness/Negativeness

Together with users’ interaction history, fine-granular information on declinism (e.g., for different content categories) could help create *personalized long-term declinism models*, used to tailor recommendations.

- **Conclusion:**
  - Clear overall trend towards more positive and less negative emotions in the past

Cognitive Biases

- Feature-Positive Effect
- IKEA Effect
- (Cultural) Homophily
- Conformity Bias
- Declinism
- **Primacy/Recency Effects, Position Bias**
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

Primacy/Recency Effects, Position Bias

Example:

- Which of the animals shown on the previous slides can you name?

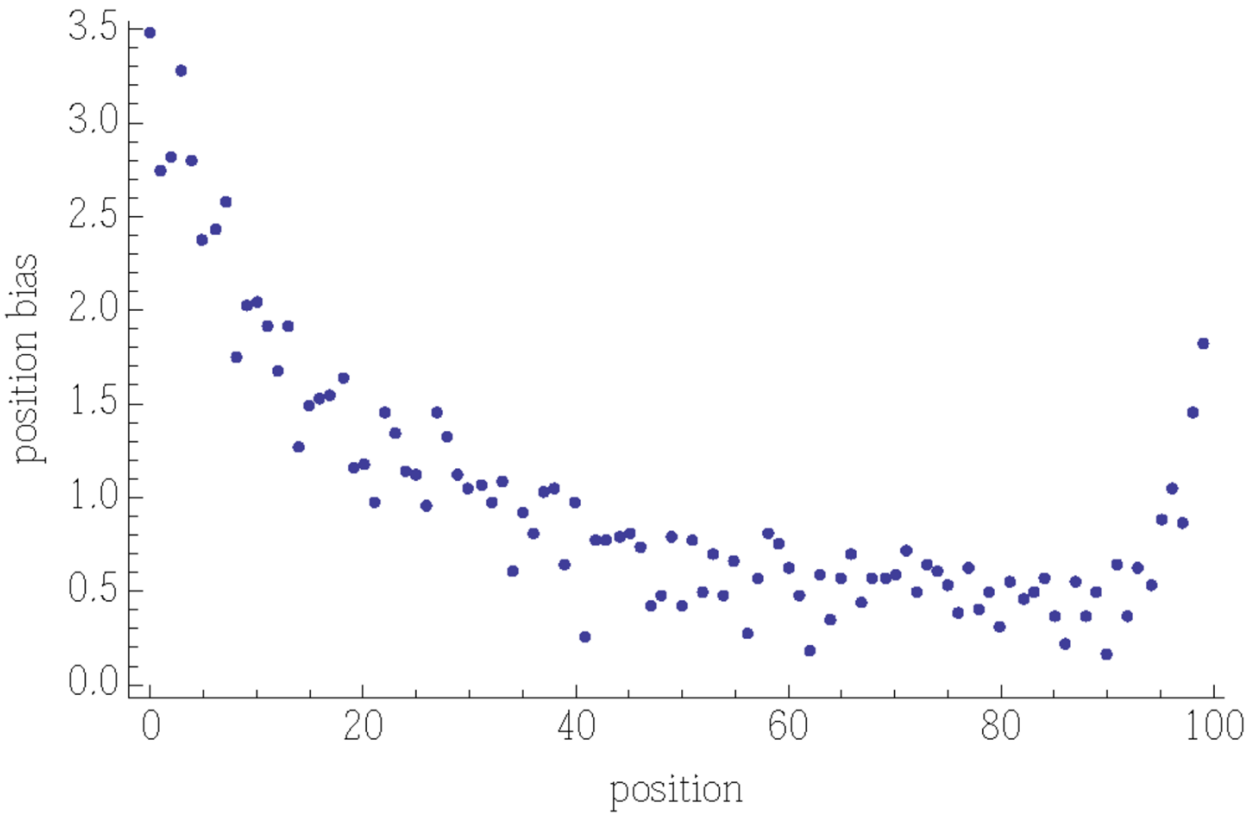
Definition/Meaning:

- Human tendency to easier recall first and last items from a sequence as opposed to the items from the middle of the sequence

Possible manifestations and uses in the context of ML systems:

- Users are more likely to interact with items appearing at the beginning (primacy effect) and at the end (recency effect) of a list of recommendations or retrieved documents
- Negative effect in terms of expose for mid-ranked items
- To which extent does position bias depend on the algorithm, recommendation task, and presentation of results (UI)? (e.g., top-N recommendations vs. endless list)
- Can we counteract this effect by algorithmic in-processing or post-processing techniques (e.g. reranking)?

