Trustworthy Algorithmic Ranking Systems

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

JOHANNES KEPLER UNIVERSITÄT LINZ

Emilia Gómez
Joint Research Centre, European Commission
Universitat Pompeu Fabra, Spain
emilia.gomez-gutierrez@ec.europa.eu | https://emiliagomez.com | @emiliagogu

European Commission
Joint Research Centre

Elisabeth Lex
Graz University of Technology, Austria
elisabeth.lex@tugraz.at | https://elisabethlex.info

TU Graz

About Markus Schedl

JOHANNES KEPLER UNIVERSITÄT LINZ

INSTITUTE OF COMPUTATIONAL PERCEPTION

MULTIMEDIA MINING AND SEARCH GROUP

- Full Professor at Johannes Kepler University (JKU) Linz, Austria
- Head of *Multimedia Mining and Search* (MMS) group at Institute of Computational Perception
- Head of *Human-centered Artificial Intelligence* (HCAI) group at Linz Institute of Technology (LIT), AI Lab
- Lab: https://hcai.at | https://www.jku.at/en/institute-of-computational-perception
- Interests: recommender systems, user modeling, information retrieval, machine learning, natural language processing, multimedia, data analysis, and web mining

Contact: markus.schedl@jku.at | www.mschedl.eu | @m_schedl

About Emilia Gómez

European Commission
Joint Research Centre

upf. Universitat Pompeu Fabra Barcelona

MTG Music Technology Group

ISMIR

- PI HUMAINT & Lead Scientist at European Centre for Algorithmic Transparency, European Commission, Joint Research Centre, European Commission.
- Guest Professor, Music Information Research Lab, Universitat Pompeu Fabra, Barcelona. https://www.upf.edu/web/mtg/
- Bsc/MSc in Engineering, PhD in Information Retrieval.
- Interests: music IR, content-based description, recommender systems, social and ethical impact, science for policy and regulatory approaches.

Contact: emilia.gomez-gutierrez@ec.europa.eu | emiliagomez.com | @emiliagogu

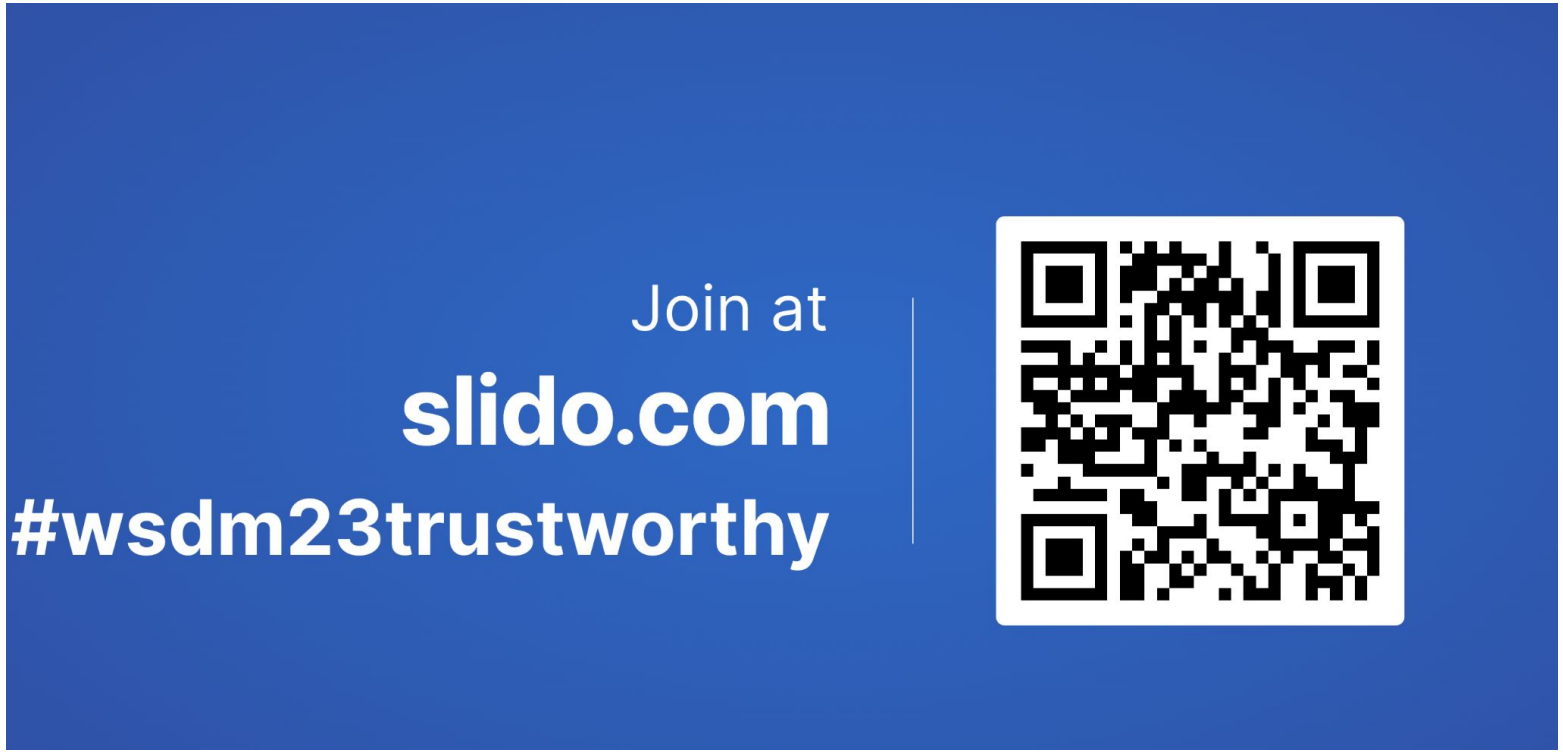
About Elisabeth Lex

- Assoc. Prof at Graz University of Technology, Austria
- PI Recommender Systems & Social Computing Lab at Institute of Interactive Systems and Data Science (ISDS)
- Lab page: https://socialcomplab.github.io/
- Interests: user modeling, recommender systems, information retrieval, natural language processing, computational social science

Contact: elisabeth.lex@tugraz.at | https://elisabethlex.info | @elisab79

What about you?

- Share your views on our tutorial on Twitter: **#wsdm23trustworthy**
- Interact, ask/vote questions, on site & online



Overview

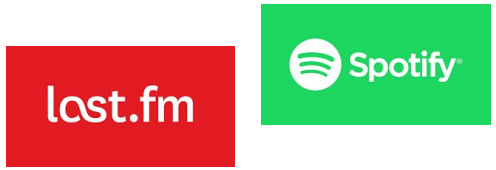
1. **Introduction** - 15’
   Background, motivation, objectives, relevance to community
2. **Trustworthy AI** - 45’
   Introduction, requirements, from ethics guidelines to regulation
3. **Fairness and non-discrimination** - 45’ + + break (30)’
   Categories of bias and fairness, relation to non-discrimination, definition and measurement of bias and fairness, algorithms to mitigate biases and improve fairness
4. **Transparency** - 45’
   Categories of transparency, explainability and justification, traceability and auditability, documentation
5. **Open Challenges** - 15’

Tutorial Slides: https://socialcomplab.github.io/Trustworthy-ARS-Tutorial-WSDM22

Part 1: Introduction

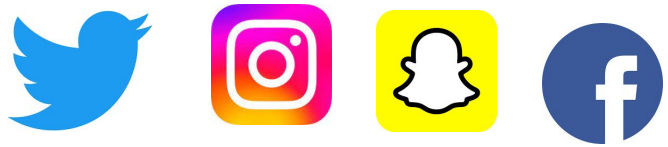
Information Retrieval (IR) and Recommender Systems (RS) are Ubiquitous Algorithmic Ranking Systems

e-commerce/products



music

jobs



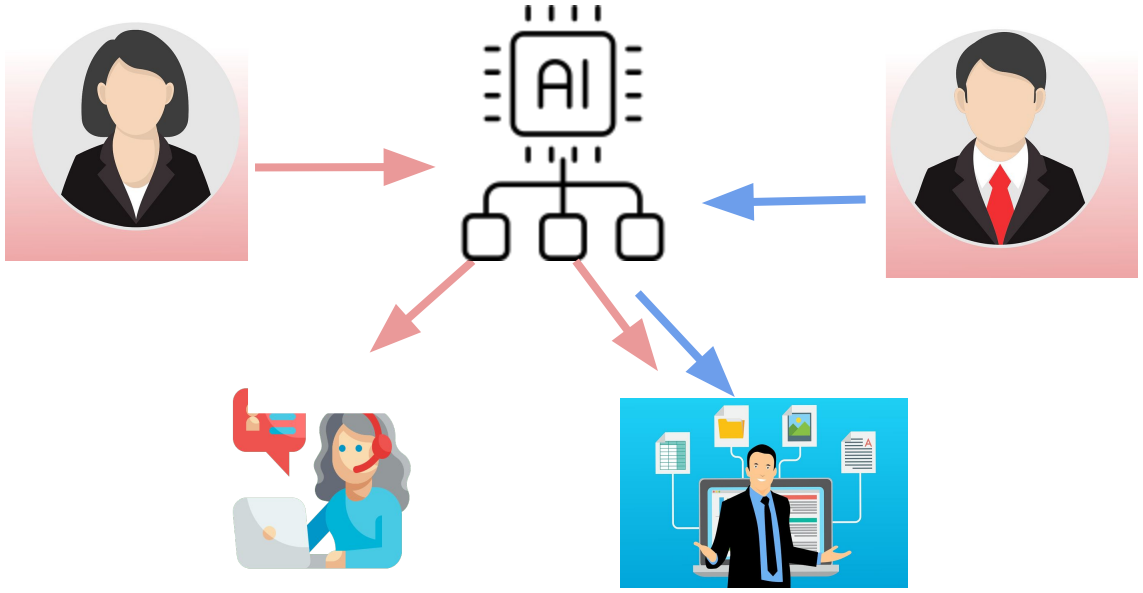
social networking
/ information

movies / tv

travel

Societal Impacts of IR & RS

- From decision support / information seeking tools → socio-technical systems
- Create, control, limit exposure & access, shape opinion, influence behaviour:
  - e.g., jobs, products, information, opportunities



Raises Ethical Questions

Not Only a Technological or Algorithmic Problem

- **Multidisciplinary perspective**: law, ethics, sociology, economics, psychology, etc.
- EU Charter of Fundamental Rights (https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en)
- RS & IR/search engines as part of Artificial Intelligence:
  - EU Ethical Principles for Trustworthy AI (https://op.europa.eu/s/pXjd)
  - EU Regulatory Framework proposal on AI - 2021 (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai)
  - EU Digital Services Act (more details later) https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en

Chinese AI Governance Approaches

THREE APPROACHES TO CHINESE AI GOVERNANCE		
Organization	**Focus of Approach**	**Relevant Documents**
Cyberspace Administration of China	Rules for online algorithms, with a focus on public opinion	- Internet Information Service Algorithmic Recommendation Management Provisions - Guiding Opinions on Strengthening Overall Governance of Internet Information Service Algorithms
China Academy of Information and Communications Technology	Tools for testing and certification of "trustworthy AI" systems	- Trustworthy AI white paper - Trustworthy Facial Recognition Applications and Protections Plan
Ministry of Science and Technology	Establishing AI ethics principles and creating tech ethics review boards within companies and research institutions	- Guiding Opinions on Strengthening Ethical Governance of Science and Technology - Ethical Norms for New Generation Artificial Intelligence

https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127

US Initiatives

- The Artificial Intelligence Initiative Act (116th Congress 2019-2020, S.1558): https://www.congress.gov/bill/116th-congress/senate-bill/1558/text

- White House's Office of Science and Technology Policy released a draft *Guidance for Regulation of Artificial Intelligence Applications*: https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

- Regulations in different states, e.g. California on Automated Decision Systems for Employment and Housing.

  https://www.dfeh.ca.gov/wp-content/uploads/sites/32/2022/03/AttachB-ModtoEmployRegAutomated-DecisionSystems.pdf

Part 2:
Trustworthy AI

Trustworthy /ˈtrʌstˌwə-ði/

: able to be relied on to do or provide what is needed or right; deserving of trust; worthy of confidence.

Paradigm change

System-oriented

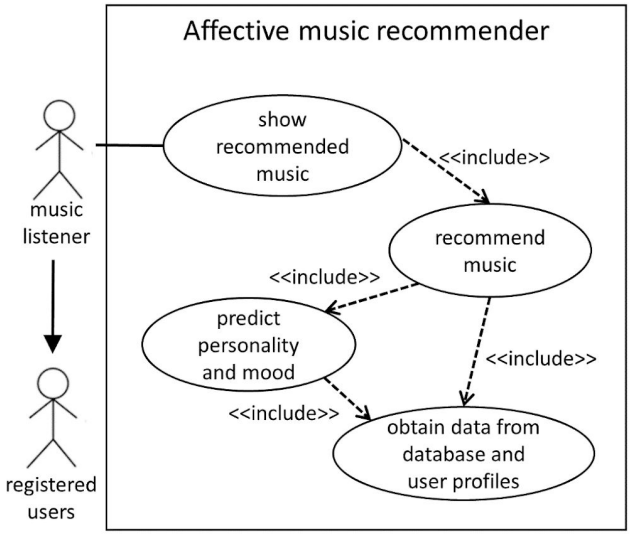
Extraction and inference of meaningful information from large collections.

Socio-technical system

Interaction with the social system, e.g. business processes, organizations, society (law, culture).

Markus Schedl, Emilia Gómez and Julián Urbano (2014), "Music Information Retrieval: Recent Developments and Applications", Foundations and Trends in Information Retrieval: Vol. 8: No. 2-3, pp 127-261. http://dx.doi.org/10.1561/1500000042

From systems to use cases: example



Hupont, I., & Gomez, E. (2022). Documenting use cases in the affective computing domain using Unified Modeling Language. Affective Computing and Intelligence Interaction https://arxiv.org/abs/2209.09666v1

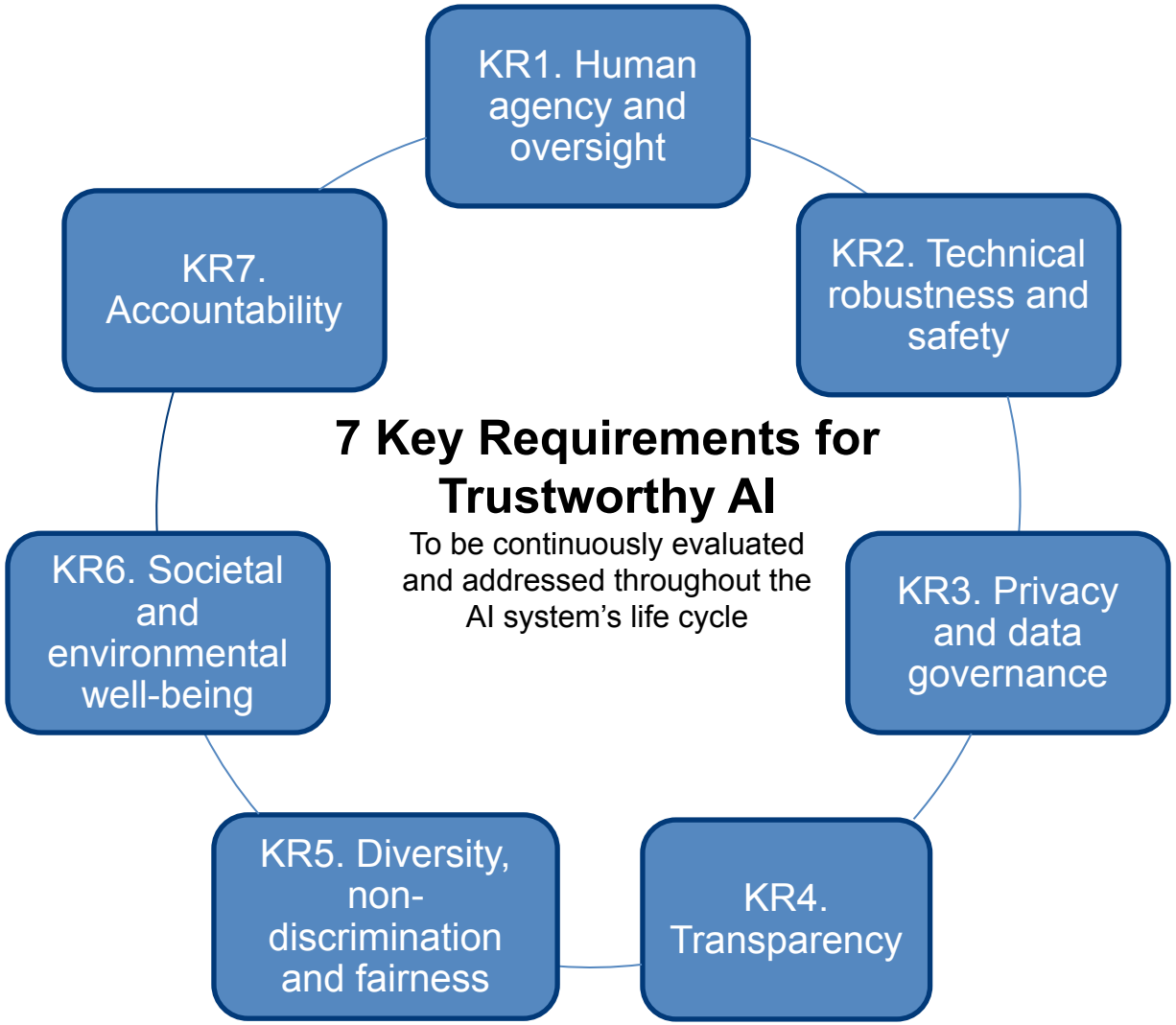
USE CASE	**Show recommended music**	
Context of use	The user is subscribed to a music platform, which recommends the most appropriate and enjoyable tracks according to her personality and current mood. Personality and mood are predicted based on the data in the user's profile (voluntary provided by the user) and the historical music data she has listened to. The system also takes into account the music tracks played by other users with a similar profile to make recommendations. The user accesses the music platform through an application installed in her mobile phone.	
Intended purpose	Recommend a list of songs to the user according to her personality, current mood and music preferences.	
Application areas	Entertainment and leisure	
User	Music listener	
Target persons	**Person**	**Description**
	Registered users	Other users registered on the platform and whose profile and music preferences are used to make recommendations.
Success end condition	A list of 20 recommended music tracks is shown to the user in the application's graphic interface.	
Failure protection	A default personality- and mood-neutral list of 20 songs is shown to the user in the application's graphic interface.	
Trigger	The user presses the "recommend music" button in the application.	
Main course	**Step**	**Action**
	1	The application calls the recommender algorithm.
	2	The current mood of the user is predicted based on her profile information and recently played songs.
	3	The personality of the user is predicted based on her profile information and historical music playlists.
	4	The recommender ranks songs according to predicted mood, personality and music playlists of other registered users with similar profile.
	5	The application displays the 20 top-ranked recommended tracks for the user.
Extensions	**Step**	**Branching action**
	2a	If no song has been played yet, the system assigns the user a neutral mood.
	3a	If there is no historical music data, personality prediction is based on the user's profile information exclusively.
Misuses	The recommender shall not propose pieces of music pre-conceived to exploit vulnerabilities, manipulate, distort or induce certain emotions or behaviour in users, e.g. for marketing purposes.	

Ethics guidelines for Trustworthy AI (2019)

1. **Lawful** - respecting all applicable laws and regulations.

2. **Ethical** - respecting ethical principles and values.

3. **Robust** - both from a technical perspective while taking into account its social environment. Avoid intentional/unintentional harm.


KR1. Human agency and oversight
KR2. Technical robustness and safety
KR3. Privacy and data governance
KR4. Transparency
KR5. Diversity, non-discrimination and fairness
KR6. Societal and environmental well-being
KR7. Accountability
7 Key Requirements for Trustworthy AI
To be continuously evaluated and addressed throughout the AI system's life cycle

KR1. Human agency and oversight

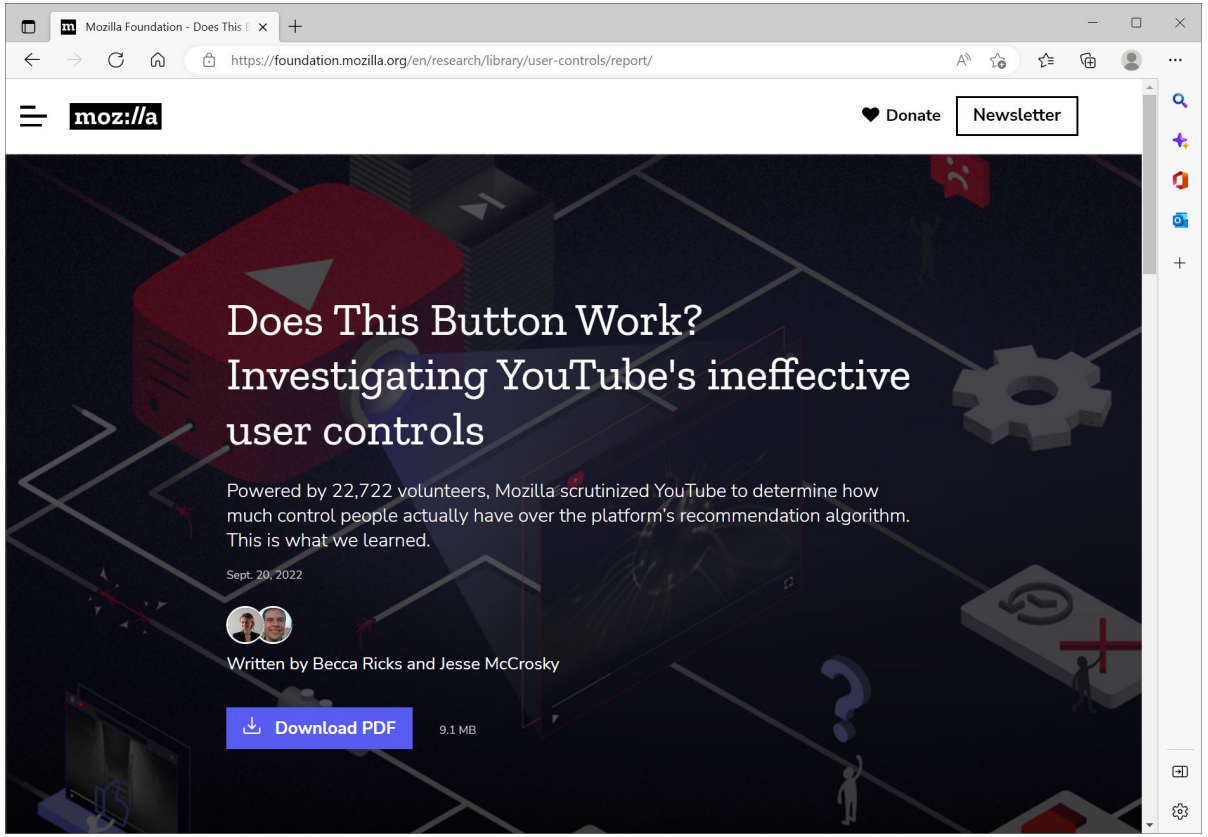
Description

*AI systems should **empower** human beings, allowing them to make informed decisions and fostering their fundamental rights.*

*At the same time, proper **oversight mechanisms** need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches.*

Related topics

- *Relevant user interfaces and HCIs.*
- *Ways to override or reverse the system output.*
- *Expertise needed to operate a system, how to evaluate its correct operation.*



KR1. Which are relevant human oversight mechanisms for a web search engine?

slido

KR1. Which are relevant human oversight mechanisms for a web search engine?KR1. Which are relevant human oversight mechanisms for a web search engine?KR1. Which are relevant human oversight mechanisms for a web search engine?KR1. Which are relevant human oversight mechanisms for a web search engine?

ⓘ Start presenting to display the poll results on this slide.

KR2. Technical robustness and safety

Description

*AI systems need to be **resilient** and **secure**. They need to be **safe**, ensuring a fall back plan in case something goes wrong, as well as being **accurate**, **reliable** and **reproducible**. That is the only way to ensure that also unintentional harm can be minimized and prevented.*

Related topics

- Accuracy level, metrics,
- Robustness tests: unintended (e.g. noise) or intended errors (e.g. attacks).
- Consideration of edge cases.
- Reproducibility, open science.

Considerations in information retrieval and search engines:

- Evaluation metrics vary.
- Robustness tests not widely applied.
- Reproducibility fostered in IR research, but limited in real-world scenarios.

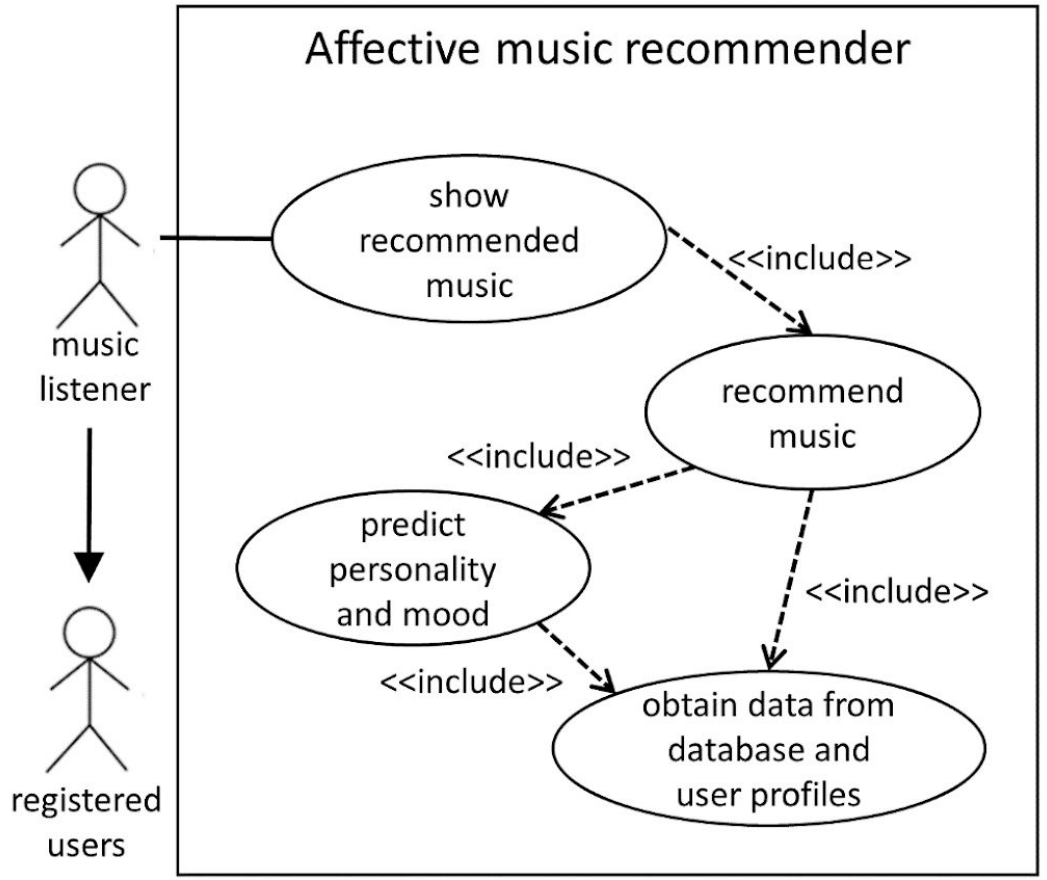
KR3. Privacy and data governance

Description

Besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.

Related topics

- *Protection of personal data and data minimization, anonymization.*
- *Privacy-preserving algorithms.*
- *Define how to collect, label, process, audit and monitor the data that goes into developing models.*
- *Define which data is really needed.*
- *Datasets used for training and validation should be relevant, without errors and representative: ensure data is up to date, matches demographics, carry out sanity checks.*



KR3. Privacy and data governance

- Different legal approaches (e.g. EU - GDPR vs US).
- Different levels of personal data:
  - Nominative data leading to the identification of natural persons.
  - Data leading to identification through “complex” processes.

Challenges:

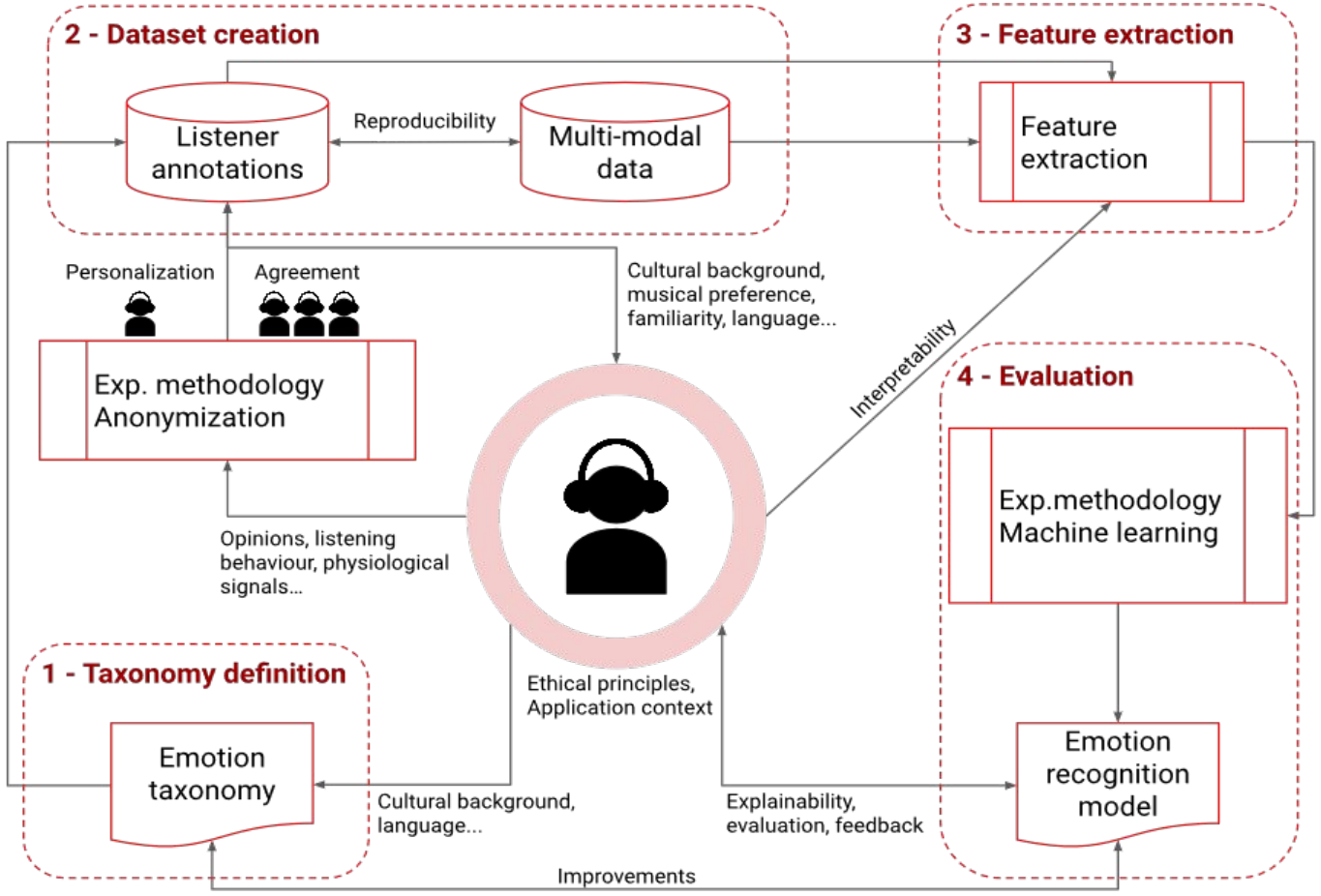
- Privacy vs personalization.
- Implicit and explicit user data.
- Consent: understandability and control.
- Time dimension.
- Minimization of personal data.

Pierre Saurel, Francis Rousseaux, & Marc Danger. (2014). On The Changing Regulations of Privacy and Personal Information in MIR. Proceedings of the 15th International Society for Music Information Retrieval Conference, 597–602. https://doi.org/10.5281/zenodo.1416638

KR3. Privacy and personalization

J. S. Gómez-Cañón et al., "Music Emotion Recognition: Toward new, robust standards in personalized and context-sensitive applications," in IEEE Signal Processing Magazine, vol. 38, no. 6, pp. 106-114, Nov. 2021, doi: 10.1109/MSP.2021.3106232.

KR3. Do you handle any personal data in your research?



slido

KR3. Do you handle any personal data in your research?

ⓘ Start presenting to display the poll results on this slide.

KR4. Transparency

Description

- *The **data, system and business models** **linked to AI** should be transparent. Traceability mechanisms can help achieving this.*
- *AI systems and their decisions should be **explained** in a manner adapted to the stakeholder concerned.*
- *Humans need to be aware that they are interacting with an AI system, and must be informed of the system's **capabilities and limitations**.*

Related topics

How the system is built and evaluated, training data, limitations.
How the system is monitored, e.g. logs.
How the system is controlled, e.g. how to interpret its outputs. Potential misuse.
Pre-determined changes, expected lifetime and
necessary maintenance/care measures.
Communication to users, e.g. generative models.

https://algorithmic-transparency.ec.europa.eu/index_en

KR4. Transparency: generation

- Transparency can serve to empower people to challenge AI systems.
- Users → fake news, impersonation.
- Artists & creators → intellectual property, e.g. training data has copyright, infringement on the engineer.

Gómez, E., Blaauw, M., Bonada, J., Chandna, P., & Cuesta, H. (2018). Deep learning for singing processing: Achievements, challenges and impact on singers and listeners. arXiv preprint arXiv:1807.03046.
Sturm B., Iglesias M, Ben-Tal O, Miron M, Gómez E. Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis. Arts. 2019; 8(3):115. https://doi.org/10.3390/arts8030115

MOTHERBOARD

‘Deep Voice’ Software Can Clone Anyone's Voice With Just 3.7 Seconds of Audio

Using snippets of voices, Baidu's ‘Deep Voice’ can generate new speech, accents, and tones.

SHARE TWEET

KR4. Which information should the user know about a web search engine or list of retrieved items in order to challenge it?

slido

KR4. Which information should the user have about a search/IR engine in order to challenge it?KR4. Which information should the user have about a search/IR engine in order to challenge it?

ⓘ Start presenting to display the poll results on this slide.

KR5. Diversity, non-discrimination and fairness

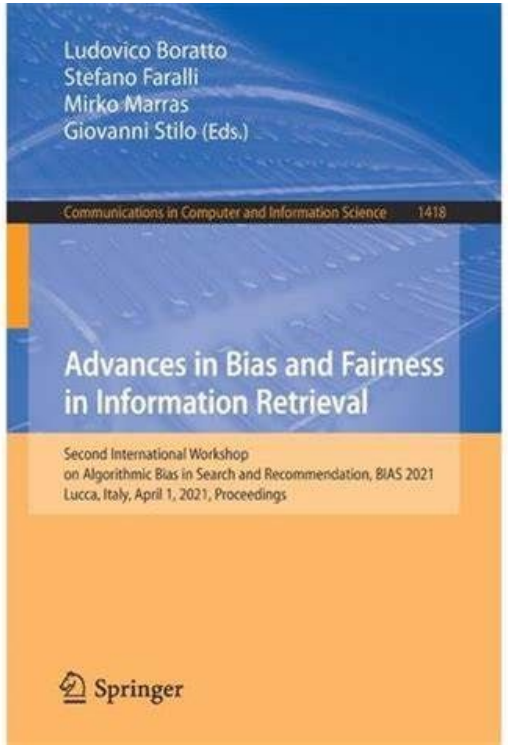
Description

Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life cycle.

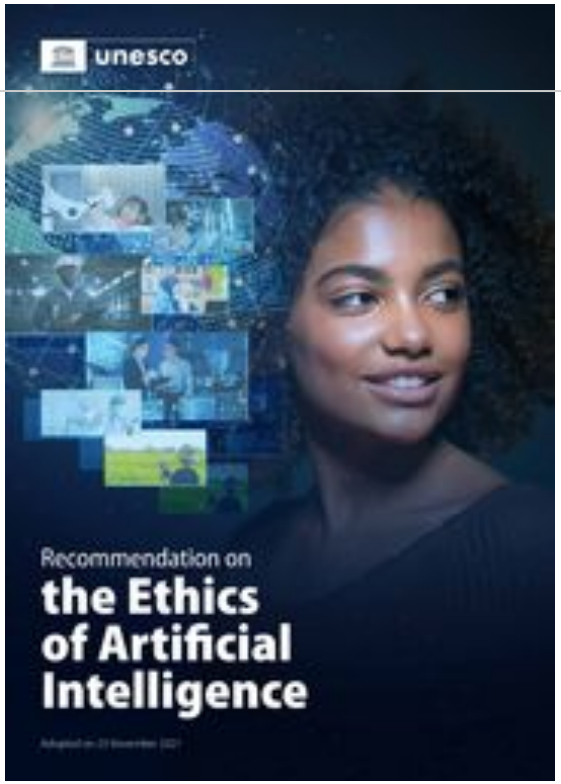
Related topics

Demographics, gender, sex, age, ethnicity, culture, religion, sexual orientation, political orientation, culture….

Strong research on bias and fairness, Part 2.



KR5. Diversity: motivation



== Innovation, creativity; Lack of diversity == bias

Ensuring diversity and inclusion (UNESCO)

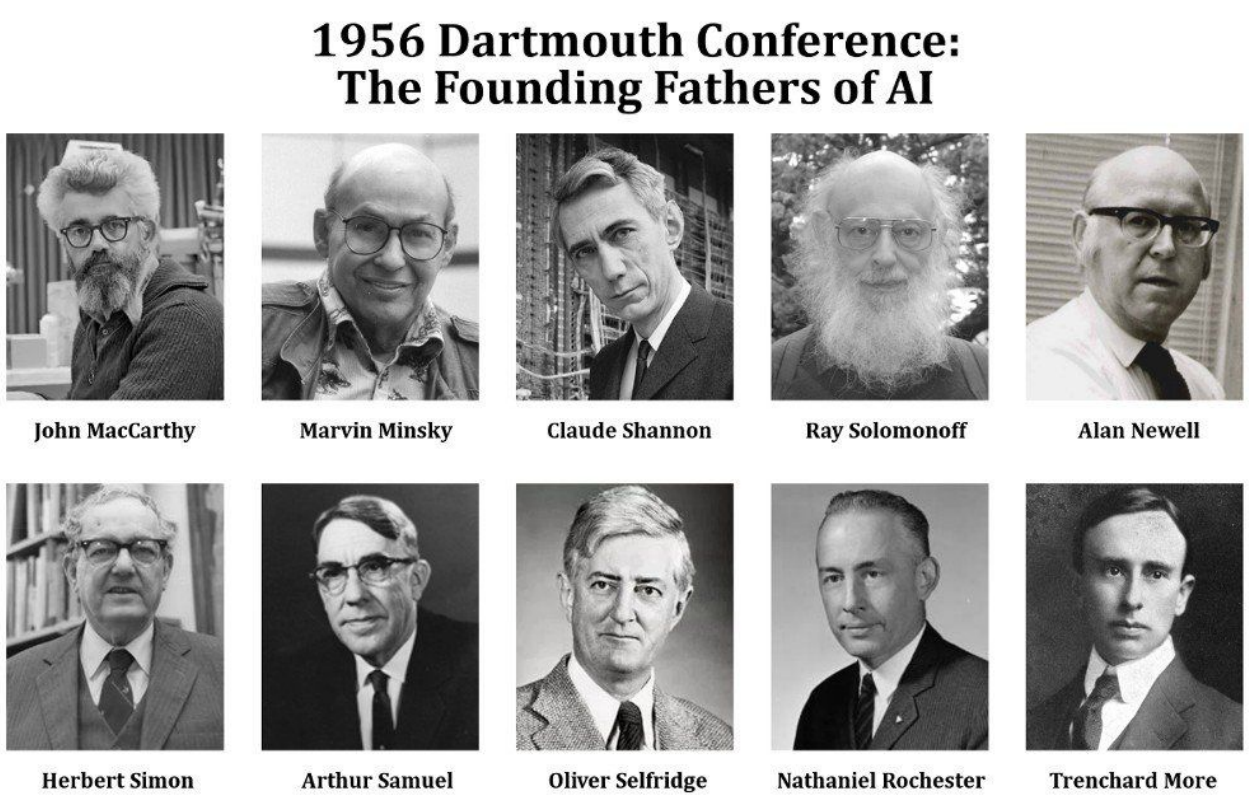
- Respect, protection and promotion of diversity.
- Consider personal choices, including the optional use of AI systems and its co-design.
- Overcome lack of necessary technological infrastructure, education and skills, as well as legal frameworks.
- Dimensions: gender, age, cultural origin, language, political opinion, …

Relevant topics

- Demographic diversity
- Diversity by design

Recommendation on the Ethics of Artificial Intelligence, UNESCO https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
Porcaro, L., Castillo, C. and Gómez, E., 2021. Diversity by Design in Music Recommender Systems. Transactions of the International Society for Music Information Retrieval, 4(1), pp.114–126. DOI: http://doi.org/10.5334/tismir.106

KR5. Demographic diversity



https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

4 **AI has a diversity challenge:** In 2019, 45% new U.S. resident AI PhD graduates were white—by comparison, 2.4% were African American and 3.2% were Hispanic.

Zhang, D., et al. The AI Index 2021 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.

How diverse is our group now?

slido

KR5. What best describes your affiliation?

ⓘ Start presenting to display the poll results on this slide.

slido

KR5. Country of affiliation

ⓘ Start presenting to display the poll results on this slide.

slido

KR5. Country of origin

ⓘ Start presenting to display the poll results on this slide.

slido

KR5. Gender

ⓘ Start presenting to display the poll results on this slide.

slido

KR5. Age

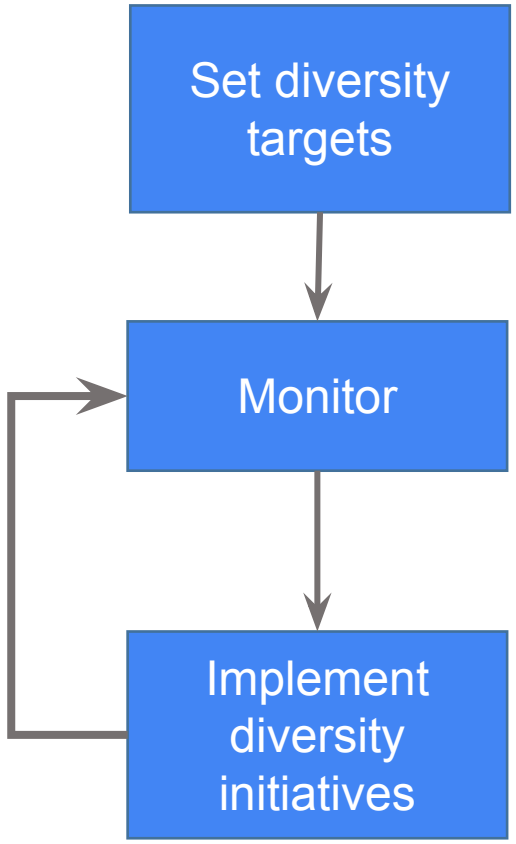
ⓘ Start presenting to display the poll results on this slide.

slido

Which keywords describe your research topic?