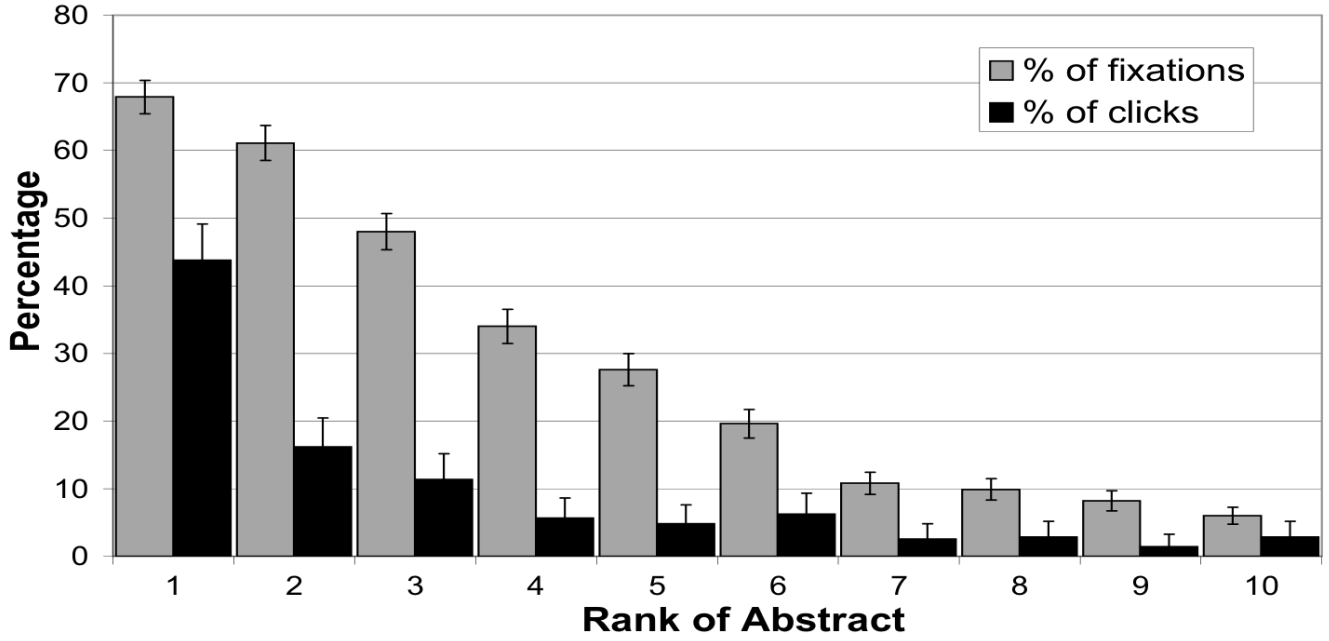
Primacy/Recency Effects in Story Recommendation



Relative increase or decrease in number of ratings (votes) for each position of an item (story) in the recommendation list, compared to the average number

Lerman, K. and Hogg, T. **Leveraging Position Bias to Improve Peer Recommendation**, PLoS ONE 9(6): e98914, 2014.
https://doi.org/10.1371/journal.pone.0098914

Primacy/Recency Effects in Web Search



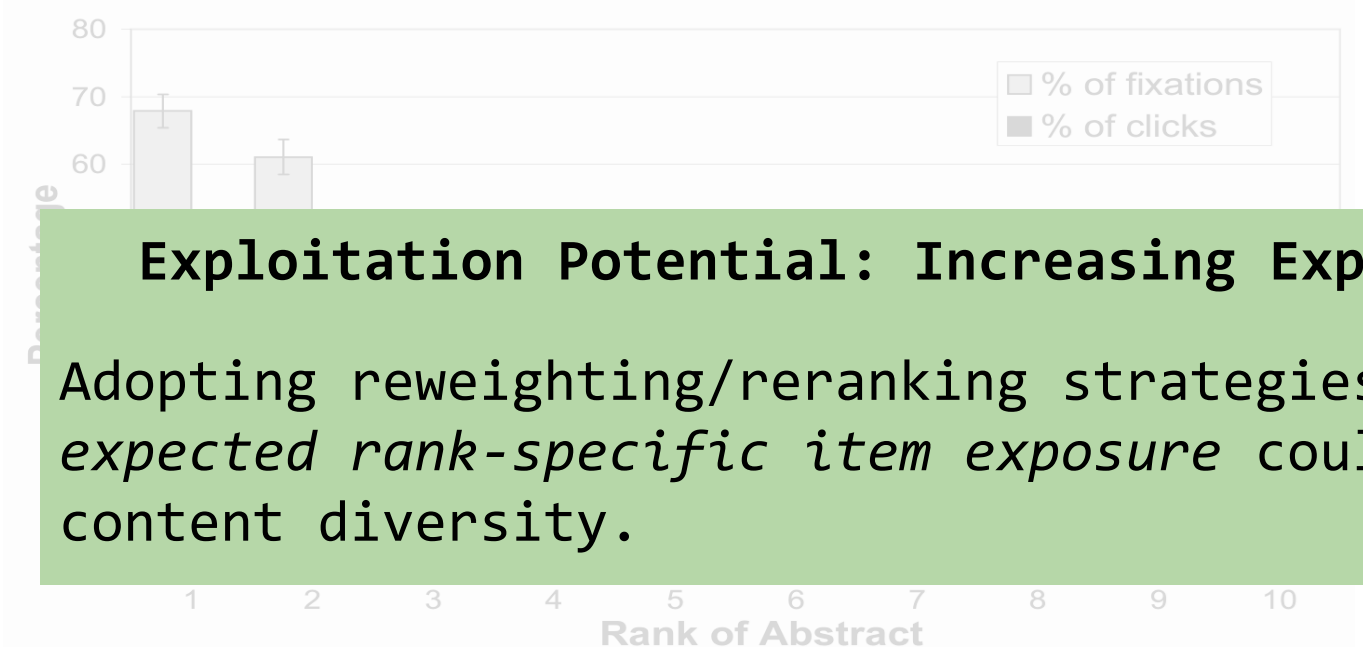
Percentage of times an abstract was viewed or clicked, depending on the rank of the retrieved document (using Google Search)