ⓘ Start presenting to display the poll results on this slide.

KR5. Demographic diversity

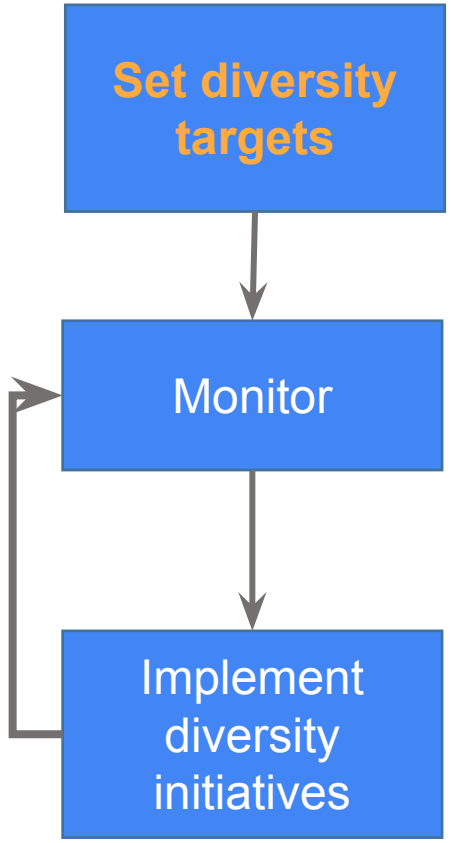


1

I. Hupont, S. Tolan, P. Frau, L. Porcaro and E. Gómez, "Measuring and fostering diversity in Affective Computing research," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2023.3244041 .

KR5. Demographic diversity

- Dimensions:
  - Gender, sexual orientation
  - Age, seniority
  - Racial, ethnicity / geographical origin or location
  - Institution type: academia, industry, government,…
  - Disabilities
  - Topics: disciplines, methodologies, aspects
- Targets: increase diversity, a collective decision?



I. Hupont, S. Tolan, P. Frau, L. Porcaro and E. Gómez, "Measuring and fostering diversity in Affective Computing research," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2023.3244041 .

KR5. Demographic diversity

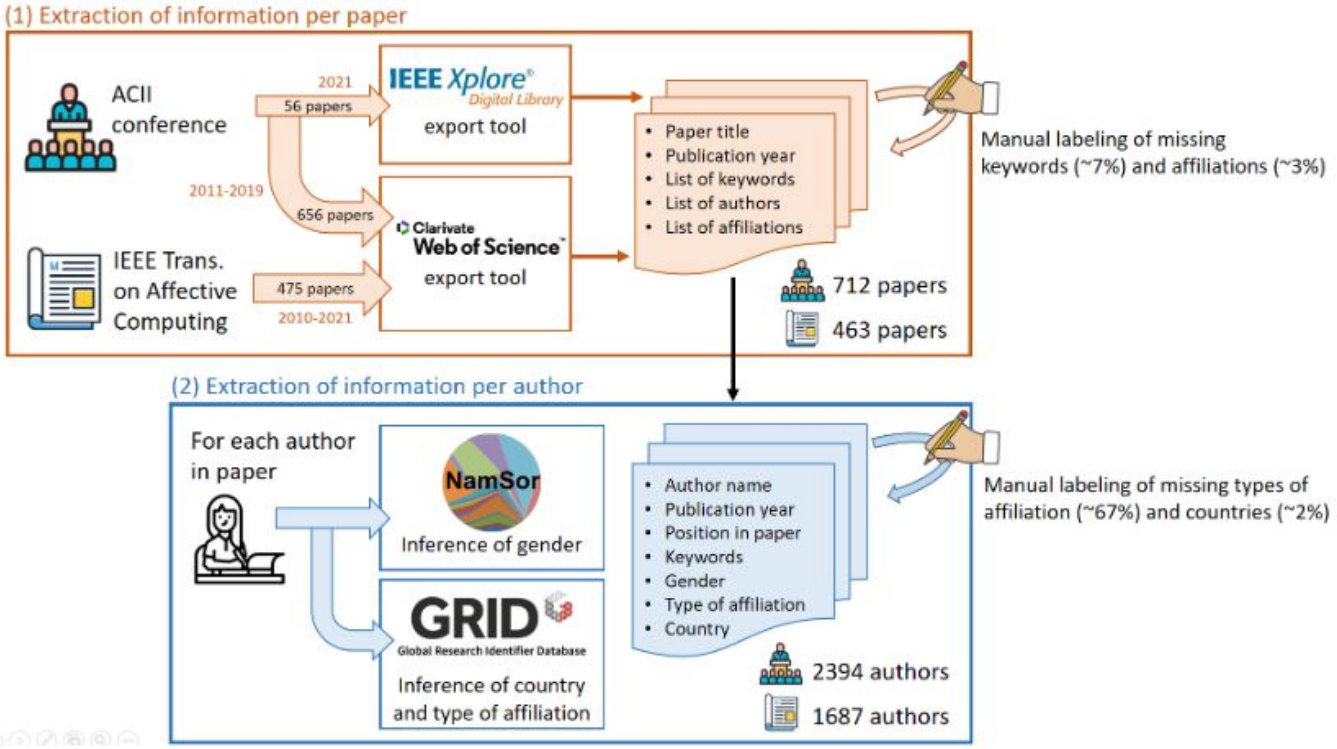
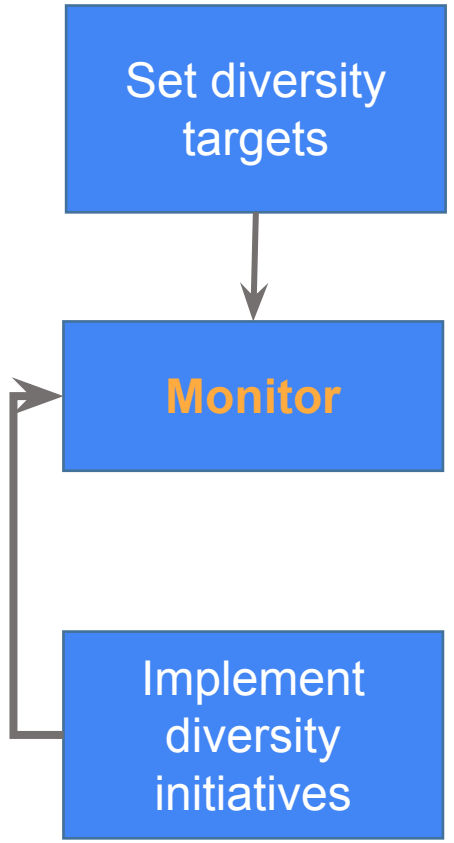


Fig. 1: Semi-automated process followed to collect per-paper and per-author data from ACII conferences and the IEEE Transactions on Affective Computing journal (years 2011 to 2021).

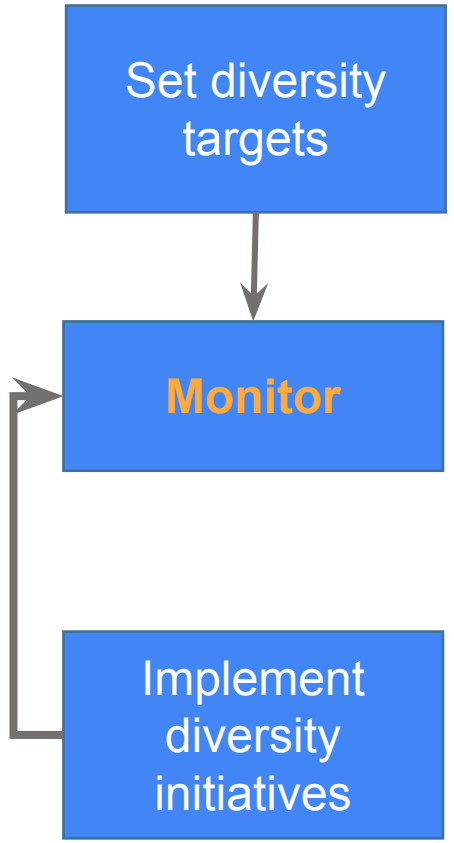


I. Hupont, S. Tolan, P. Frau, L. Porcaro and E. Gómez, "Measuring and fostering diversity in Affective Computing research," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2023.3244041 .

KR5. Demographic diversity

MAIN FOCUS	YEARS	METRICS	DIVERSITY DIMENSIONS							
			Gender	Sex. orient.	Ethnicity	Age	Countries	Institutions	Topics	Cross
BiasWatchNeuro [13] Neuroscience: keynote speakers in >50 conferences and 4 journals	2015-2021	Women rate with respect to "base" rate	×							
Neuroscience: speakers in 18 conferences [14]	2019-2020	Percentages			×					
Geoscience: 9 societies, 25 journals and 10 conferences (organisation committee members) [15]	2016	Percentages	×			×		×		
Geoscience: speakers at 1 conference [16]	2017	Percentages	×		×				×	×
STEM: 1 society and 1 conference (speakers, attendees and poster presenters) [18]	2011-2015	Percentages	×						×	×
Medicine: speakers at 1 conference [20]	2016-2018	Percentages, speaking time	×							
AI Index Report [22] AI: survey data obtained from under-represented group members (women, queer, black) and participants in 1 workshop	2015-2020	Percentages	×	×	×		×	×		
AI Watch Index [23] AI: authors, keynote speakers and PC members in 5 top-tier conferences	2016-2020	4 diversity indexes	×				×	×		×
Affective Computing: authors, keynote speakers and PC members in ACII conference [11]	2005-2019	4 diversity indexes, percentages	×				×	×		×
This work Affective Computing: ACII conference (authors, keynote speakers and PC members), TAFFC journal and AAAC association	**2011-2021**	**8 diversity indexes, percentages, clustering**	×				×	×	×	×

TABLE 1: Main focus, years, metrics and dimensions analysed in state-of-the-art diversity studies. The last row corresponds to the current work.

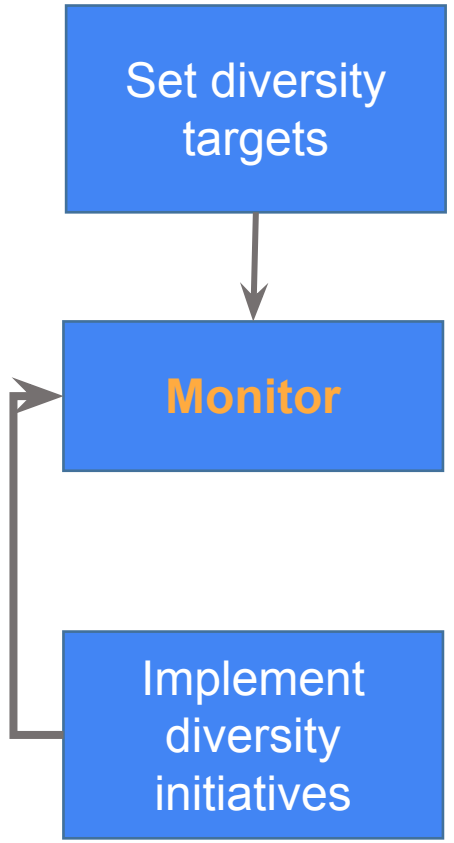


I. Hupont, S. Tolan, P. Frau, L. Porcaro and E. Gómez, "Measuring and fostering diversity in Affective Computing research," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2023.3244041 .

KR5. Demographic diversity

Indicators:

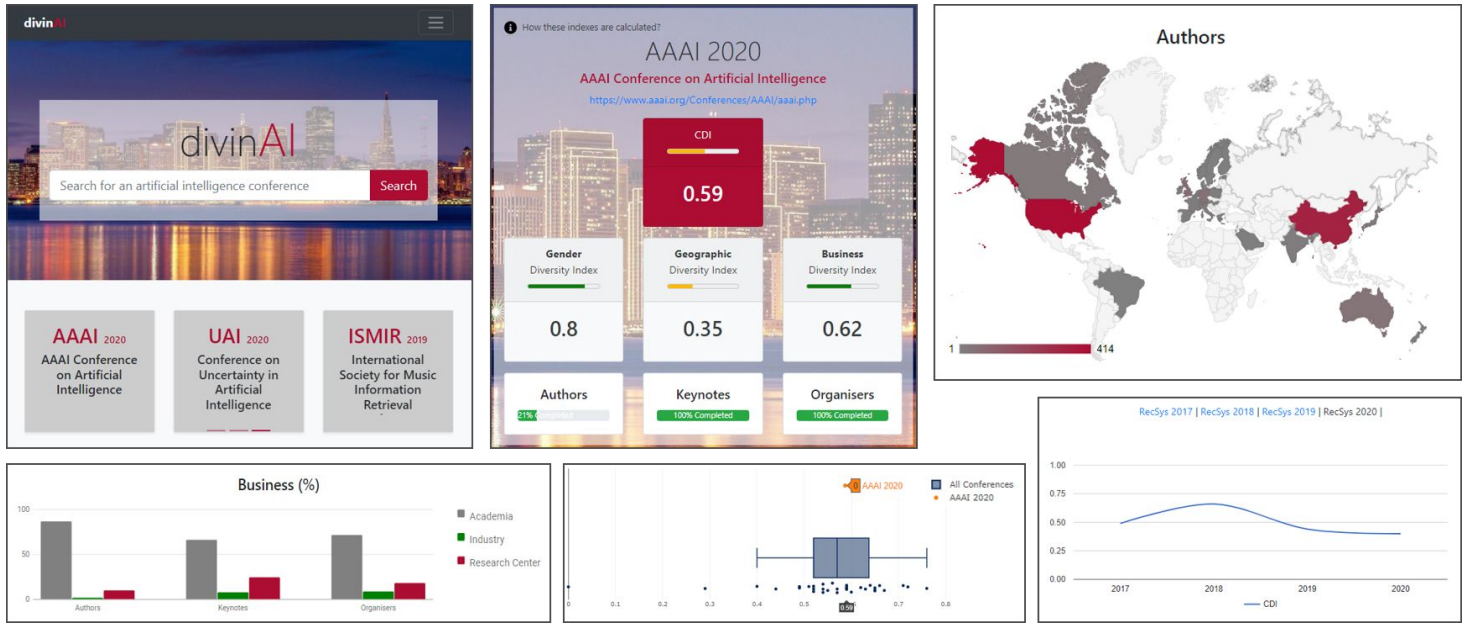
- Dual-concept diversity (McDonald and Dimmick, 2003)
  a. Variety: number of categories in a population.
  b. Balance: evenness of distribution across categories.
- Examples: *Shannon, Pielou, Simpson, Herfindahl-Hirschman*.
- 3rd dimension to account for similarity among categories - disparity: Rao-Stirling index (Stirling, 2007)
- Weighting of different dimensions.

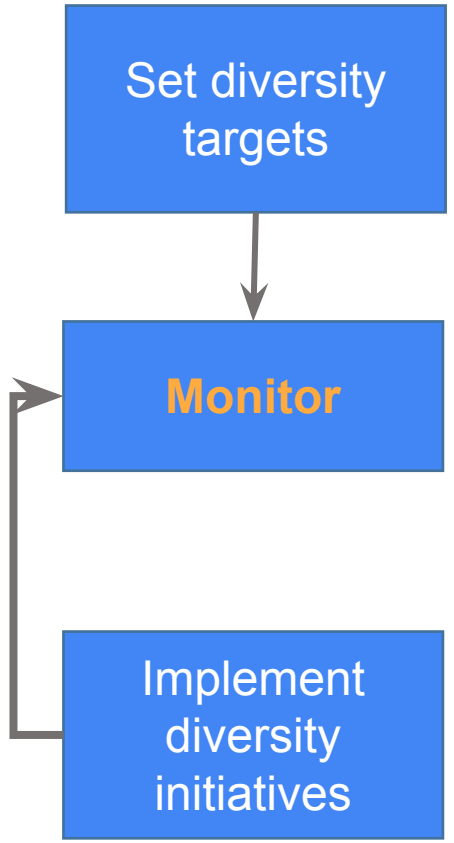


I. Hupont, S. Tolan, P. Frau, L. Porcaro and E. Gómez, "Measuring and fostering diversity in Affective Computing research," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2023.3244041 .

KR5. Demographic diversity

- Lack of curated data (country, gender, institution type, topics)
- Ethical concerns: privacy, labeling.



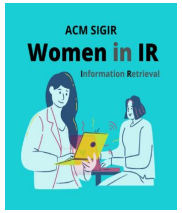


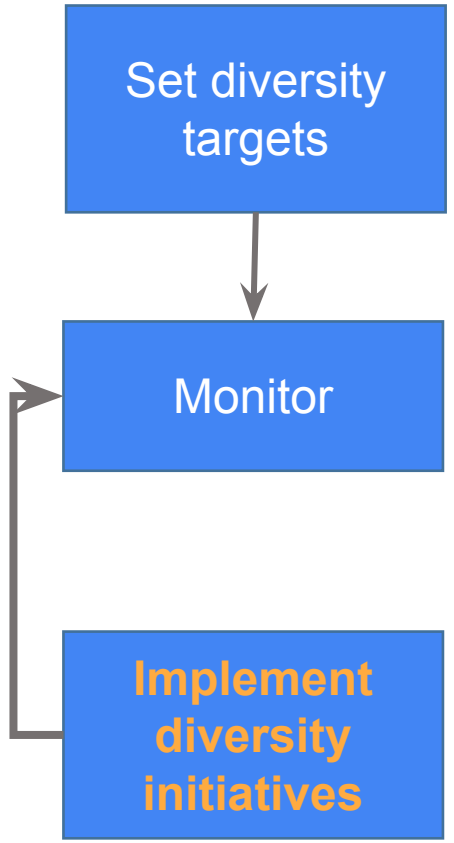
https://divinai.org/

I. Hupont, S. Tolan, P. Frau, L. Porcaro and E. Gómez, "Measuring and fostering diversity in Affective Computing research," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2023.3244041 .

KR5. Demographic diversity

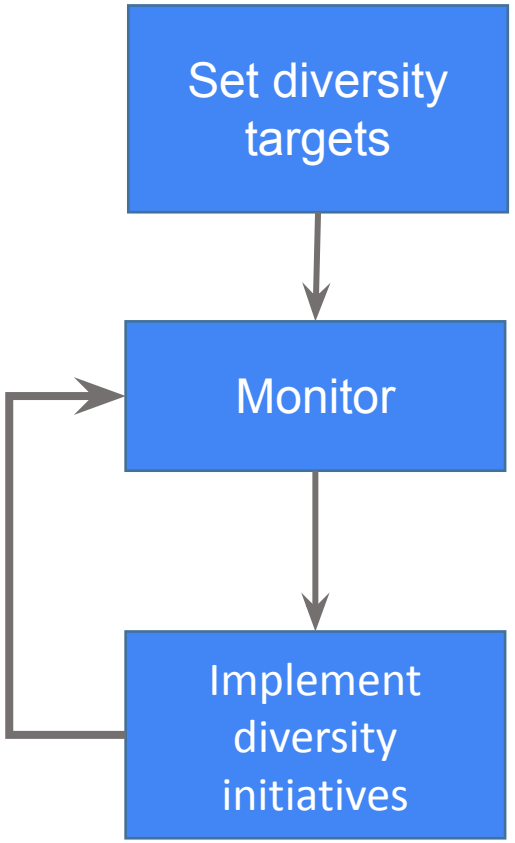
AFFINITY GROUP	SINCE	FOCUS
Women in ML (WiML)[19]	2007	Enhance the experience of women in ML, in order to help them succeed professionally and increase their impact in the community.
Women in MIR (WiMIR)[20]	2012	Promote the role of, and increase opportunities for, women, trans or non-binary at any career stage in the field of music information retrieval.
Women in RecSys[21]	2014	Foster diversity and celebrate female role models in the recommender systems research community.
Women in CV (WiCV)[22]	2015	Foster the carreer and mitigate the isolation of female researchers working on computer vision.
Black in AI[23]	2017	Increasing the presence and inclusion of Black people in the field of AI.
Widening NLP (WiNLP)[24]	2017	Help to promote and support ideas and voices of under-represented groups in the Natural Language Processing community.
LatinX in AI[25]	2018	Latin professionals working on AI, ML and Data Science.
Queer in AI[26]	2019	People with diverse non-normative sexual orientations, romantic orientations and/or genders, corresponding to acronyms like LGBTQ+.
{Dis}Ability in AI[27]	2019	All those who experience barriers in accessing education due to having or being considered to have an impairment (e.g. physical or sensory impairments, people with learning difficulties, people with mental health or autism spectrum conditions).
Indigenous AI[28]	2019	Design and create AI from an ethical position that centers Indigenous concerns. The Indigenous term covers diverse communities in Aotearoa, Australia, North America and the Pacific.
African Women in AI (AWAI)[29]	2022	Promote knowledge sharing within the African women AI and ML community.





I. Hupont, S. Tolan, P. Frau, L. Porcaro and E. Gómez, "Measuring and fostering diversity in Affective Computing research," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2023.3244041 .