Trust Bias:
More clicks on links ranked highly by Google, even if those abstracts are less relevant than other abstracts the user viewed

Quality Bias:
Users' clicking decision is not only influenced by the relevance of the clicked link, but also by the overall quality of the other abstracts in the ranking

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G. **Accurately Interpreting Clickthrough Data as Implicit Feedback**, Proceedings of SIGIR, pp. 154-161, 2005. https://doi.org/10.1145/1076034.1076063

Primacy/Recency Effects



Percentage of times an abstract was viewed or clicked, depending on the rank of the retrieved document (using Google Search)

Trust Bias:
More clicks on links ranked highly by Google, even if those abstracts are less relevant than

Exploitation Potential: Increasing Exposure of Underexposed Contents

Adopting reweighting/reranking strategies to *balance relevance and expected rank-specific item exposure* could increase creator fairness and content diversity.

Cognitive Biases

- Feature-Positive Effect
- IKEA Effect
- (Cultural) Homophily
- Conformity Bias
- Declinism
- Primacy/Recency Effects, Position Bias
- **Bandwagon Effect, Popularity Bias**
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

Bandwagon Effect, Popularity Bias

Example:

- *“I don’t really like this new fashion style, but it has become so popular that I can’t resist.”*
- *“Should I buy that stock? Many others bought it, so it must be great even if it's overpriced.”*

Definition/Meaning:

- Human tendency of adopting certain behaviors or beliefs because many other people do the same (“hop on the bandwagon”)

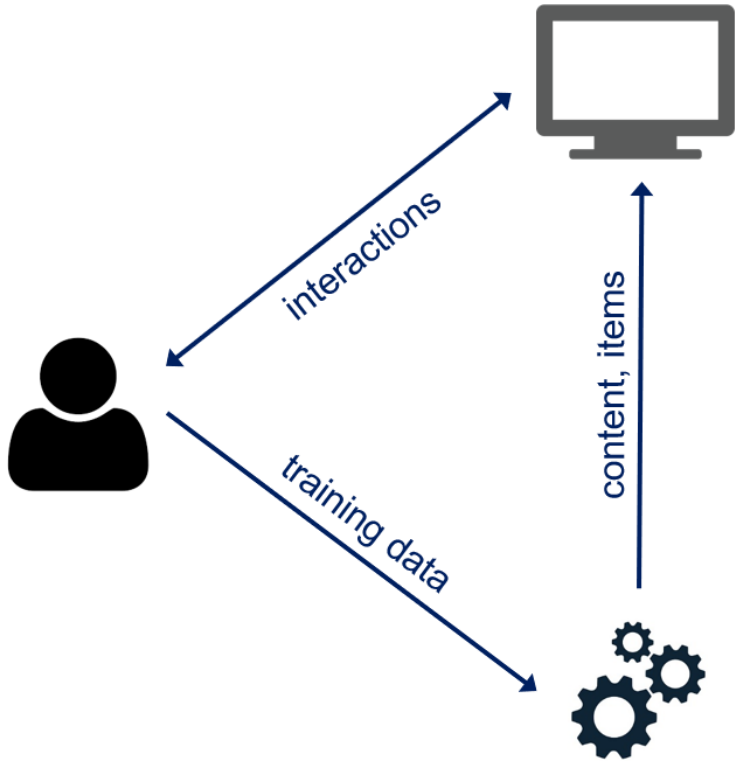
Possible manifestations and uses in the context of ML systems:

- Overly many user-item interactions with popular items (in training data) may result in a popularity-biased ranking model, which in turn favors already popular content
- Due to their higher exposure to popular content during training, LLMs could pick up this bias and reproduce it at generation stage

[Kiss and Simonovits, 2014] Identifying the Bandwagon Effect in Two-round Elections, Public Choice 160, 327-344, 2014
[Shyam Sundar, S. et al., 2008] The Bandwagon Effect of Collaborative Filtering Technology, CHI Extended Abstracts, 3453-3458, 2008
[Knyazev and Oosterhuis, 2022] The Bandwagon Effect: Not Just Another Bias, Proceedings of ICTIR 2022: 243-253

Popularity Bias in Recommendation

Problem: Reinforcing already popular items/content, while limiting exposure of less popular ones (harmful for content creators and users)
→ “Rich-get-richer effect”



Di Noia, T., Tintarev, N., Fatourou, P., and Schedl, M. **Recommender Systems Under European AI Regulations**, Communications of the ACM, volume 65, issue 4, pp. 69-73, 2022. https://dl.acm.org/doi/10.1145/3512728

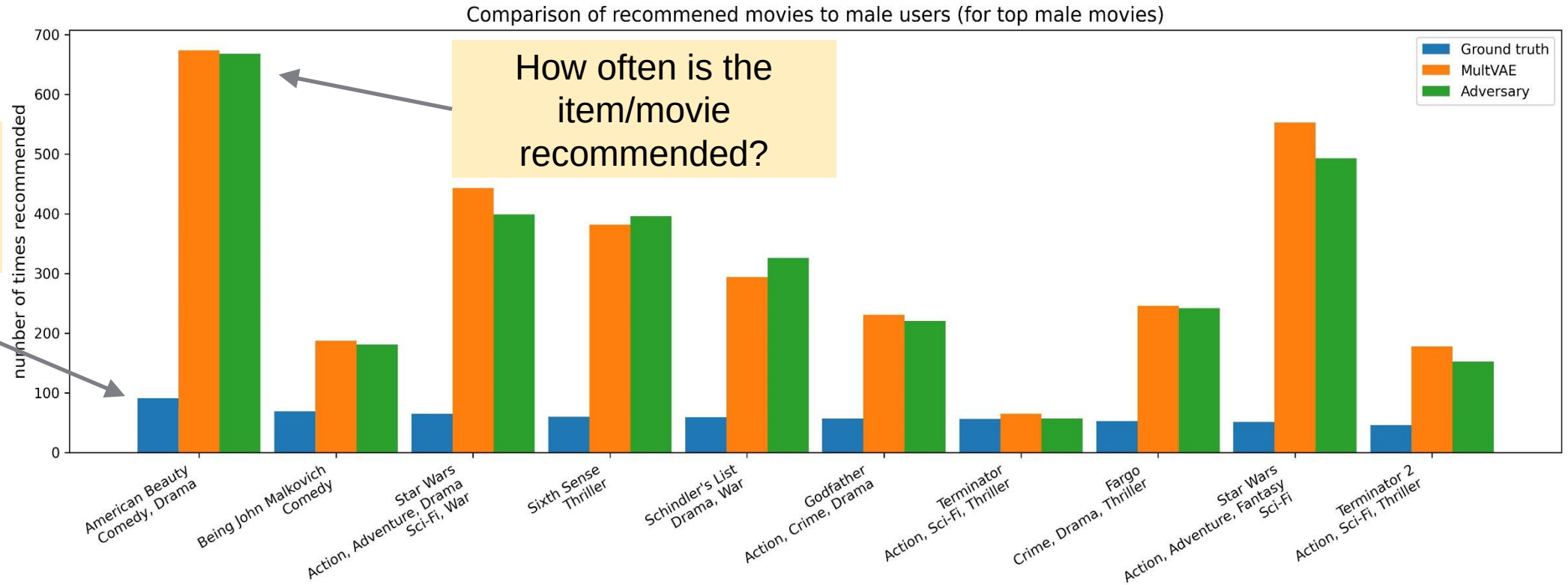
Popularity Bias: How to Measure It?

Ad-hoc variant: Difference between an item's recommendation frequency and consumption frequency in user profiles

Shortcoming: Does not take into account the user's individual preference for popular content

How often is the item/movie consumed?