KR5. Demographic diversity



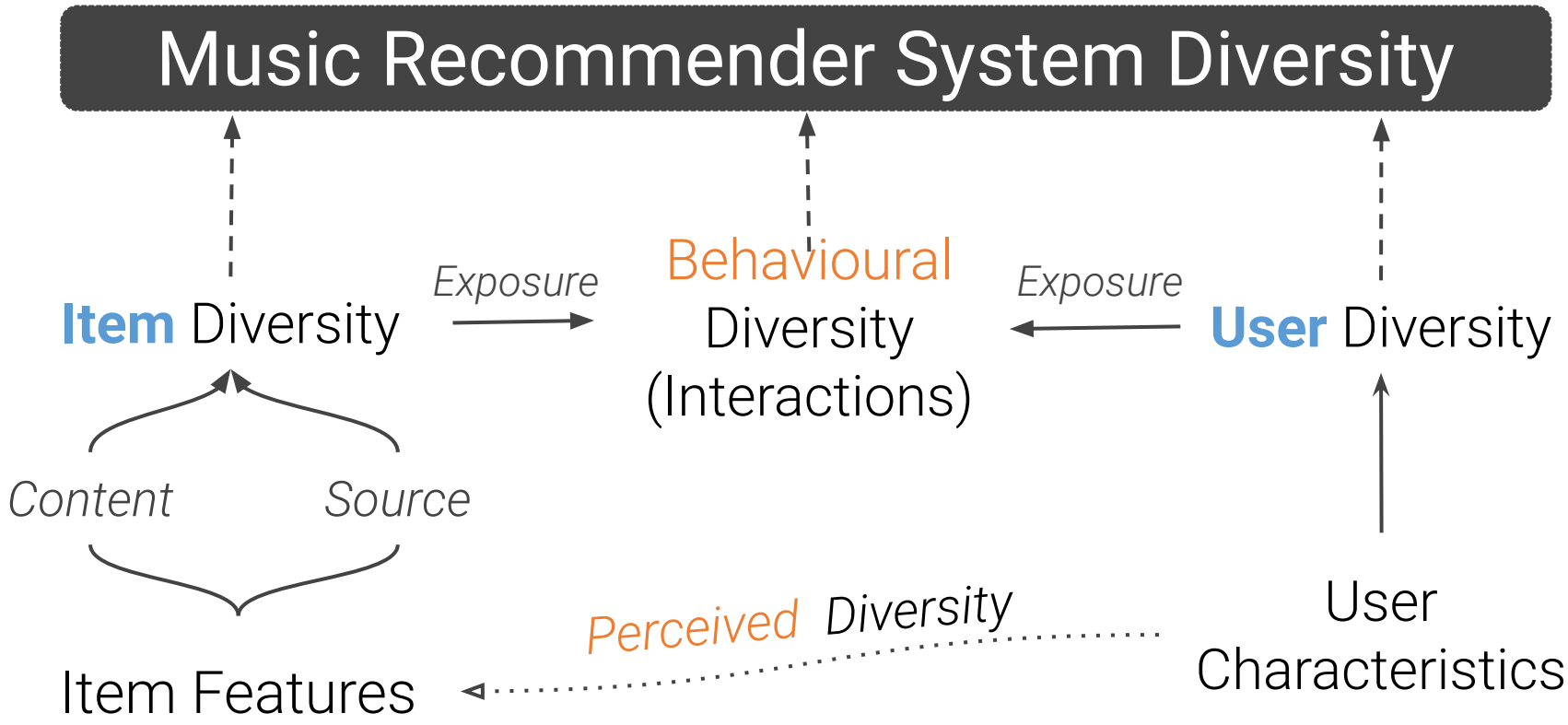
Dimensions
Desired targets

Lack of tools/metrics
Needed for impact assessment

Context-dependent
Linked to targets

I. Hupont, S. Tolan, P. Frau, L. Porcaro and E. Gómez, "Measuring and fostering diversity in Affective Computing research," in *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2023.3244041 .

KR5. Diversity by design



Lorenzo Porcaro, Assessing the Impact of Music Recommendation Diversity on Listeners, PhD thesis, Universitat Pompeu Fabra, 2022.

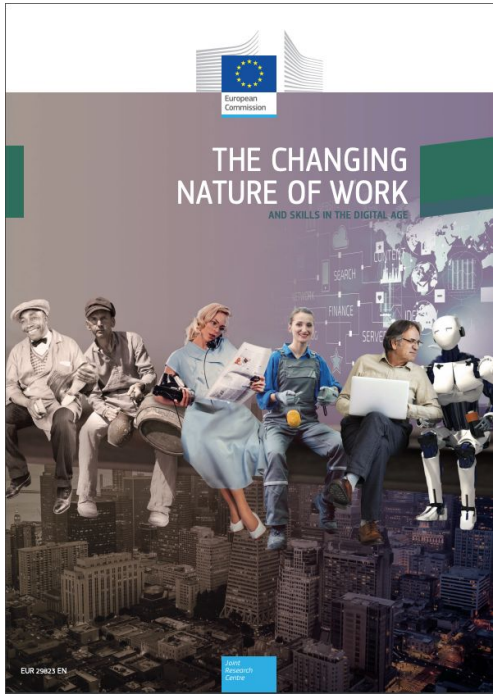
KR6. Societal and environmental well-being

Description

*AI systems should **benefit** all human beings, including future generations. It must hence be ensured that they are sustainable and **environmentally** friendly. Moreover, they should take into account the environment, including other living beings, and their social and **societal impact** should be carefully considered.*

Related topics

- Impact on jobs and skills.
- Computing ressources.



KR6. How will IR/search engines affect workspaces?

Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J., & Gómez, E. (2021). Measuring the occupational impact of AI: tasks, cognitive abilities and AI benchmarks. *Journal of Artificial Intelligence Research*, *71*, 191-236.
Emilia Gómez (2020). Human and Machine Intelligence: A Music Information Retrieval Perspective. Keynote speech, 11th International Conference on Computational Creativity.

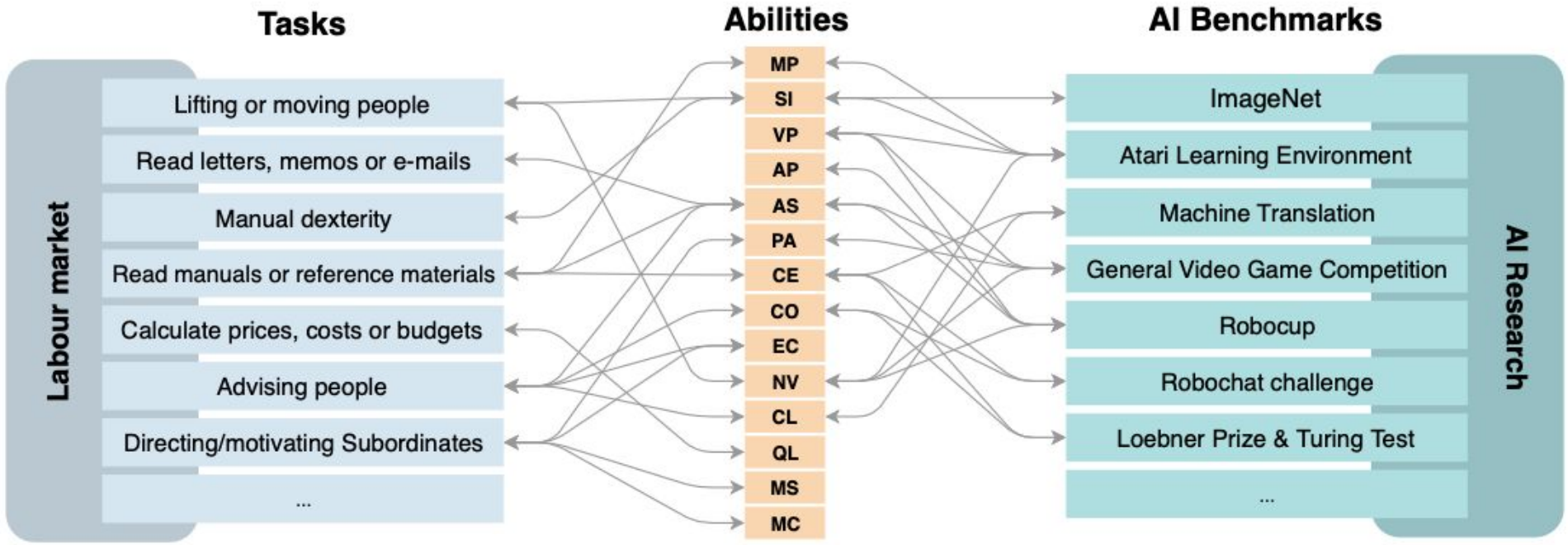
slido

KR6. How is web search and data mining affecting workspaces? KR6. How is web search and data mining affecting workspaces?

ⓘ Start presenting to display the poll results on this slide.

KR6. Societal well-being: jobs

Songul Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernandez-Macias, José Hernández-Orallo & Emilia Gómez, "Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks", JRC Working Papers on Labour, Education and Technology 2020-02, Joint Research Centre (Seville site).

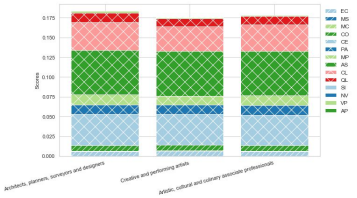


- Creative and performing artists
- Architects, planners, surveyors, designers
- Artistic, cultural and culinary associate professionals

- Creativity and resolution
- Accounting
- Business

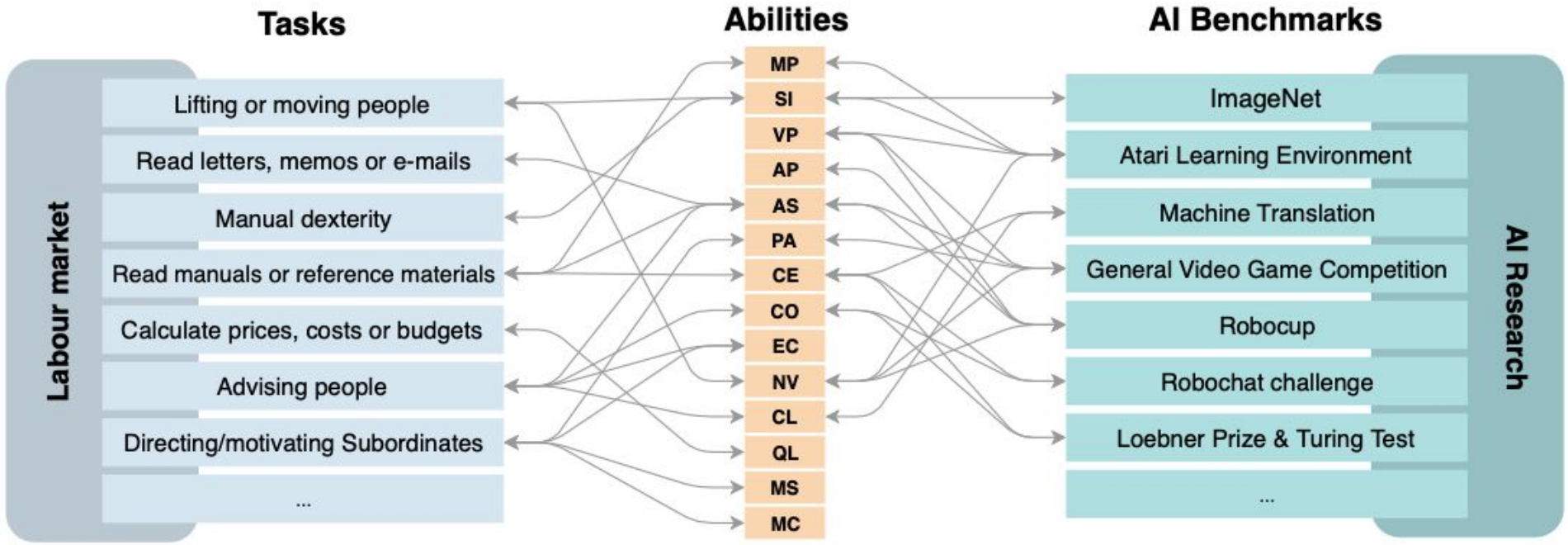
- Communication
- Conceptualization
- Text comprehension
- Attention and search
- Quantitative reasoning

- AI exposure score
- Few benchmaking initiatives on creative systems



KR6. Societal well-being: jobs

Songul Tolan, Annarosa Pesole, Fernando Martínez-Plumed, Enrique Fernandez-Macias, José Hernández-Orallo & Emilia Gómez, "Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks", JRC Working Papers on Labour, Education and Technology 2020-02, Joint Research Centre (Seville site).

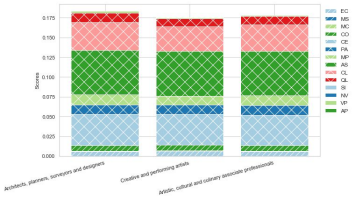


- Creative and performing artists
- Architects, planners, surveyors, designers
- Artistic, cultural and culinary associate professionals

- Creativity and resolution
- Accounting
- Business

- Communication
- Conceptualization
- Text comprehension
- Attention and search
- Quantitative reasoning

- AI exposure score
- Few benchmaking initiatives on creative systems

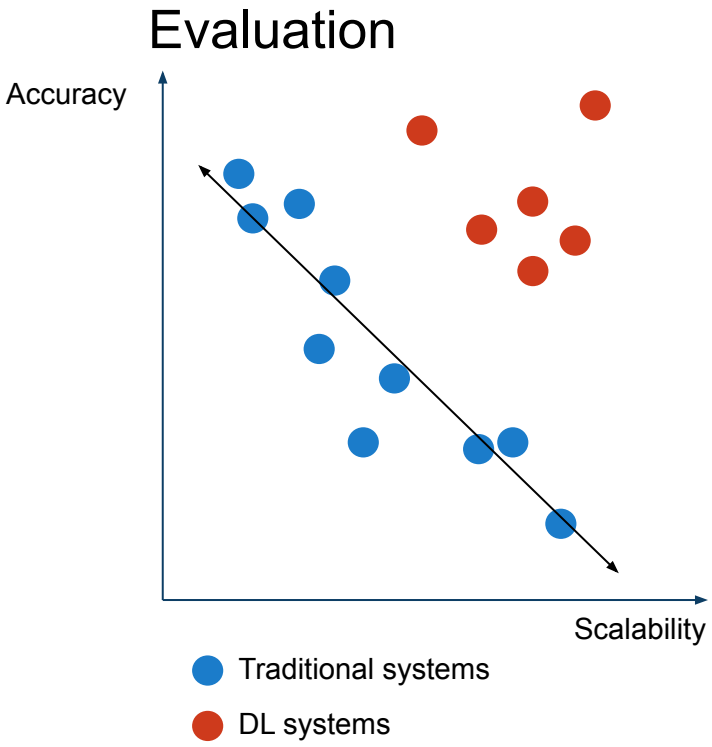


KR6. Environmental well-being

Training

Evaluation: version identification, accuracy-scalability plane.
Embedding distillation techniques: less storage, faster retrieval and similar accuracy (Yesiler et al. 2022).

NIME **Conference** Environmental Statement
https://eco.nime.org/



F. Yesiler, J. Serrà, E. Gómez. Less is more: Faster and better music version identification with embedding distillation, ISMIR 2020.

128x respectively, with a codebook size of 2048 for each level. The VQ-VAE has 2 million parameters and is trained on 9-second audio clips on 256 V100 for 3 days. We used exponential moving average to update the codebook following Razavi et al. (2019). For our prior and upsampler models, we use a context of 8192 tokens of VQ-VAE codes, which corresponds to approximately 24, 6, and 1.5 seconds of raw audio at the top, middle, and bottom level, respectively. The upsamplers have one billion parameters and are trained on 128 V100s for 2 weeks, and the top-level prior has 5 billion parameters and is trained on 512 V100s for 4 weeks. We use Adam with learning rate 0.00015 and weight decay of 0.002. For lyrics conditioning, we reuse the prior and add a small encoder, after which we train the model on 512 V100s for 2 weeks. The detailed hyperparameters for our models and training are provided in Appendix B.3.

Jukebox, Open AI (2020)

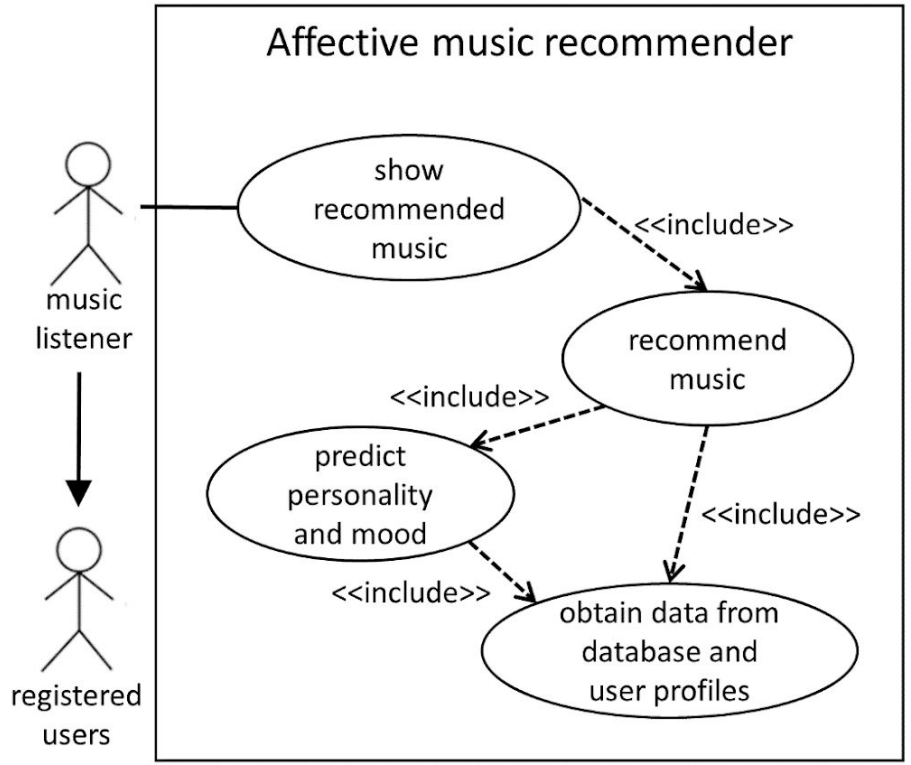
KR7. Accountability

Description

Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured.

Related topics

- Reproducibility and open data, code.
- Algorithmic audits.
- Specially difficult in complex systems.



K7: Which are the challenges to reproduce an existing paper, e.g. audit an IR system?

slido

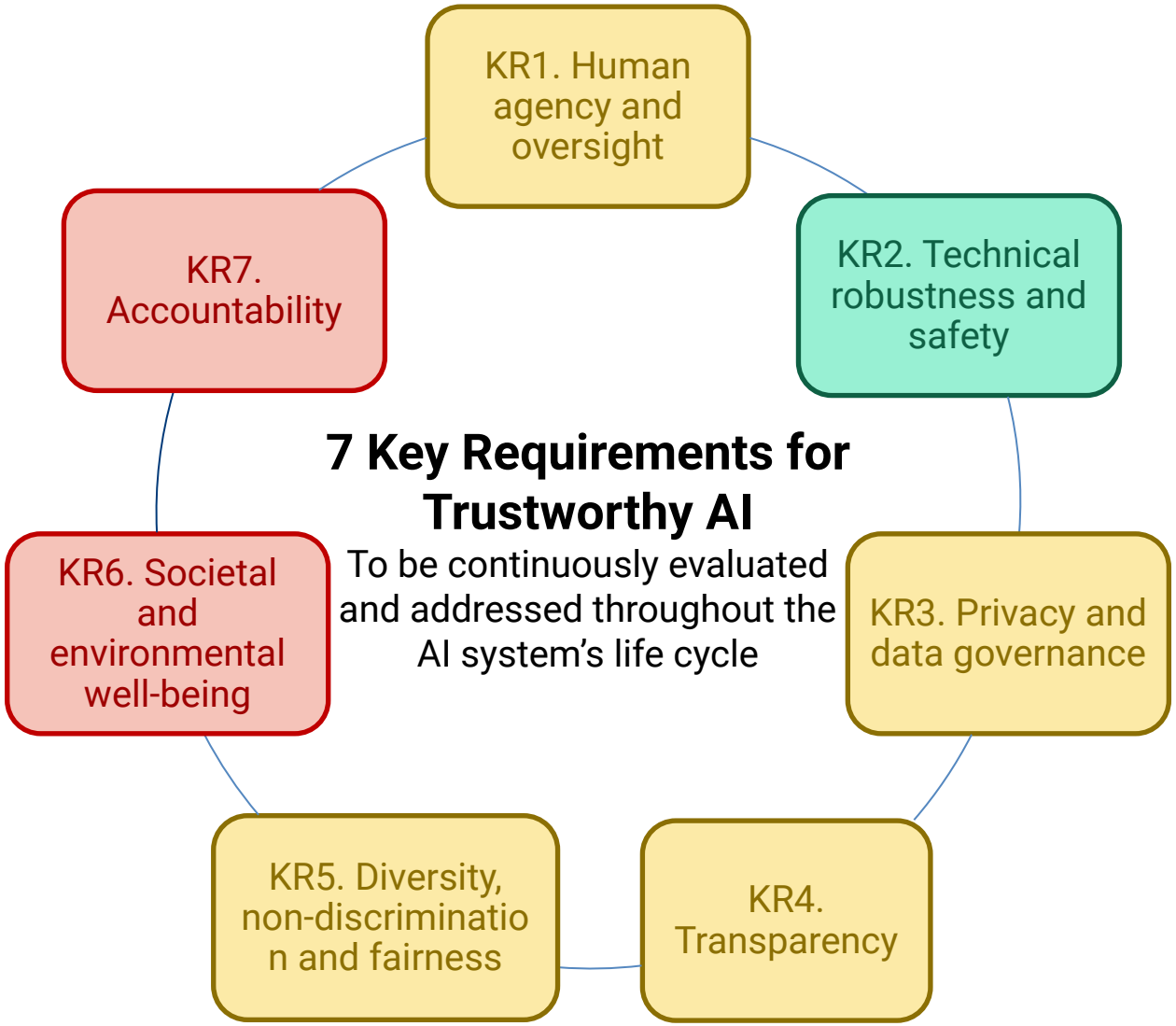
K7: Which are the challenges you have found to reproduce an algorithmic system?K7: Which are the challenges you have found to reproduce an algorithmic system?