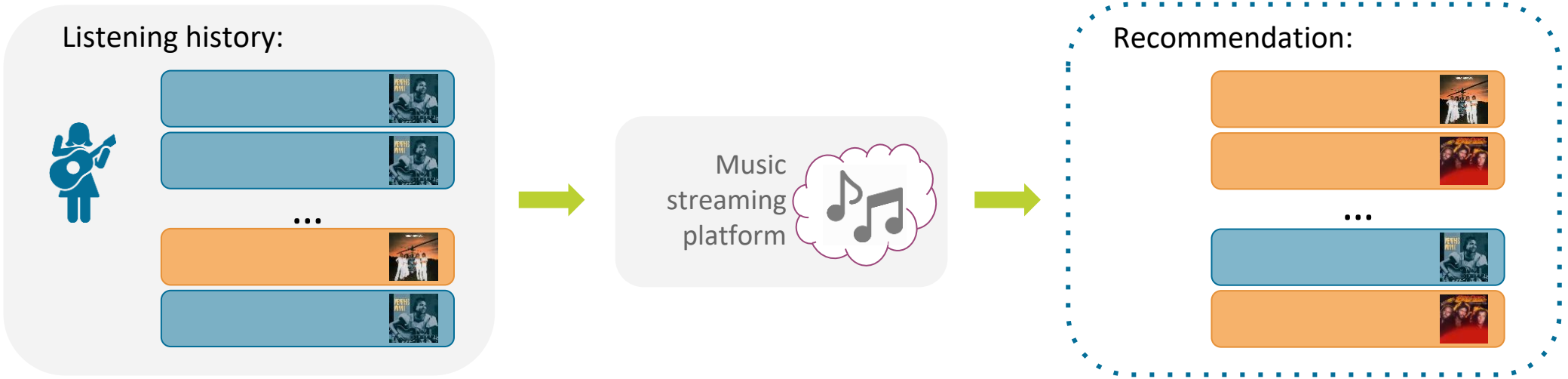
How often is the item/movie recommended?

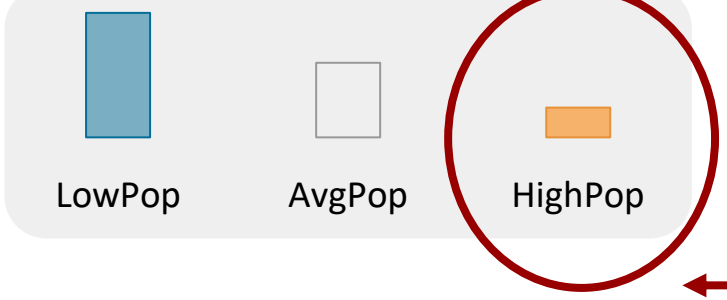


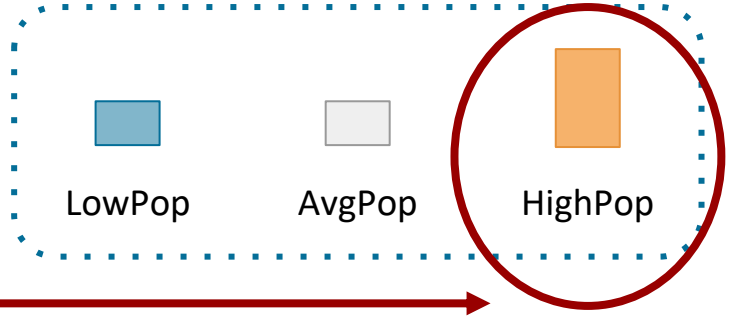
Popularity Bias in Music Recommendation

27.4M monthly listeners on Spotify

0.3M monthly listeners on Spotify







Popularity bias/miscalibration

JOHANNES KEPLER UNIVERSITY LINZ

Measuring Popularity Bias

Assumption: Users prefer “calibrated” recommendations, i.e., the RS should mimic the interaction distribution w.r.t. an attribute (popularity in our case): $pop(H_{u_i}) \sim pop(R_{u_i})$.

pop	some measure of popularity (e.g., total number of interactions with items, number of interacting users)
H_{u_i}	historical interaction list of user u_i’s over items
R_{u_i}	recommendation list created for user u_i (top recommendations at fixed cut-off)

Delta metrics: *statistical moments* of popularity differences between items in H_{u_i} and R_{u_i}

Distribution-based metrics: difference between popularity distributions (e.g., KL divergence or Kendall’s τ)

Lesota, O., Melchiorre, A., Rekabsaz, N., Brandl, S., Kowald, D., Lex, E., and Schedl, M. **Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected?**, Proceedings of the 15th ACM Conference on Recommender Systems (RecSys 2021), Amsterdam, the Netherlands, October-November 2021. https://dl.acm.org/doi/10.1145/3460231.3478843

Measuring Popularity Bias: Delta Metrics

$$\%\Delta\xi(u_i) = \frac{\xi\left(R_{u_i}\right) - \xi\left(H_{u_i}\right)}{\xi\left(H_{u_i}\right)} \cdot 100$$

$\%\Delta\xi$ relative popularity difference between items in H_{u_i} and R_{u_i} in terms of statistical measure ξ (e.g., mean, median, variance, skew)

Aggregate over all users (bias of the RS): $\%\Delta\xi = \ Median\left(\%\Delta\xi(u_i)\right)$

→ Positive $\%\Delta Mean$ and $\%\Delta Median$ indicate that more popular items are recommended to user u_i than warranted given his or her consumption history ("miscalibration").
→ Positive $\%\Delta Variance$ indicate that recommendation list is more diverse w.r.t. covering differently popular items than user u_i's consumption history.

Measuring Popularity Bias: Distribution-based

$$JSD\left(H_{u_i}, R_{u_i}\right) = \frac{1}{2} \cdot \sum_p H_{u_i}(p) \cdot \log_2 \frac{2H_{u_i}(p)}{H_{u_i}(p) + R_{u_i}(p)} + \frac{1}{2} \cdot \sum_p R_{u_i}(p) \cdot \log_2 \frac{2R_{u_i}(p)}{H_{u_i}(p) + R_{u_i}(p)}$$

JSD Jensen-Shannon Divergence quantifies distribution mismatch of popularity distributions

Popularity Bias: Results on LFM-2b

Alg.	Users	%ΔMean	%ΔMedian	%ΔVar.	%ΔSkew	%ΔKurtosis	KL	Kendall's τ	NDCG@10
RAND	*All*	−91.8	−87.2	−99.5	11.5	15.3	3.904	0.165	0.000
	Δ*Female*	−1.8	−3.5	−0.2	+0.0	−3.5	+0.976	−0.189	−0.000
	Δ*Male*	+0.5	+1.1	+0.1	−0.0	+1.3	−0.281	+0.053	+0.000
POP	*All*	432.5	975.2	487.0	−58.0	−87.0	6.023	0.057	0.045
	Δ*Female*	+11.0	+282.1	−172.2	−2.1	−1.9	+1.626	−0.033	+0.003
	Δ*Male*	−2.8	−115.8	+55.9	+0.5	+0.5	−0.380	+0.016	−0.001
ALS	*All*	121.8	316.6	72.6	−25.2	−43.9	4.368	0.046	0.184
	Δ*Female*	+9.9	+27.4	−7.1	−3.2	−5.4	+0.467	+0.110	−0.017
	Δ*Male*	−2.7	−6.6	+1.6	+0.8	+1.5	−0.121	−0.023	+0.005
BPR	*All*	−49.0	−3.7	−87.4	−14.8	−29.4	1.202	0.268	0.129
	Δ*Female*	+5.2	+7.7	+2.1	−1.4	−3.9	+0.476	−0.043	−0.011
	Δ*Male*	−1.1	−1.9	−0.6	+0.4	+1.1	−0.110	+0.010	+0.003
ItemKNN	*All*	9.6	4.6	5.7	−14.3	−29.0	0.175	0.423	0.301
	Δ*Female*	+2.0	+5.8	−2.6	−2.1	−3.2	+0.128	−0.037	−0.042
	Δ*Male*	−0.5	−1.3	+0.9	+0.8	+0.9	−0.020	+0.008	+0.012
SLIM	*All*	49.8	99.8	56.0	−12.5	−26.0	0.424	0.189	0.365
	Δ*Female*	−6.4	−13.1	−17.4	−1.7	−4.6	+0.217	+0.052	−0.048
	Δ*Male*	+1.9	+3.9	+5.6	+0.6	+1.1	−0.029	−0.012	+0.014
VAE	*All*	303.9	736.3	351.0	−45.2	−70.1	4.823	−0.028	0.191
	Δ*Female*	+10.1	+56.4	−69.3	−6.2	−6.6	+0.633	+0.146	−0.020
	Δ*Male*	−2.3	−20.4	+17.3	+1.8	+2.1	−0.161	−0.042	+0.006