ⓘ Start presenting to display the poll results on this slide.

Newly addressed

Some research

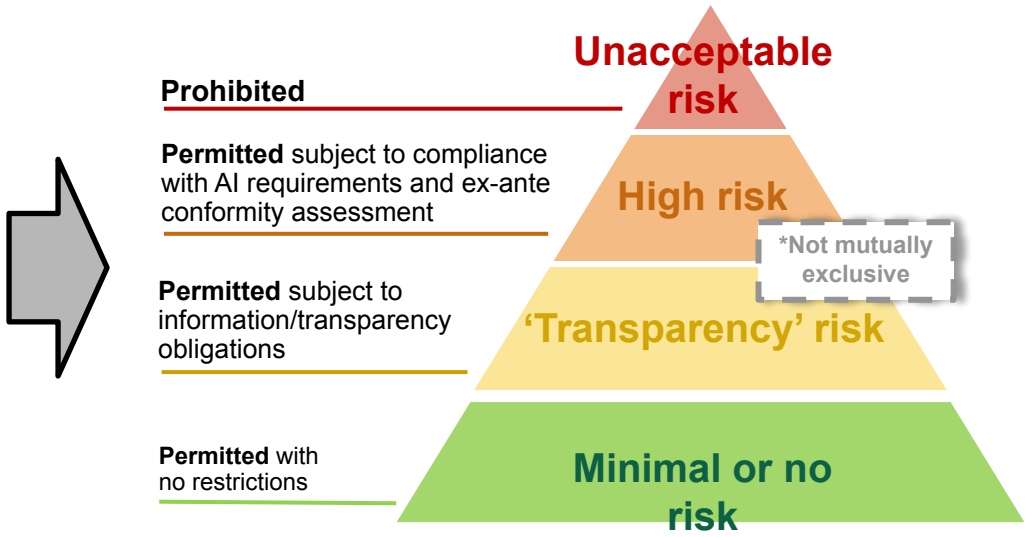
Strong background



From ethical guidelines to legal requirements

- KR1. Human agency and oversight
- KR2. Technical robustness and safety
- KR3. Privacy and data governance
- KR4. Transparency
- KR5. Diversity, non-discrimination and fairness
- KR6. Societal and environmental well-being
- KR7. Accountability

AI Act - Scope: AI systems (software products)



Legal requirements:

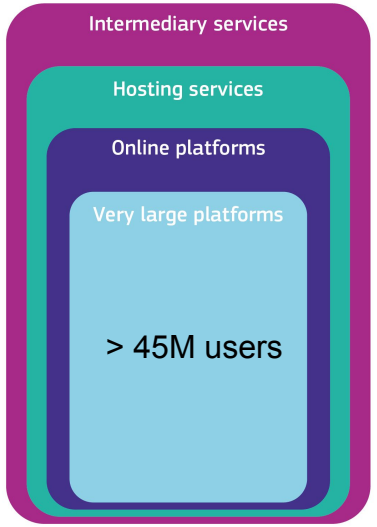
- Robustness, accuracy and cybersecurity.
- Human oversight (measures built into the system and/or to be implemented by users).
- Ensure appropriate degree of transparency and provide users with information on capabilities and limitations of the system and how to use it.
- Technical documentation and logging capabilities (traceability and auditability).
- High-quality and representative training, validation and testing data.
- Risk management.

AI Act: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 **Under negotiation**

From ethical guidelines to legal requirements

Digital Services Act - Scope: digital services (e.g. search engines, online platforms)

- Risk management.
- Transparency of recommender systems, online advertisement.
- External & independent auditing, internal compliance function and public accountability.
- Data sharing with authorities and researchers.
- Crisis response cooperation.



Digital Services Act:
https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en
Currently entering into force!

Setting the tone for a safer online space.

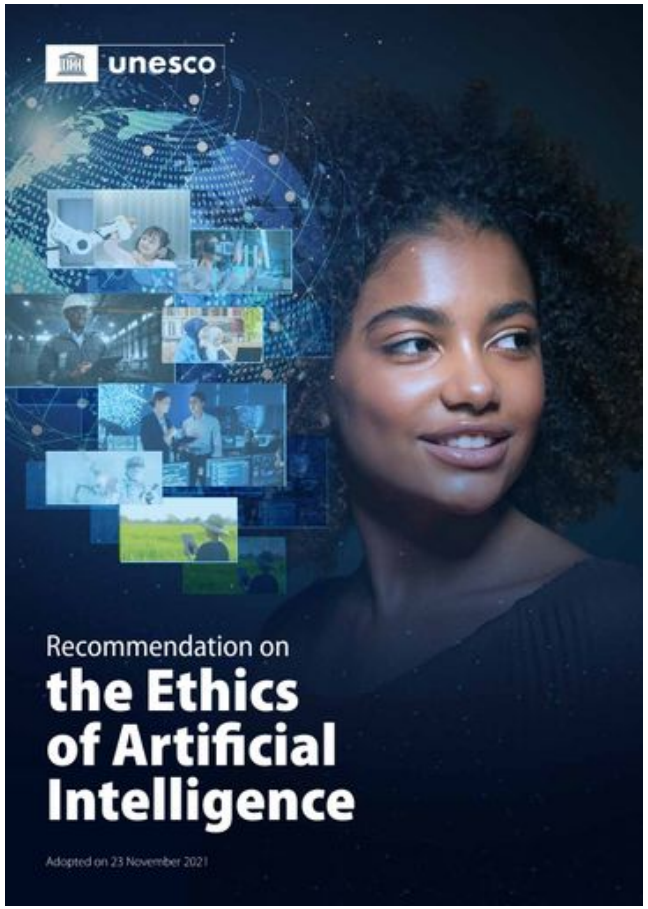
European Centre for Algorithmic Transparency

#DigitalEU

Towards worldwide recommendations

PRINCIPLES

- Proportionality and Do No Harm
- Safety and security
- Fairness and non-discrimination
- Sustainability
- Right to Privacy, and Data Protection
- Human oversight and determination
- Transparency and explainability
- Responsibility and accountability
- Awareness and literacy
- Multi-stakeholder and adaptive governance and collaboration



https://unesdoc.unesco.org/ark:/48223/pf0000381137

slido

Audience Q&A Session

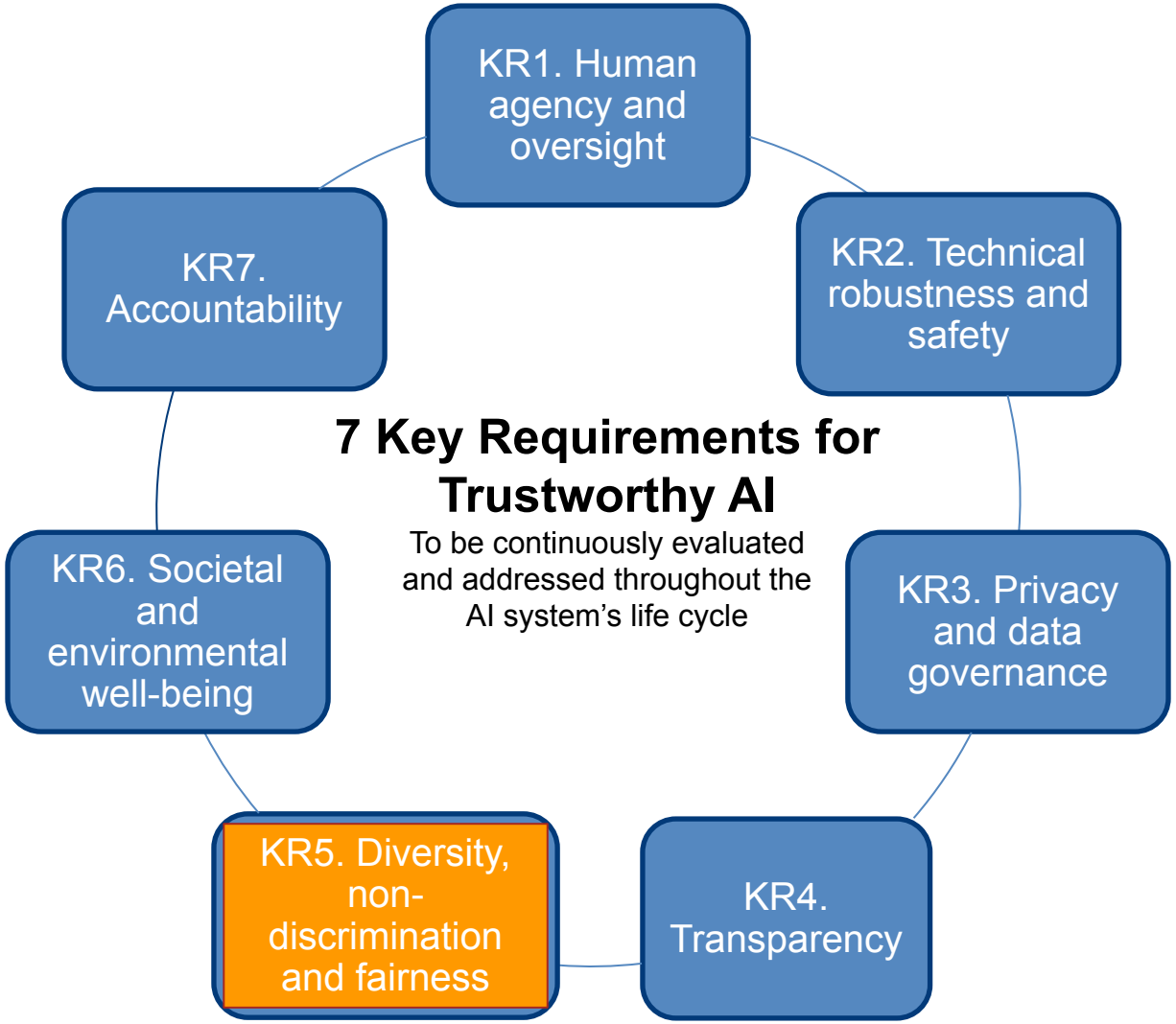
ⓘ Start presenting to display the audience questions on this slide.

Part 3:
Bias, Fairness, and Non-discrimination

Outline

- EU Regulation
- Bias from various perspectives
- Relation to fairness and non-discrimination
- Measuring biases (demographics, personality, popularity)
- Strategies to mitigate bias and improve fairness

Non-discrimination and Fairness are Key Requirements for Trustworthy AI



EU Regulations

- EU Regulatory Framework for AI (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai)
- EU Charter of Fundamental Rights (https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en)

Article 21: **Non-discrimination**

1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.
2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited.

Article 23: **Equality between women and men**

1. Equality between women and men must be ensured in all areas, including employment, work and pay.
2. The principle of equality shall not prevent the maintenance or adoption of measures providing for specific advantages in favour of the under-represented sex.

Biases from a High-level Perspective

- **Societal Bias:** Discrepancy between how the world should be and how it actually is (e.g., equal representation of genders in jobs/positions vs. actual over/underrepresentation of genders)

Biases from a High-level Perspective

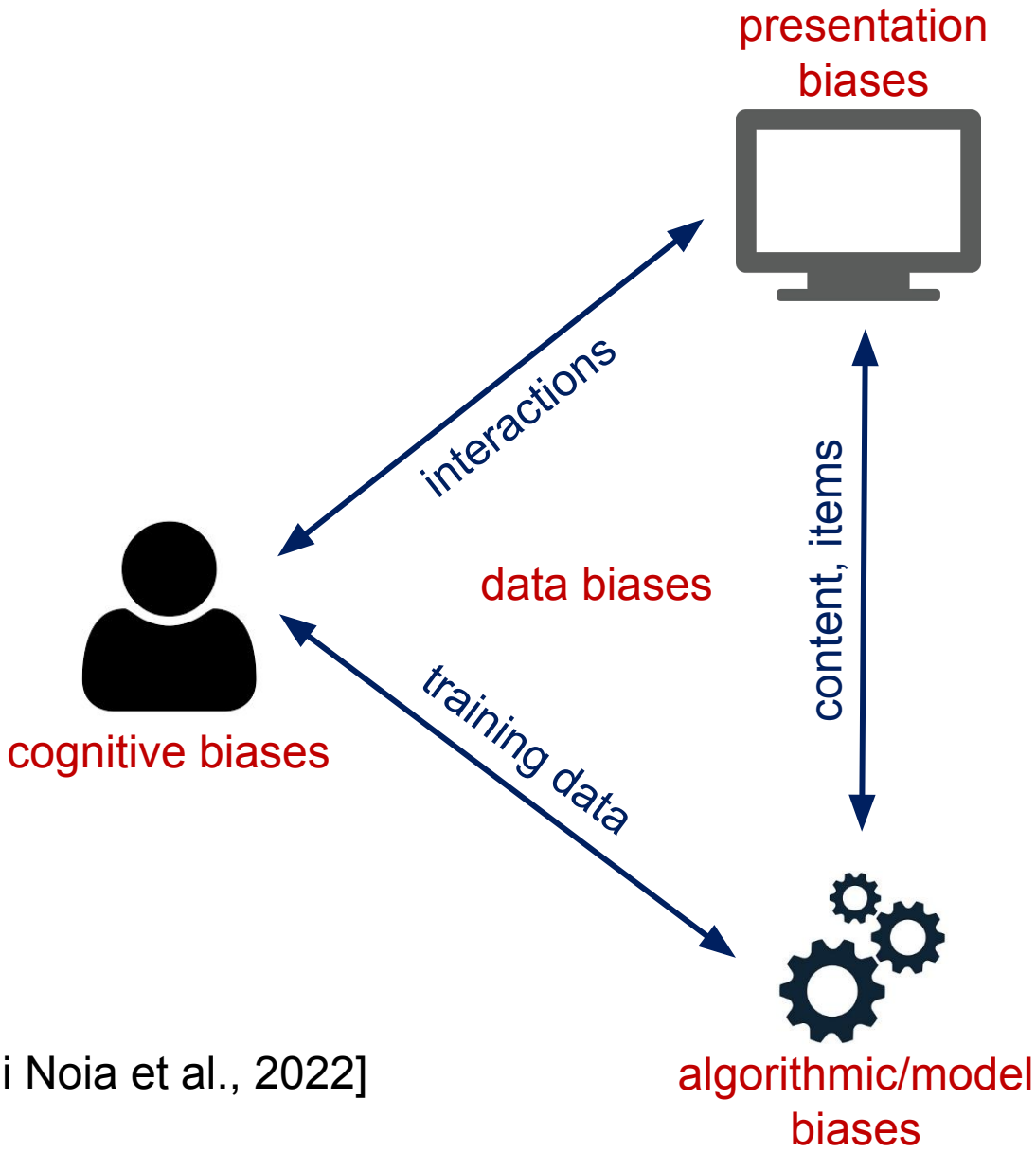
- **Societal Bias:** Discrepancy between how the world should be and how it actually is (e.g., equal representation of genders in jobs/positions vs. actual over/underrepresentation of genders)

- **Statistical Bias:** Discrepancy between how the world is and how it is encoded in the system or created machine learning model (e.g., data does not reflect population at large; in RSs often a community bias)

Biases in Retrieval and Recommender Systems

Decisions made by IR and RSs are affected by various biases (influencing each other), originating from:

- *Data:* e.g., unbalanced dataset w.r.t. group of users → demographic bias, community bias
- *Algorithms:* e.g., reinforcing stereotypes or amplify already popular content ("rich get richer" effect) → popularity bias
- *Presentation:* e.g., positions of recommended items on screen
- *User cognition or perception:* e.g., serial position effect, confirmation bias

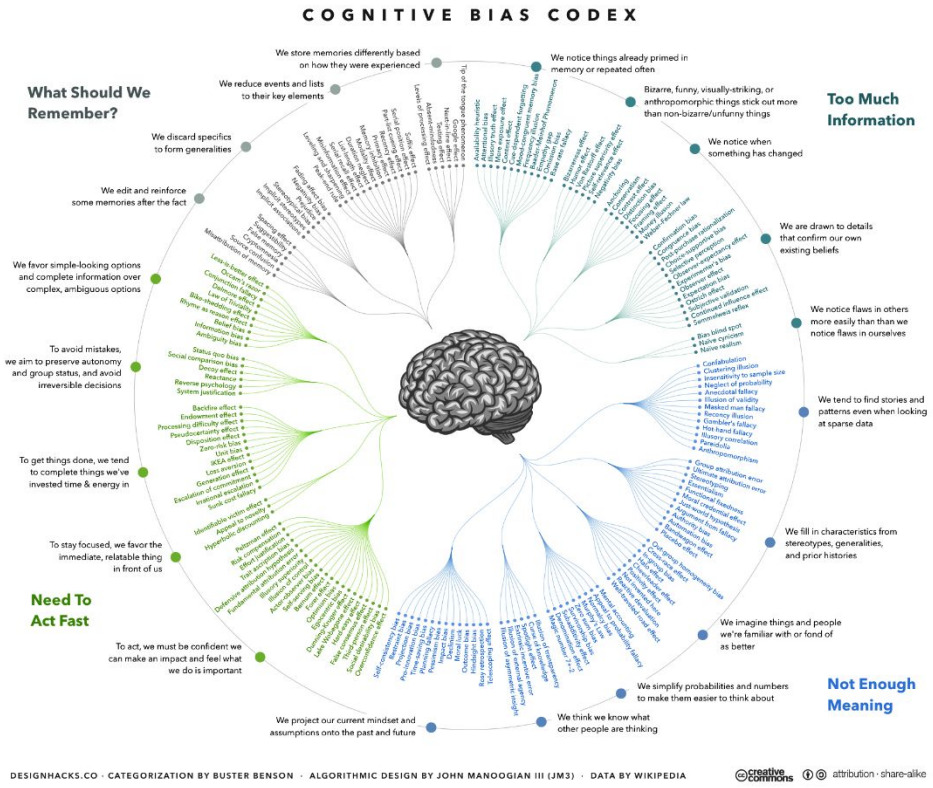


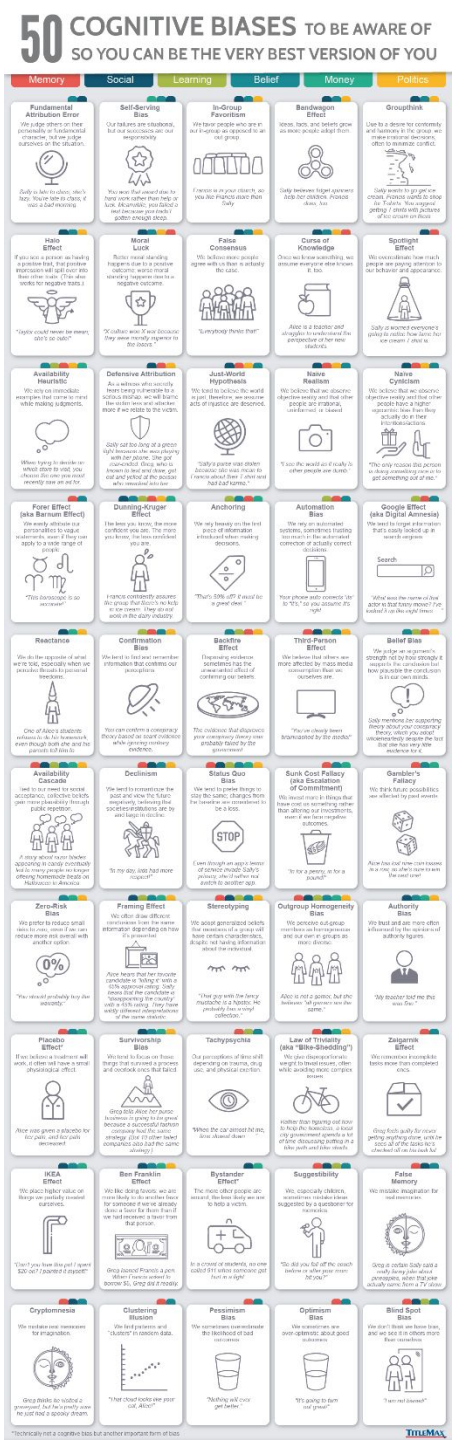
[Di Noia et al., 2022]

Biases in IR and RSs

Additional cognitive biases:

- https://en.wikipedia.org/wiki/List_of_cognitive_biases
- https://commons.wikimedia.org/wiki/File:Cognitive_bias_codex_en.svg
- https://www.visualcapitalist.com/50-cognitive-biases-in-the-modern-world





When are Biases Problematic?

Biases can result in different treatment of users or groups of users

> *“The system systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others.”* [Friedman and Nissenbaum, 1996]

When are Biases Problematic?

Biases can result in different treatment of users or groups of users

> *"The system systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others."* [Friedman and Nissenbaum, 1996]

However, not all biases are bad…

- Trade-off between personalization and fairness, i.e., the RS has to favor items that the user is likely to consume
- Case study: **Popularity bias** (i.e., overrepresentation of popular content)
  - Should a system recommend *all* content items with the same likelihood?
  - Should the popularity of items in the recommendation list match the popularity of items in the user's consumption history ("calibration")?
  - Should it match with the item popularity in the consumption history of all users of the system?