- Most RS algorithms are prone to **popularity bias** ($\%\Delta Mean$)
- Some algorithms are affected more than others
- Most RSs create a higher popularity bias for female than male users, pointing to **demographic bias** (+/- values are relative to values in row *All*)

https://ars.electronica.art/planetb/en/black-holes-of-popularity

Black Holes of Popularity

- Artistic/scientific project presented at Ars Electronica Festival of Media Arts 2022
- Raising awareness of artist popularity bias in music recommendation
- Exploration of music via genre, using metaphor of a universe
- Cosmic bodies represent songs with varying levels of popularity (planets, stars, black holes)
- User interacts by means of a lifebuoy with planets and stars, selecting which ones to save from being eaten by the black hole
- Influence of user's song saving activities is computed by in/decrease of fairness score, shown to the user
- Explanatory video: https://bit.ly/3VBAbqT

JOHANNES KEPLER UNIVERSITY LINZ

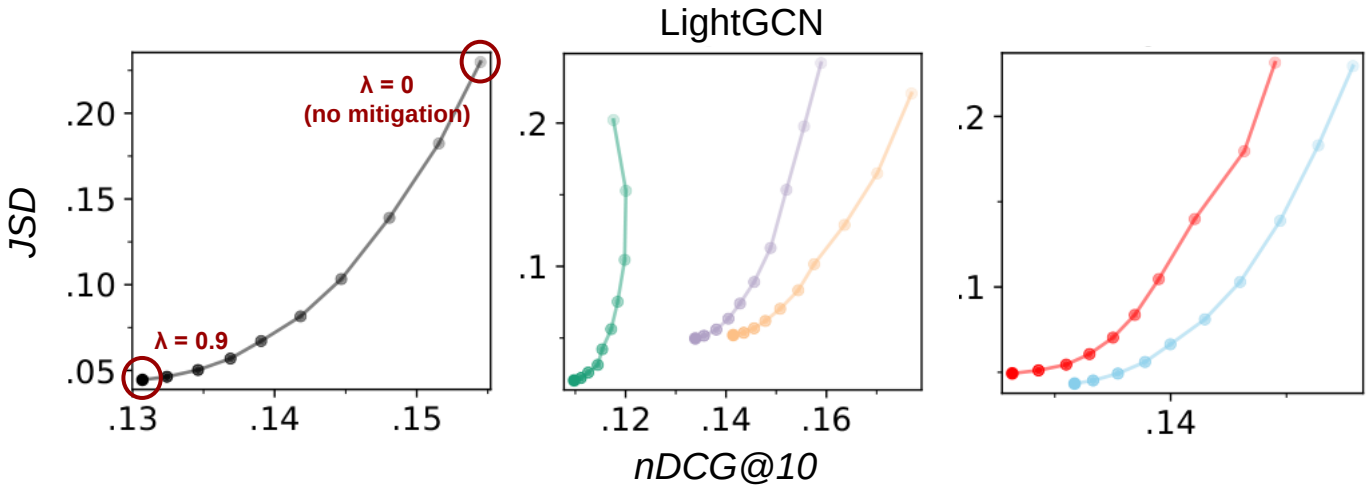
Mitigating Popularity Bias (Post-processing)

Idea: Reduce difference in popularity distribution of items in user u_i's historical interactions H_{u_i} and recommendation list R_{u_i}

Method: Create a personalized popularity-aware recommendation list $R^*_{u_i}$ by optimizing

$$R^*_{u_i} = \arg\max_{L_{u_i}} (1-\lambda) \cdot Rel(L_{u_i}) - \lambda \cdot JSD\left(H_{u_i}, L_{u_i}\right),\ L_{u_i} \subset R_{u_i},$$

λ strength of bias mitigation



- Trade-off between popularity bias (JSD) and recommendation accuracy (NDCG@10) is different for users preferring **HighPop**, **LowPop**, or **AvgPop** content; as well as for **male** and **female** users
- λ can be adjusted depending on the user group to optimize trade-off

Cognitive Biases: Examples

- Feature-Positive Effect
- IKEA Effect
- (Cultural) Homophily
- Conformity Bias
- Declinism
- Primacy/Recency Effects, Position Bias
- Bandwagon Effect, Popularity Bias
- Anchoring, Decoy Effect
- Confirmation Bias
- Authority Bias
- Halo Effect

Conclusions and Open Challenges

- Strong evidence of various cognitive biases in algorithmic decision making processes
- Most studies face several limitations (e.g., only single of few domains, standard top-N recommendation scenario, ignoring confounding factors)

- How to (mathematically) *formalize* accurate models of cognitive biases?
- Which CoBis are *intertwined* and how does their entanglement manifest?
- Which CoBis are important for different RecSys *stakeholders*?
- What role does the *user interface* play?
- How do CoBis manifest in *other retrieval and recommendation tasks and domains*, e.g., sequential recommendation; video, travel, people?

We advocate for a holistic discussion of *both negative and positive effects of cognitive biases*, and for new approaches to algorithmic decision making that mitigate the former while leveraging the latter.

We advocate for a holistic discussion of *both negative and positive effects of cognitive biases*, and for new approaches to algorithmic decision making that mitigate the former while leveraging the latter.

Thank You!

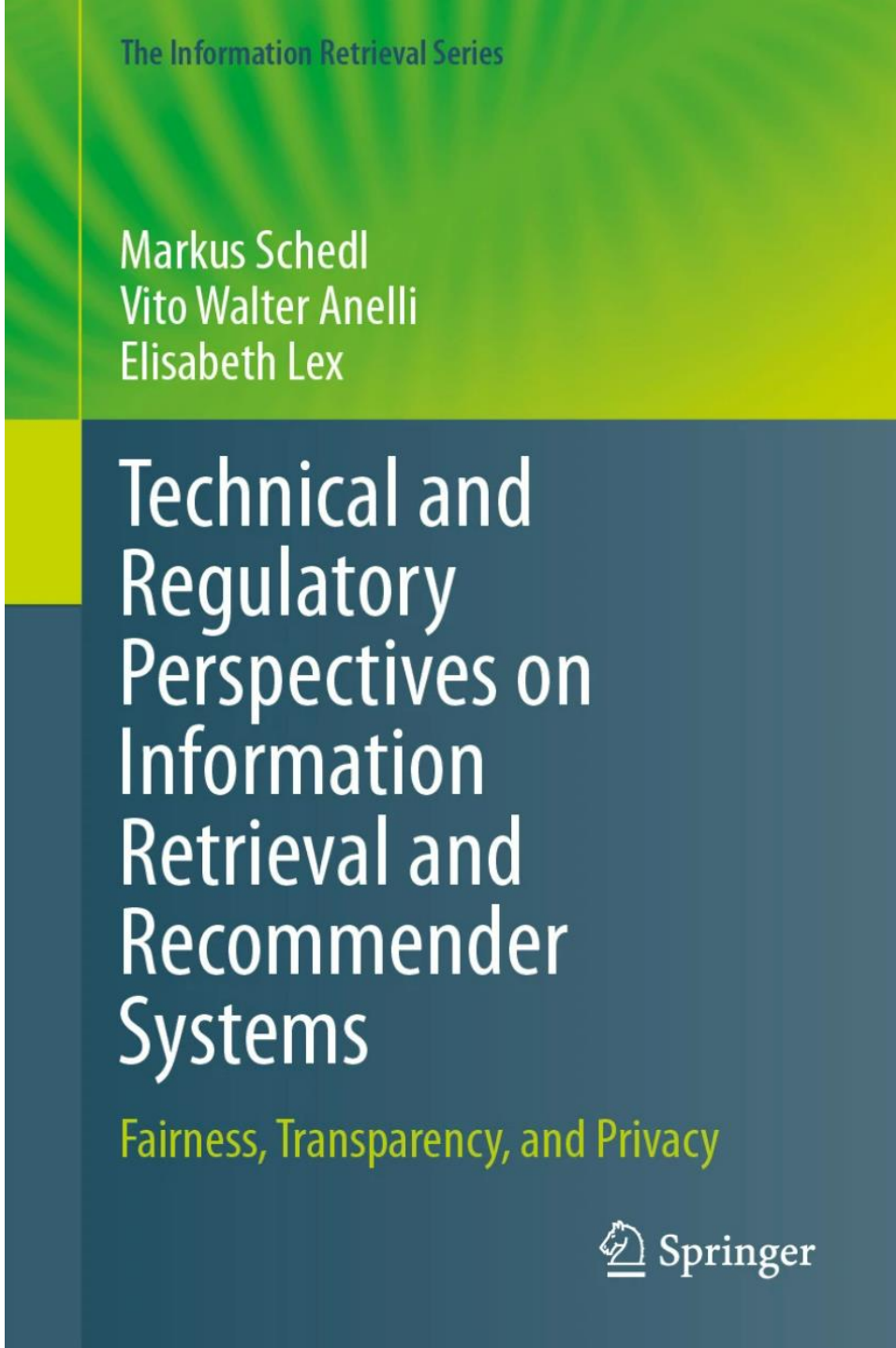
Markus Schedl
Johannes Kepler University Linz, Austria
Linz Institute of Technology, Austria
markus.schedl@jku.at | www.mschedl.eu | www.hcai.at

JOHANNES KEPLER UNIVERSITY LINZ

Springer © 2025

Especially the bias/fairness part strongly relates to cognitive (and other) biases and their relation to fairness of IR and RSs

https://link.springer.com/book/10.1007/978-3-031-69978-8



JOHANNES KEPLER UNIVERSITY LINZ