When are Biases Problematic?

However, not all biases are bad…

- Trade-off between personalization and fairness, i.e., the RS has to favor items that the user is likely to consume
- Case study: **Popularity bias** (i.e., overrepresentation of popular content)
  - Should a system recommend *all* content items with the same likelihood?
  - Should the popularity of items in the recommendation list match the popularity of items in the user's consumption history ("calibration")?
  - Should it match with the item popularity in the consumption history of all users of the system?

Making things even more complicated: **multiple stakeholders** are involved
(e.g., content producers, content consumers, platform providers, policymakers)

→ Finding an optimal level of popularity in recommendation results is tricky!
(often, *popularity calibration* is aimed for) e.g. [Abdollahpouri et al., 2021; Lesota et al., 2021]

slido

Have you experienced popularity bias when using retrieval or recommender systems?

ⓘ Start presenting to display the poll results on this slide.

slido

In which recommender systems or search engines have you already experienced popularity bias?

ⓘ Start presenting to display the poll results on this slide.

Fair for Whom?

- **Individual fairness:**
  Similar users are treated in a similar fashion (e.g., users with similar skills receive job recommendations within the same pay grade)

- **Group fairness:**
  Different groups of users defined by some sensitive or protected attribute (e.g., gender, age, or ethnicity) are treated in the same way. Accordingly, unfairness is defined as “systematically and unfairly discriminat[ing] against certain individuals or groups of individuals in favor of others.”

Bias Measurement

User Demographic Bias

[Melchiorre et al., 2021]

Metric: *RecGap* measures performance difference of the RS for different user groups

$$RecGap^{\mu} = \frac{\sum_{\langle g,g' \rangle \in G^{pair}} \left| \frac{\sum_{u \in U_g} \mu(u)}{|U_g|} - \frac{\sum_{u' \in U_{g'}} \mu(u')}{|U_{g'}|} \right|}{|G^{pair}|}$$

Average difference in performance metric μ between all pairs of user groups G^{pair}

μ precision, recall, NDCG, or beyond-accuracy metrics (e.g., coverage or diversity)

U_g set of users in group g, e.g. defined by gender, ethnicity, age, country

→ *RecGap* considers a RS to be fair if it performs *equally good* across the groups

User Demographic Bias

[Melchiorre et al., 2021]

Metric: *RecGap* measures performance difference of system for different user groups

Model	Scenario	All	M/F	*RecGap*
POP	STANDARD	.046	.045/.049	.004 (f)
	RESAMPLED	.045	.044/.051	.007 (f) †
ItemKNN	STANDARD	.301	.313/.259	.054 (m) †
	RESAMPLED	.292	.304/.250	.054 (m) †
BPR	STANDARD	.127	.129/.117	.012 (m) †
	RESAMPLED	.123	.124/.116	.008 (m)
ALS	STANDARD	.241	.251/.205	.046 (m) †
	RESAMPLED	.238	.248/.204	.044 (m) †
SLIM	STANDARD	.364	.378/.315	**.063** (m) †
	RESAMPLED	.359	.372/.312	**.060** (m) †
MultiVAE	STANDARD	.192	.197/.173	.024 (m) †
	RESAMPLED	.183	.188/.166	.023 (m) †

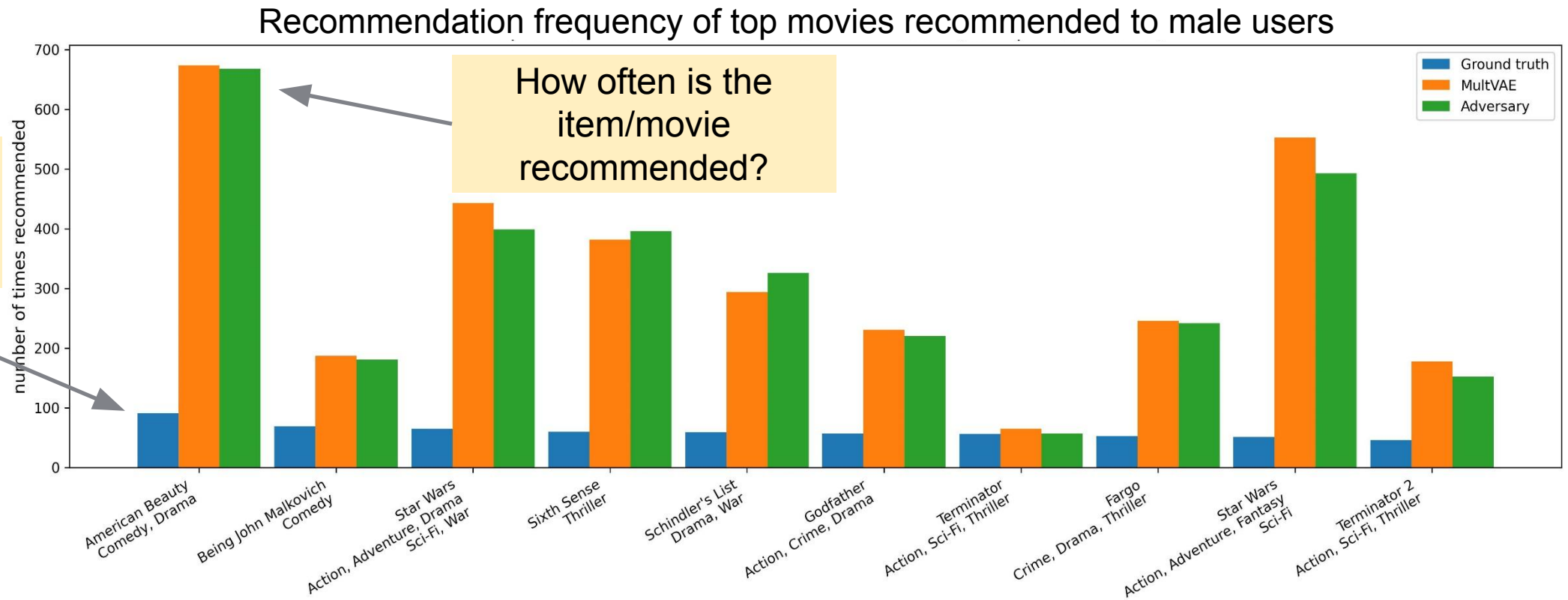
- Majority of CF-based algorithm provide worse recommendations to female than to male users (w.r.t. NDCG and Recall)
- Mostly inverse relationship between accuracy (NDCG, Recall) and fairness

Popularity Bias: Simple Example

[Lesota et al., 2021]

Metric: Difference between an item's recommendation frequency and consumption frequency in user profiles

How often is the item/movie consumed?



Popularity Bias: More Formal / Delta Metrics

[Lesota et al., 2021]

Metrics: *"Delta" metrics* and *distribution-based metrics*
Assumption: Users prefer "calibrated" recommendations, i.e., the RS should mimic the input distribution w.r.t. an attribute (popularity in our case): $pop\big(H_{u_i}(p_j)\big) \sim pop\big(R_{u_i}(p_j)\big)$

pop some measure of popularity
(e.g., number of interactions with item p_j, over all users, or number of users)
H_{u_i} list of user u_i's interaction history (over items p_j)
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., Kullback-Leibler divergence or Kendall's τ)

Popularity Bias: Delta Metrics

[Lesota et al., 2021]

Metrics: *"Delta" metrics*

$\%\Delta\xi$ "percent Delta Xi" ~ relative popularity difference in terms of statistical measure ξ

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

ξ statistical measure or moment of interest (mean, median, variance, skew, etc.)

→ Positive $\%\Delta Mean$ and $\%\Delta Median$ indicate that more popular tracks are recommended to user u_i than warranted given his or her consumption profile ("miscalibration")

→ Positive $\%\Delta Variance$ indicate that recommendation list is more diverse w.r.t. covering differently popular items than user u_i's consumption profile

Aggregate over all users (a RS's bias): $\%\Delta\xi = Median\left(\%\Delta\xi(u_i)\right)$

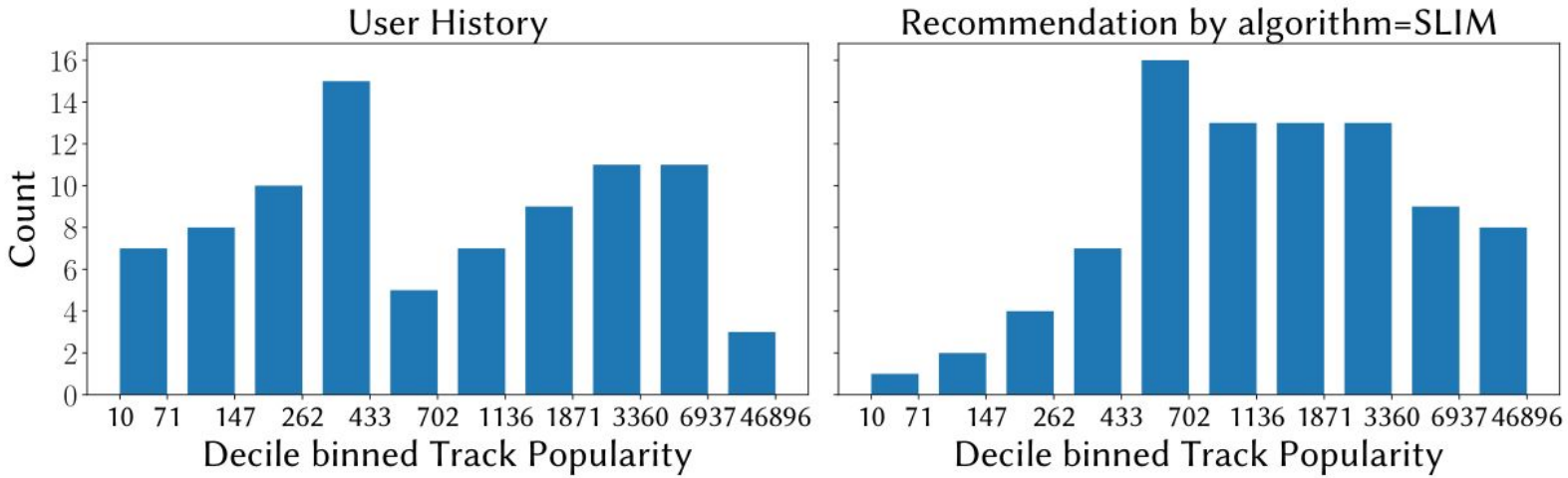
Popularity Bias: Distribution-based Metrics

[Lesota et al., 2021]

Metrics: *Distribution-based metrics*

Considers the binned and normalized item popularities as (probability) distribution and computes:

- Kullback-Leibler (KL) divergence: ~dissimilarity between the two distributions
- Kendall's τ: ~degree to which the order of bins is the same for the two distributions when ranked according to the respective counts



Popularity Bias: Empirical Results

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 ($\%\Delta Mean$)
- ALS and VAE particularly
- ItemKNN least
- ALS and VAE increase also diversity ($\%\Delta Var.$)

[Lesota et al., 2021]

Popularity Bias: Empirical Results

Popularity Bias can be combined with User Demographic Bias

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 create an even higher popularity bias for female users than for male users (+/- values are relative to values in row *All*)

[Lesota et al., 2021]

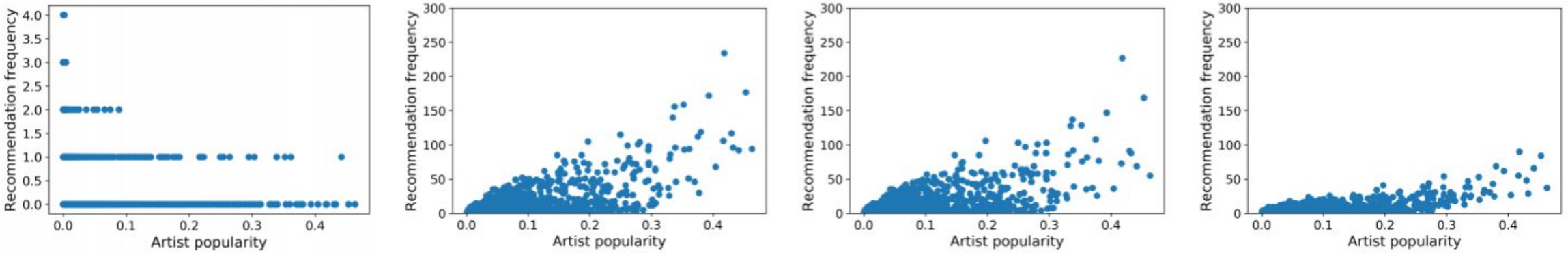
slido

As a user of a recommender system, would you rather accept positive or negative popularity bias (i.e., overly popular items or overly unpopular items) in your recommendation list?

ⓘ Start presenting to display the poll results on this slide.

Popularity Bias: Another Variant

- RQ: Is the popularity level/*mainstreaminess* of users' listening preferences accurately reflected in recommendations made by algorithms?
- ~3K Last.fm users of different mainstreaminess (low, medium, high), selected from LFM-1b (dataset of 1B music listening records from Last.fm)
- Algorithms: User-based CF (KNN), NMF, UserItemAvg, Random, Most Popular
- Correlation between (artist) popularity and frequency of recommendation:

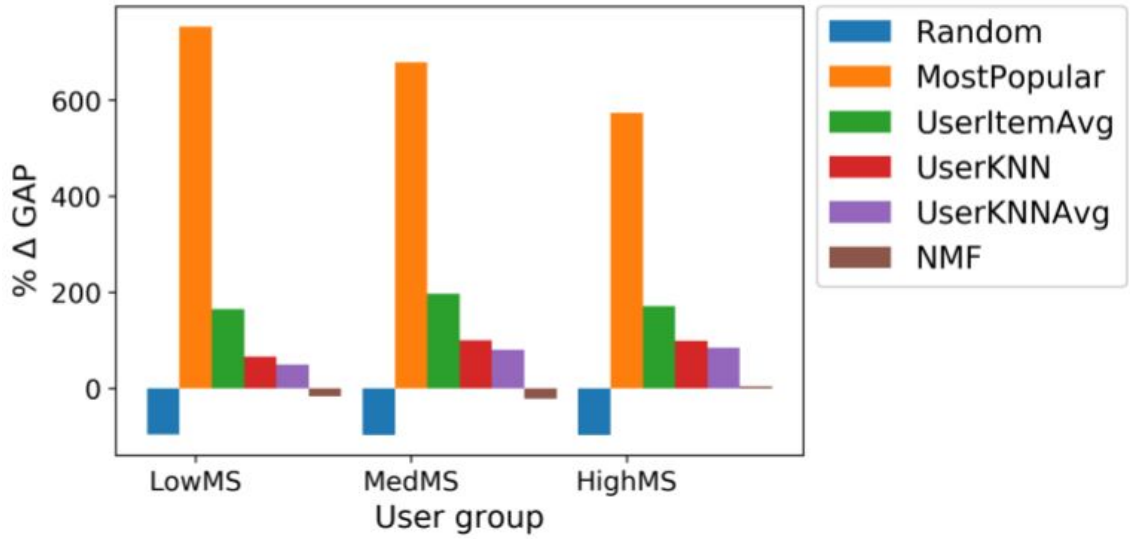


All RS algorithms favor popular artists (except for Random), irrespective of user preferences

[Kowald et al., 2020]

Popularity Bias: Another Variant

- RQ: Is the popularity level/*mainstreaminess* of users' listening preferences accurately reflected in recommendations made by algorithms?
- Taking user preferences towards popular artists (mainstreaminess) into account:
- Metric: Difference in GAP (Group Average Popularity)
- $\Delta GAP = \frac{GAP(group,rec) - GAP(group,pref)}{GAP(group,pref)}$
- Measures extent to which popularity of recommendations exceed popularity of items in user profile



Most RS algorithms favor popular artists, irrespective of user preferences

Personality Bias

- RQ: Do (music) recommender algorithms treat users with different personality traits equally?
- ~18K Twitter users (extracted music listening events; inferred personality traits from posts)
- Traits (high/low groups): openness, conscientiousness, extraversion, agreeableness, neuroticism
- Algorithms: SLIM, EASE (shallow AE), Mult-VAE

Neurotic people seem to have a narrow music taste

Group		Agr.	Con.	Ext.	Neu.	Ope.
High	No. unique tracks/user (mean and std.)	19.1 ± 24.4	19.2 ± 25.5	20.0 ± 26.3	16.2 ± 18.4	19.5 ± 24.9
	No. unique tracks	15,694	15,674	15,655	15,429	15,652
	No. listening events	208,054	206,179	217,895	177,892	209,741
Low	No. unique tracks/user (mean and std.)	17.3 ± 21.7	17.2 ± 20.4	16.4 ± 19.2	20.3 ± 26.9	16.9 ± 21.1
	No. unique tracks	15,664	15,695	15,672	15,607	15,619
	No. listening events	187,002	188,877	177,161	217,164	185,315

[Melchiorre et al., 2020]

Personality Bias: Empirical Results

- RQ: Do music recommender algorithms treat users with different personality traits equally?
- Summary of results:
  - *Open* users receive worse recommendations (than narrow-minded users)
  - *Neurotics* receive better recommendations
  - *Extraverts* receive worse recommendations
  - *Conscientious* users get worse recommendations
  - Differences for *agreeableness* not very pronounced

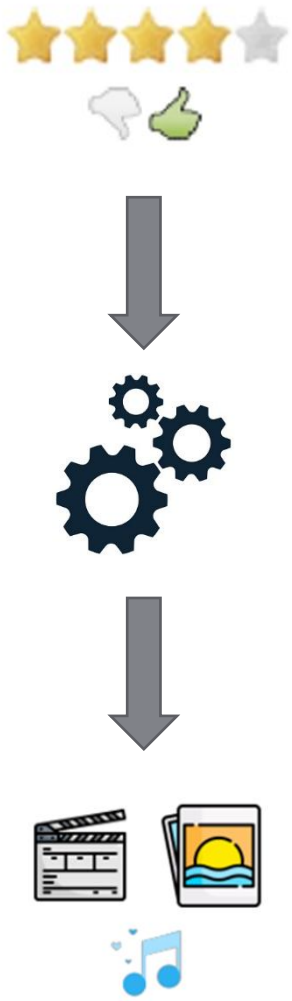
Trait	Algorithm	@5 All	@5 High	@5 Low
Agr.	EASE	0.0311	0.0295	0.0327
	SLIM	0.0279	0.0263	0.0295
	Mult-VAE	0.0380	**0.0385***	**0.0374***
Con.	EASE	0.0311	**0.0274***	**0.0349***
	SLIM	0.0279	**0.0241*****	**0.0319*****
	Mult-VAE	0.0380	0.0353	0.0407
Ext.	EASE	0.0311	**0.0266****	**0.0355****
	SLIM	0.0279	**0.0242****	**0.0317****
	Mult-VAE	0.0380	**0.0340****	**0.0417****
Neu.	EASE	0.0311	**0.0366*****	**0.0257*****
	SLIM	0.0279	**0.0335*****	**0.0224*****
	Mult-VAE	0.0380	**0.0436*****	**0.0324*****
Ope.	EASE	0.0311	**0.0221*****	**0.0400*****
	SLIM	0.0279	**0.0196*****	**0.0363*****
	Mult-VAE	0.0380	**0.0285*****	**0.0473*****

Performance of RS algorithms differs substantially between users of different personality

[Melchiorre et al., 2020]

Bias Mitigation

Strategies to Mitigating Harmful Biases



Pre-processing strategies

- *Data rebalancing* (e.g., upsample minority group, subsample majority group) e.g. [Melchiorre et al., 2021]

In-processing strategies

- *Regularization* (e.g., include bias correction term/bias metric in loss function used to train a model)
- *Adversarial learning* (e.g., train a classifier that predicts the sensitive attribute and adapt model parameters to minimize performance of this classifier) e.g. [Ganhör et al., 2022]

Post-processing strategies

- *Filter* items (e.g., remove items from overrepresented groups)
- *Reweigh/Rerank* recommendations in list e.g. [Ferraro et al., 2021]

Mitigating Harmful Biases (Pre-processing Strategy)

Ex.: Data Rebalancing

[Melchiorre et al., 2021]

Upsample data points by female user (to same amount created by male users)

Model	Scenario	All	M/F	*RecGap*
POP	STANDARD	.046	.045/.049	.004 (f)
	RESAMPLED	.045	.044/.051	.007 (f) †
ItemKNN	STANDARD	.301	.313/.259	.054 (m) †
	RESAMPLED	.292	.304/.250	.054 (m) †
BPR	STANDARD	.127	.129/.117	.012 (m) †
	RESAMPLED	.123	.124/.116	.008 (m)
ALS	STANDARD	.241	.251/.205	.046 (m) †
	RESAMPLED	.238	.248/.204	.044 (m) †
SLIM	STANDARD	.364	.378/.315	**.063** (m) †
	RESAMPLED	.359	.372/.312	**.060** (m) †
MultiVAE	STANDARD	.192	.197/.173	.024 (m) †
	RESAMPLED	.183	.188/.166	.023 (m) †

NDCG gap between male and female users narrows, but foremost due to male users' decrease in recommendation quality

Mitigating Harmful Biases (In-processing Strategy)

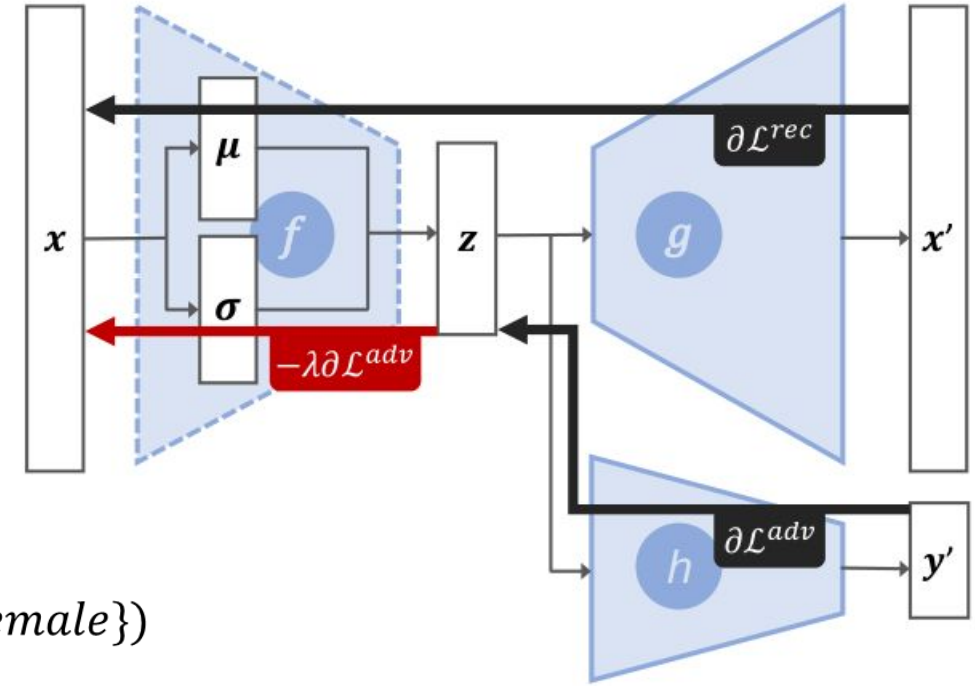
[Ganhör et al., 2022]

Ex.: Adversarial Learning

Unlearn implicit information of protected attributes while preserving accuracy

Adversarial Mult-VAE architecture:

$f(\cdot)$	encoder network
$g(\cdot)$	decoder network
$h(\cdot)$	adversarial network
x	multi-hot encoded vector of item interactions
x'	reconstruction of x
z	latent representation
y'	prediction of protected attribute (e.g., gender $\in \{male, female\}$)



$$\arg\min_{f,g} \arg\max_{h} \mathcal{L}^{\text{rec}}(\boldsymbol{x}) - \mathcal{L}^{\text{adv}}(\boldsymbol{x}, \boldsymbol{y})$$

Mitigating Harmful Biases (In-processing Strategy)

Ex.: Adversarial Learning [Ganhör et al., 2022]

Unlearn implicit information of protected attributes while preserving accuracy

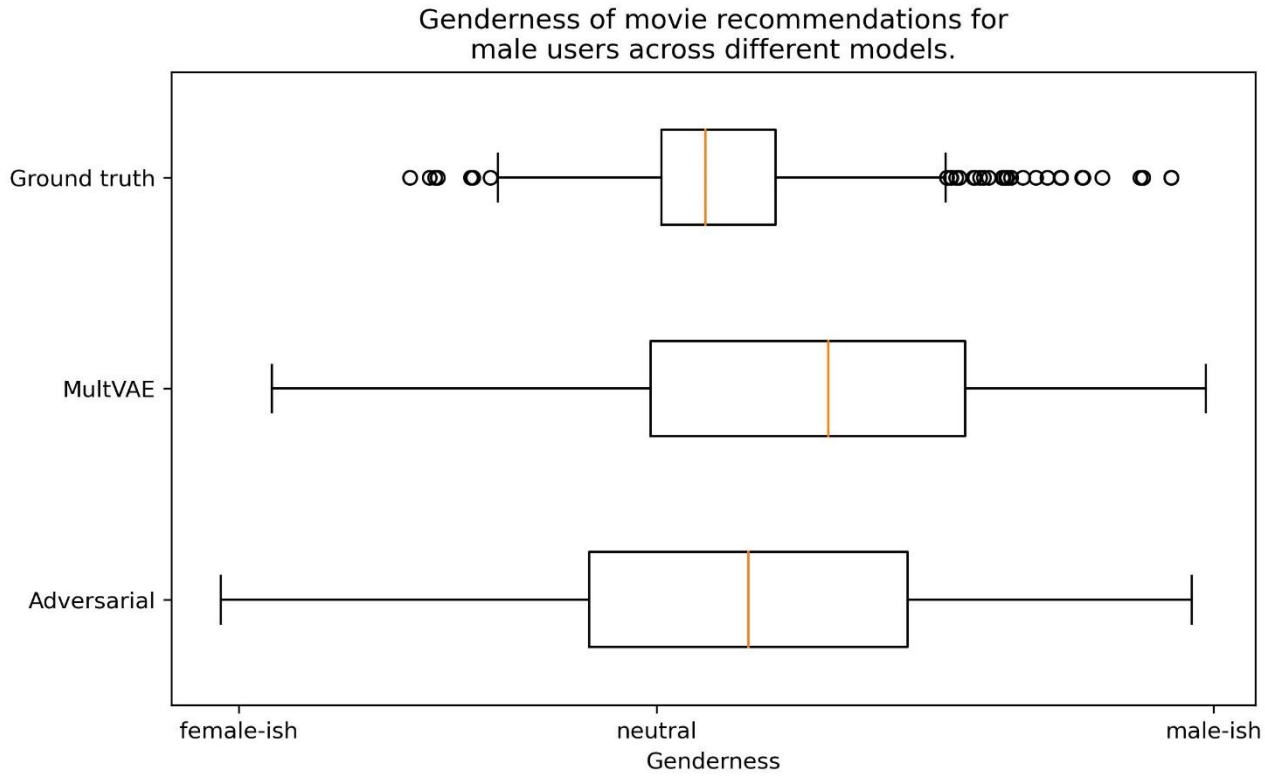
Dataset	Model	Bias↓		Performance↑	
		Acc	BAcc	NDCG	Recall
ML-1M	MultVAE$_{Best}$	0.692	0.707	**0.621**	**0.596**
	MultVAE$_{Last}$	0.699	0.693	0.591†	0.566†
	Adv-MultVAE	**0.565**	**0.572**	0.593†	0.569†
LFM2b-DB	MultVAE$_{Best}$	0.703	0.717	**0.211**	**0.192**
	MultVAE$_{Last}$	0.709	0.717	0.206†	0.189†
	Adv-MultVAE	**0.631**	**0.609**	0.206†	0.189†

Substantial reduction of encoded protected information at expense of a marginal performance decrease

Mitigating Harmful Biases (In-processing Strategy)

Ex.: Adversarial Learning [Ganhör et al., 2022]

Unlearn implicit information of protected attributes while preserving accuracy



Amount of typical female (male) content is reduced for female (male) users

Mitigating Harmful Biases (Post-processing Strategy)

Ex.: Reranking

[Ferraro et al., 2021]

Penalize/downrank content by the majority group (male artists) by λ positions in the recommendation list, created with ALS CF approach

last.fm

	Algo	Avg position 1st female	Avg position 1st male	% females rec.	
LFM-1b	ALS	6.7717	0.6142	25.44	cf. female artists in dataset: 23.25%
	POP	0.1325	1.7299	32.44	
	RND	3.3015	0.3046	23.30	
LFM-360k	ALS	8.3165	0.7136	26.27	cf. female artists in dataset: 22.67%
	POP	0.9191	0.2713	29.31	
	RND	3.3973	0.2951	22.77	

slido

Which statement, in your opinion, best describes the (un)fairness of the following RS: 23% of items in the collection have been created by females. On average, 26% of items recommended by the system have been created by women.

ⓘ Start presenting to display the poll results on this slide.

Mitigating Harmful Biases (Post-processing Strategy)

Ex.: Reranking

[Ferraro et al., 2021]

Penalize/downrank content by the majority group (male artists) by λ positions in the recommendation list, created with ALS CF approach

	Algo	Avg position 1st female	Avg position 1st male	% females rec.
LFM-1b	ALS	6.7717	0.6142	25.44
	POP	0.1325	1.7299	32.44
	RND	3.3015	0.3046	23.30
LFM-360k	ALS	8.3165	0.7136	26.27
	POP	0.9191	0.2713	29.31
	RND	3.3973	0.2951	22.77

cf. female artists in dataset: 23.25%

cf. female artists in dataset: 22.67%

Female artists tend to occur further down in the recommendation lists

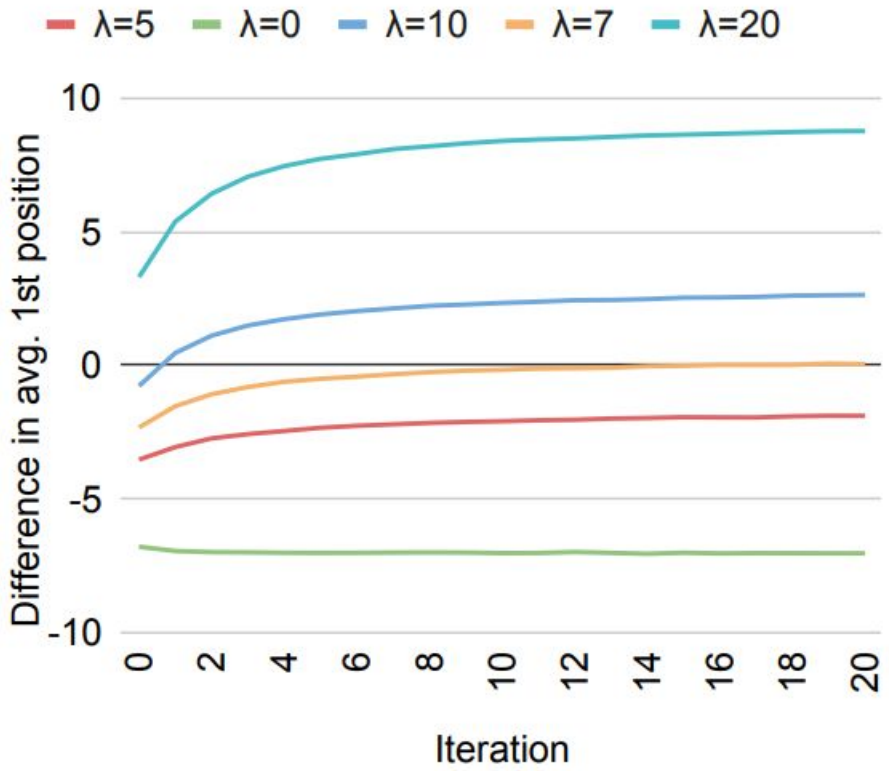
→ position bias

Mitigating Harmful Biases (Post-processing Strategy)

Ex.: Reranking

[Ferraro et al., 2021]

- Penalize/downrank content by the majority group (male artists) by λ positions
- Simulation study: In each iteration it is assumed that the top-10 recommendations are interacted with by the user, and the RS (ALS) is retrained accordingly



Positive feedback loop increases exposure of female artists

Summary

- Biases are everywhere, not only in computer systems
- All algorithmic ranking systems have to cope with a variety of biases
- Some of them are desired, because they enable personalized results
- Some of them cause unfair behavior (i.e., treat different users/stakeholders differently)
- Most researched biases include popularity bias and demographic biases
- Coping strategies include pre-, in-, and post-processing techniques
- Many open questions (e.g., perceived bias vs. offline metrics) [Ferwerda et al., 2023]

References

[Abdollahpouri et al., 2021]: *User-centered Evaluation of Popularity Bias in Recommender Systems*, Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP), Utrecht, The Netherlands, 2021.

[Chen et al., 2023]: *Bias and Debias in Recommender System: A Survey and Future Directions*, ACM Transactions on Information Systems 41(3), 67:1-39, 2023.

[Di Noia et al. 2022]: *Recommender systems under European AI regulations*. Communications of the ACM 65(4): 69-73, 2022.

[Ekstrand et al., 2021]: *Fairness and Discrimination in Information Access Systems,* CoRR abs/2105.05779, 2021.

[Ferraro et al., 2021]: *Break the Loop: Gender Imbalance in Music Recommenders*, Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR), Canberra, Australia, 2021.

[Ferwerda et al., 2023]: *I Don't Care How Popular You Are! Investigating Popularity Bias From a User's Perspective*, Proceedings of the 8th ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023),Austin, TX, USA, March 2023.

[Friedman and Nissenbaum, 1996]: *Bias in Computer Systems*, ACM Transactions on Information Systems 14(3):330-347, 1996.

[Ganhör et al., 2022]: *Mitigating Consumer Biases in Recommendations with Adversarial Training*, Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022), Madrid, Spain, 2022.

[Kowald et al., 2020]: *The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study*, Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020), Lisbon, Portugal, 2020.

[Lesota et al., 2021]: *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, 2021.

[Melchiorre et al., 2020]: *Personality Bias of Music Recommendation Algorithms*, Proceedings of the 14th ACM Conference on Recommender Systems (RecSys 2020), Virtual, 2020.

[Melchiorre et al., 2021]: *Investigating Gender Fairness of Recommendation Algorithms in the Music Domain*, Information Processing & Management, 58(5), 2021.

slido

Audience Q&A Session

ⓘ Start presenting to display the audience questions on this slide.

Break

Part 4: Transparency

Outline

- **Motivation & EU regulations**
- Categories of Transparency
- Explainability
- Traceability and Auditability
- Documentation

Motivation

- IR and RS systems should be able to *explain their decisions*
  - why are results shown to a user
  - how were results retrieved
  - help user assess whether to trust the system

  → Particularly when decision making involves sensitive aspects
- More reasons:
  - Reproducibility
  - Accountability
  - System diagnostics & performance

EU Regulations

- Transparency key feature of EU law
- Also: expression of fairness principle related to processing personal data as described in Article 8 of the Charter of Fundamental Rights of the EU
- EU General Data Protection Regulation (GDPR)
  - Transparency overarching obligation
- 3 central areas:
  - Provision of information to data subjects related to fair processing
  - How data controllers communicate with data subjects in relation to their rights under GDPR
  - How data controllers facilitate the exercise by data subjects of their rights
- Compliance with transparency required related to data processing under Directive 2016/680

https://iapp.org/media/pdf/resource_center/wp29-transparency-12-12-17.pdf

EU Regulations

- Digital Services Act
  - Online platforms & search engines need to be transparent in terms of recommender systems
  - Plus, advertisements
  - Requirements depend on size of platform measured by number of users
- Artificial Intelligence Act
  - Transparency as a key requirement
  - Besides: technical documentation for high-risk use cases

https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environmen
https://eur-lex.europa.eu/legal-content/NL/TXT/HTML/?uri=CELEX:52021PC0206&from=EN
t

One of the requirements for trustworthy AI

High-level Expert Group on AI, European Commission, Ethics Guidelines for Trustworthy AI, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

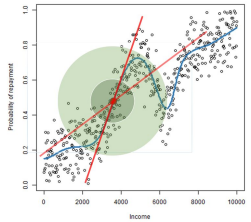
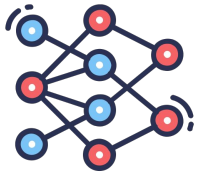
Transparency and Fairness

- Fair systems not possible if systems are opaque
  - How do algorithms work: what is in the data
  - How are end users affected
- Transparency enables audits
  - How does the system work
  - And: does system creates fair outputs
- User perceptions of fairness
  - IR /RS explanations may lead to new behavior
  - Taking fair actions; at least, informed choices

Outline

- Motivation & EU regulations
- **Categories of Transparency**
- Explainability
- Traceability and Auditability
- Documentation

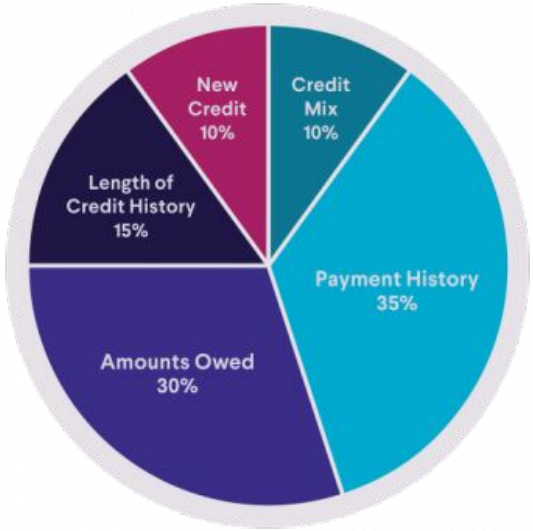
Major Aspects of Transparency

Algorithmic Transparency	Simulatability	Decomposability
		

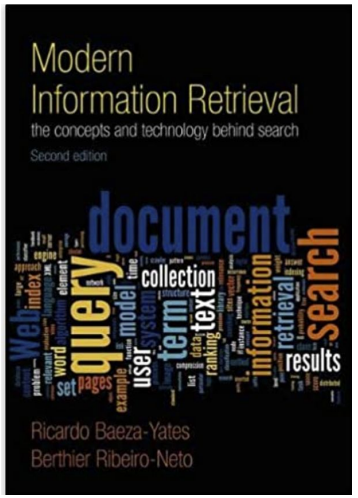
Related concepts: Explainability, Interpretability, Understandability, Black boxes

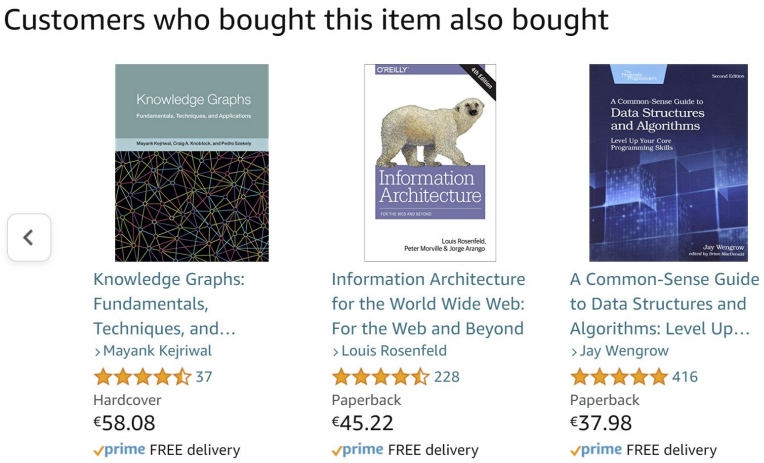
Transparency - Understandability

- Decision made by an algorithm should be understandable by those affected by the decision
  - Why was a decision reached based on a given input?

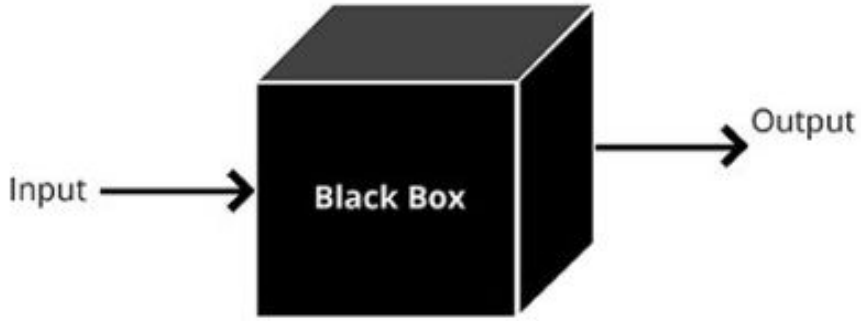


https://www.sofi.com/learn/content/do-personal-loans-hurt-credit/





The Problem of Black Boxes



- Contemporary IR & RS based on complex models: deep learning, ML
- We do not understand what is going on in the box
- Hard for users to understand why output is relevant - trust the prediction?

slido

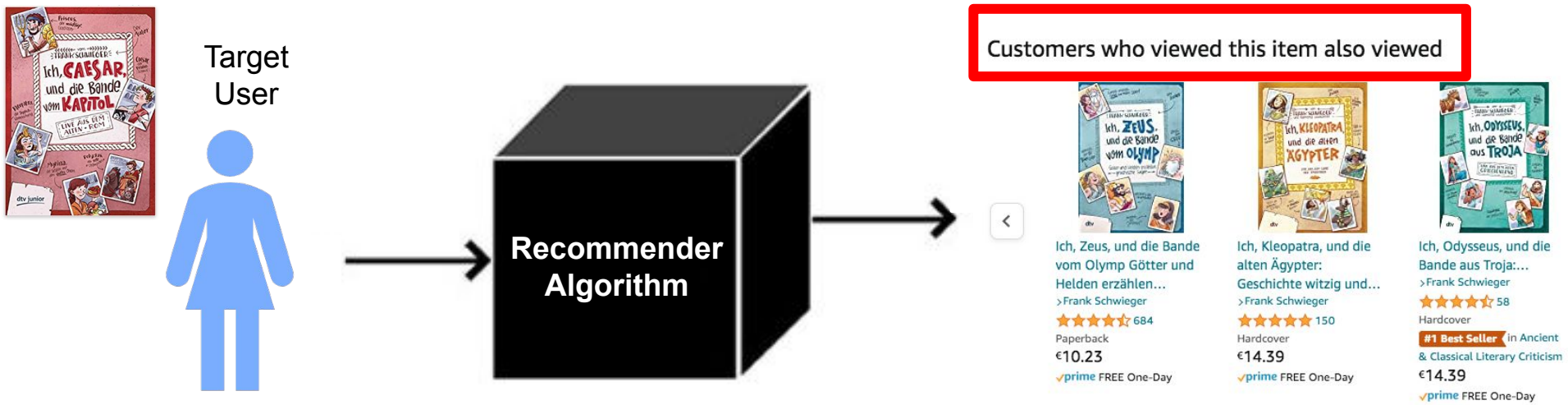
Do you think it is sufficient to disclose how algorithms came to their decision and tell how human could reverse the decision? Why yes? Why now?

ⓘ Start presenting to display the poll results on this slide.

Outline

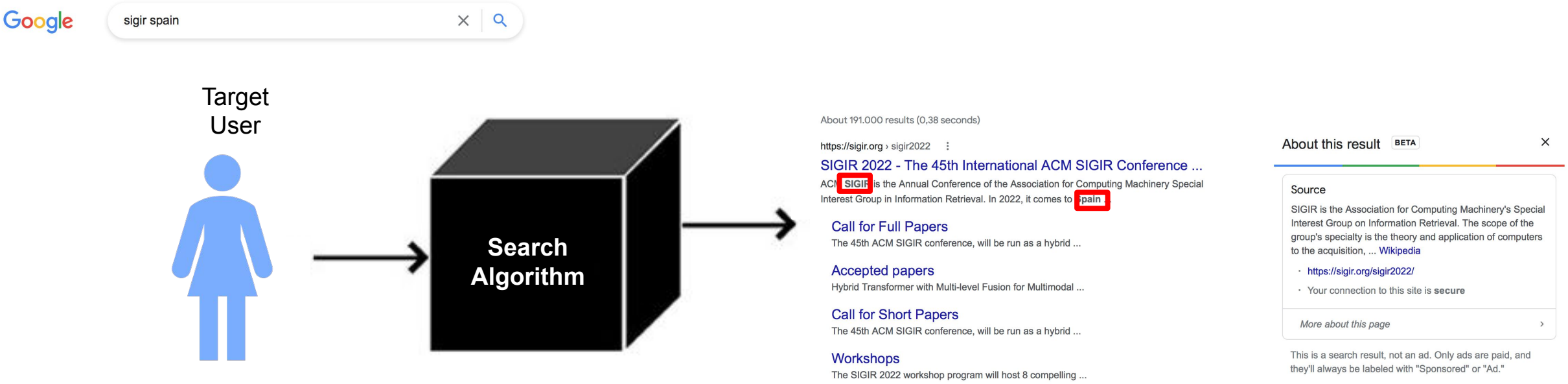
- Motivation & EU regulations
- Categories of Transparency
- **Explainability**
- Traceability and Auditability
- Documentation

Explanations in Recommender Systems



Task: Given user-item pair, provide **explanation** to justify why item is recommended to the user

Explanations in IR



Explanations in the form of search snippets, query terms highlighted

Additional information to the search result

Why Explainability?

- Increasingly important role in user interactions with systems
  - Trust in the system
  - Accountability
- Model validation
- Biases, unfairness, problems with training data, legal requirements
- Improvements of model
  - Reliability, robustness,..

slido

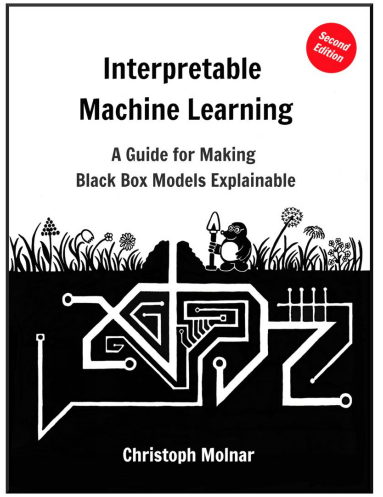
What makes a good explanation?

ⓘ Start presenting to display the poll results on this slide.

Properties of Good Explanations

- Accuracy
- Fidelity
- Consistency
- Stability
- Comprehensibility

→ see: https://christophm.github.io/interpretable-ml-book/



Explainability in Recommender Systems

“To make clear by giving a detailed description” (Tintarev et al.)

“Explainable recommendation to answer the question of why” (Zhang et al.)

Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. In Recommender systems handbook (pp. 479–510). Springer, Boston, MA
Zhang, Y. and Chen, X. (2020). Explainable recommendation: A survey and new perspectives. Found. Trends Inf. Retr., 14(1):1–101..

Explainability in Recommender Systems

Complementary information

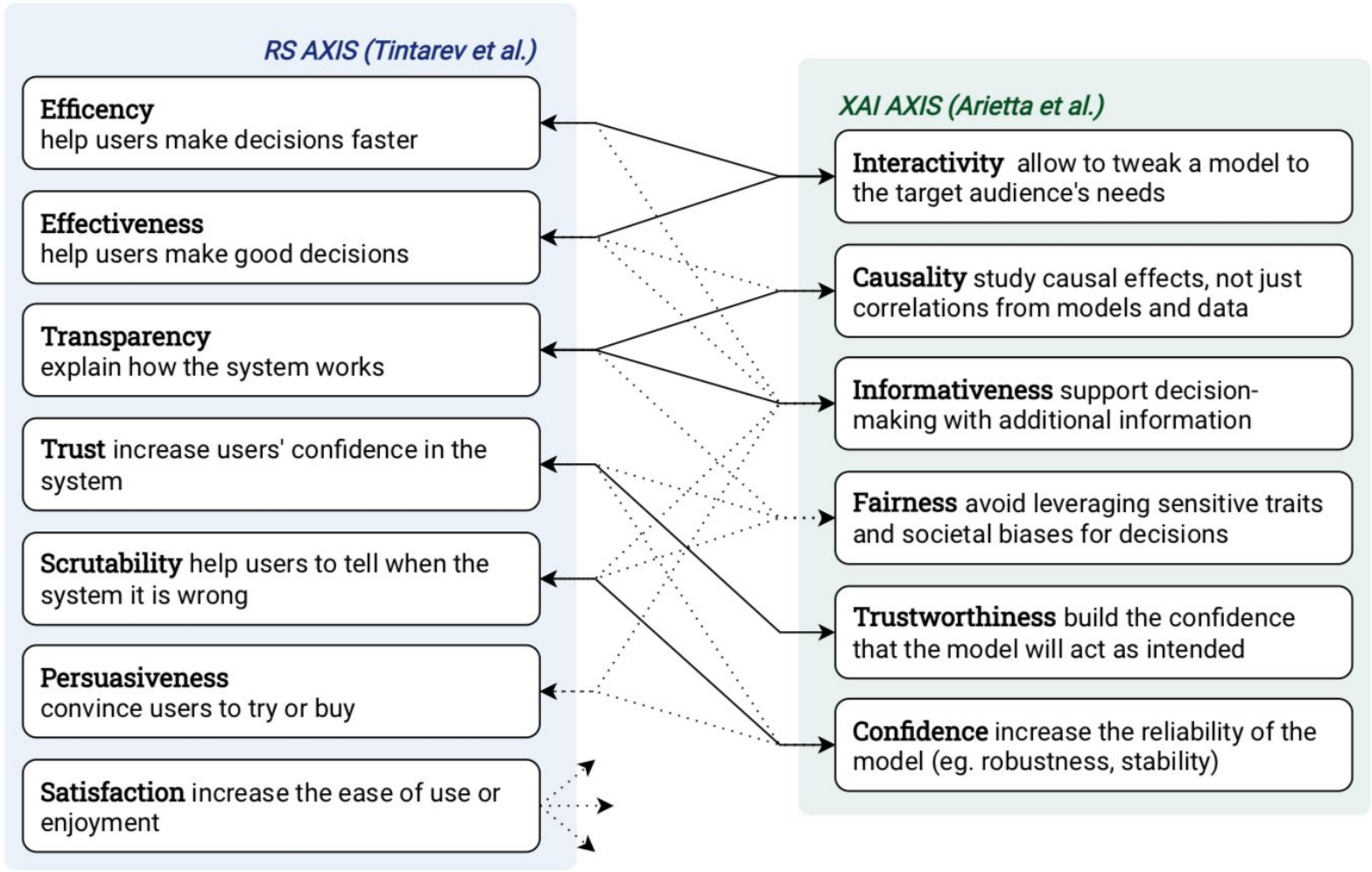
"To make clear by giving a **detailed description**" (Tintarev et al.)

"Explainable recommendation to answer the question of **why**" (Zhang et al.)

Helps ensure fairness regarding e.g. protected attributes. However: how to act upon them?

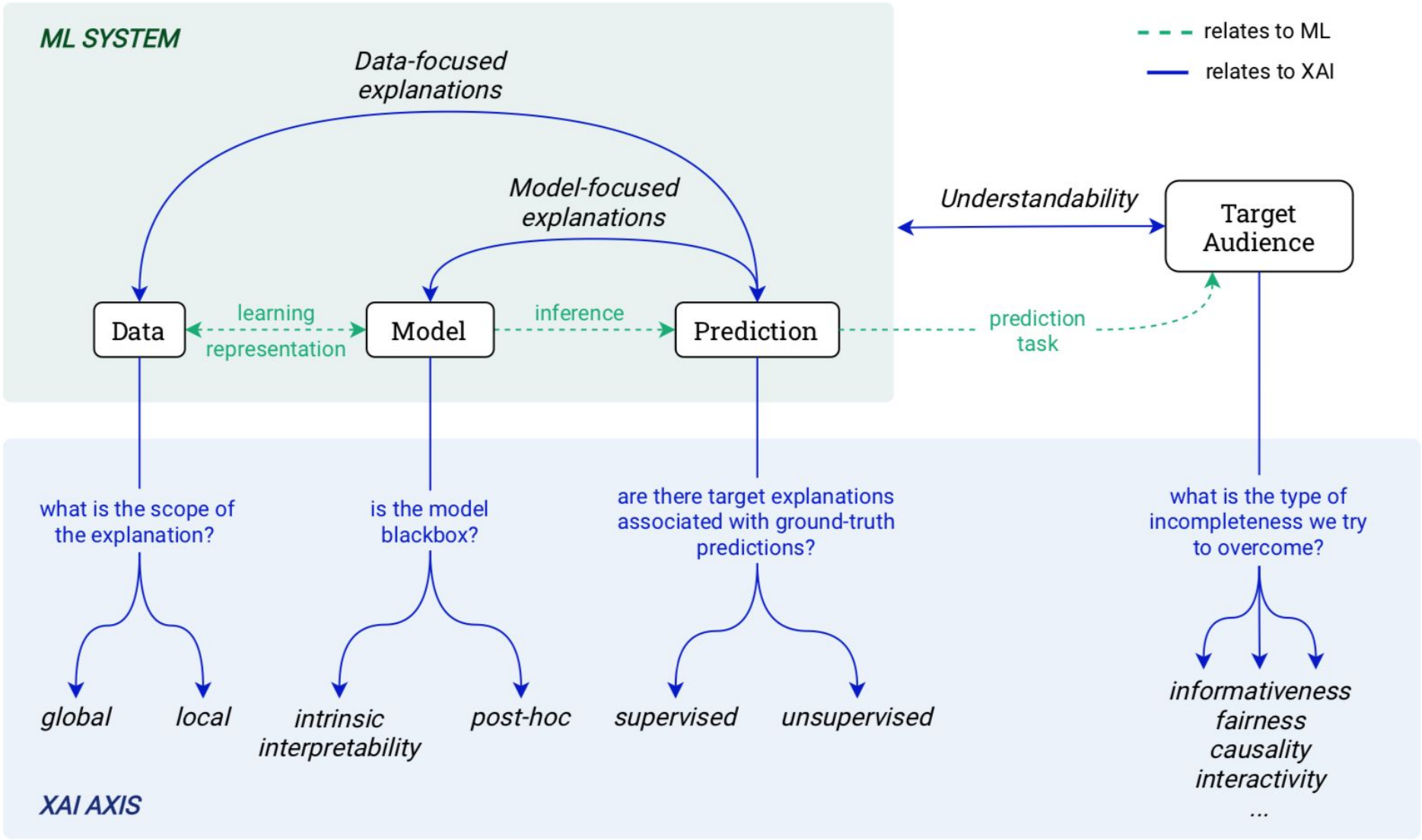
Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. In Recommender systems handbook (pp. 479–510). Springer, Boston, MA.
Zhang, Y. and Chen, X. (2020). Explainable recommendation: A survey and new perspectives. Found. Trends Inf. Retr., 14(1):1–101.

Explainability: Link to eXplainable AI (XAI)



Afchar, D., Melchiorre, A. B., Schedl, M., Hennequin, R., Epure, E. V., & Moussallam, M. (2022). Explainability in Music Recommender Systems. https://onlinelibrary.wiley.com/doi/full/10.1002/aaai.12056 & arXiv preprint arXiv:2201.10528.

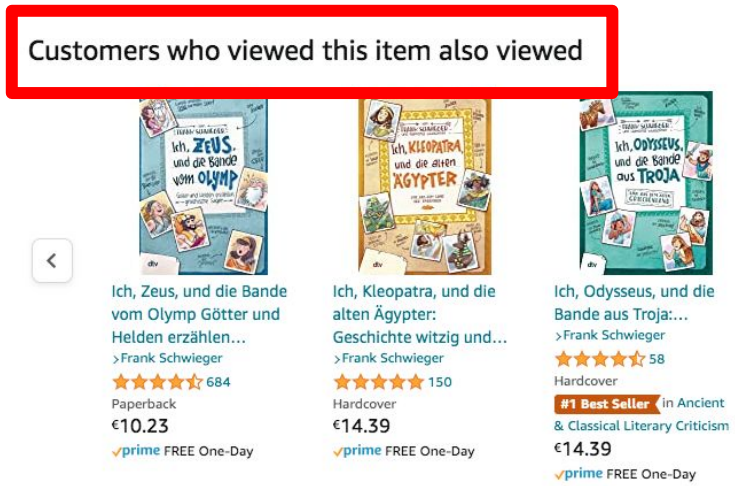
XAI Notions


ML SYSTEM
Data-focused explanations
Model-focused explanations
relates to ML
relates to XAI
Understandability
Target Audience
Data
learning representation
Model
inference
Prediction
prediction task
what is the scope of the explanation?
is the model blackbox?
are there target explanations associated with ground-truth predictions?
what is the type of incompleteness we try to overcome?
global
local
intrinsic interpretability
post-hoc
supervised
unsupervised
informativeness
fairness
causality
interactivity
...
XAI AXIS

Local vs. Global

Local: explain model decision for particular user-item pair

Explain single predictions



Global: explain model logic

Tells us about the average behavior of the model

Helps detect systematic biases of the model

Intrinsic vs. Post-hoc

Intrinsic: interpretability inherent in the model

“White-box models”

Ex.: item kNN model

“We recommend you <artist> because it is similar to <artist(s)>"

Post-hoc: apply external technique to create interpretability

Applied for black box models

“We recommend you <artist> because it has <features> that you might like"

Model vs. Data

Model: explaining learned model and parameters

Can lead to adjustments and regularization, e.g. to balance fairness and accuracy

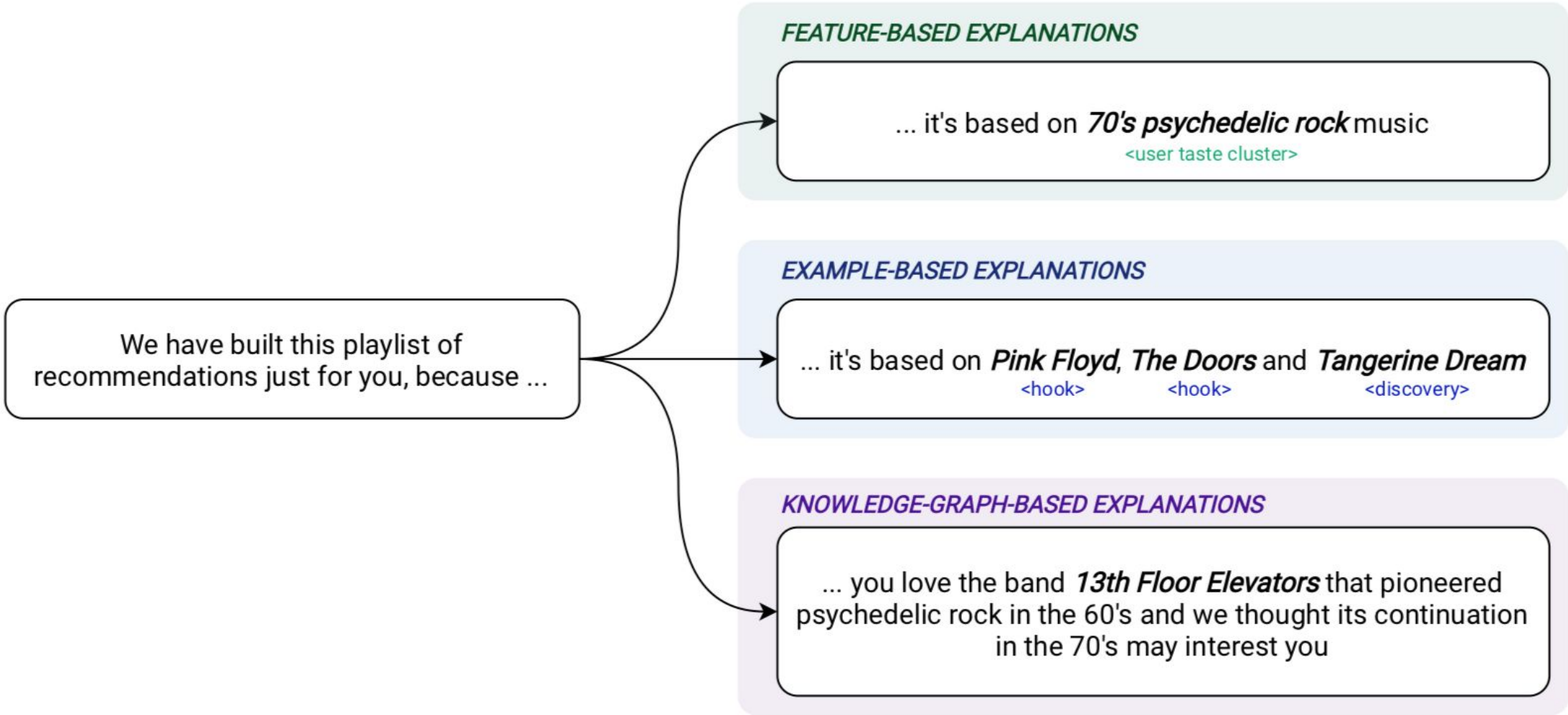
Data: explain data characteristics

Helps find irregularities in training data

“The has recommended you the item because it maximizes the probability of being co-listened with your history, considering all other users listening history"

“why are those items co-listened in the first place?”

Generating Explanations: Types



Selected Further Resources

- Afchar, D., Melchiorre, A. B., Schedl, M., Hennequin, R., Epure, E. V., & Moussallam, M. (2022). Explainability in Music Recommender Systems. https://onlinelibrary.wiley.com/doi/full/10.1002/aaai.12056 & arXiv preprint arXiv:2201.10528.
- Yongfeng Zhang and Xu Chen (2020), “Explainable Recommendation: A Survey and New Perspectives”, Foundations and Trends® in Information Retrieval: Vol. 14, No. 1, pp 1–101. DOI: 10.1561/1500000066.
- Tintarev, N., & Masthoff, J. (2022). Beyond explaining single item recommendations. In Recommender Systems Handbook(pp. 711-756). Springer, New York, NY.
- Zhang, Y., Zhang, Y., Zhang, M., & Shah, C. (2019, July). EARS 2019: The 2nd international workshop on explainable recommendation and search. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1438-1440).
- EARS tutorial: https://sites.google.com/view/ears-tutorial/

Outline

- Motivation & EU regulations
- Categories of Transparency
- Explainability
- **Traceability and Auditability**
- Documentation

Algorithm Auditing

- Area receives increased attention in various communities: CSCW, HCI, ML
- Aim: audit algorithms for biased, discriminatory, harmful behavior
  - alignment of systems with laws, regulations, ethics, …
- Inspired by audits in finance, security, employment,...
- Involves third part external experts:
  - researchers
  - developers
  - policymakers
- Helped uncover bias in search engines, housing, hiring, e-commerce → see https://arxiv.org/pdf/2105.02980.pdf for cases

Algorithm Auditing

Audit e-commerce sites for discrimination & price steering (Hannak et al., 2014)

- Web scraping + Amazon MTurk users as testers to audit e-commerce sites

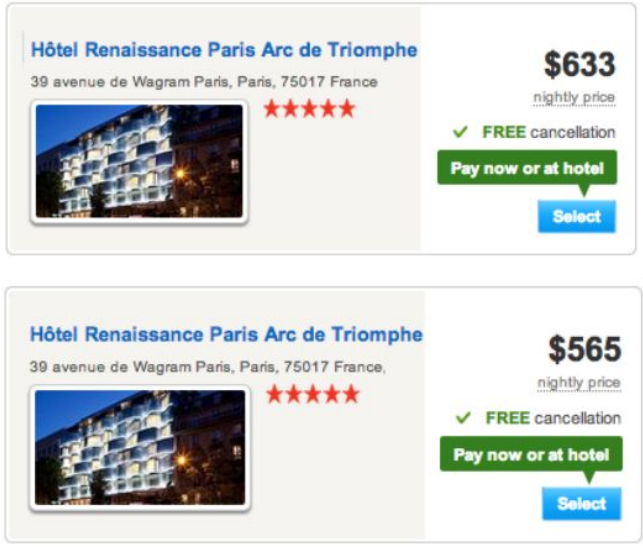


Figure 4: Example of price discrimination. The top result was served to the AMT user, while the bottom result was served to the comparison and control.

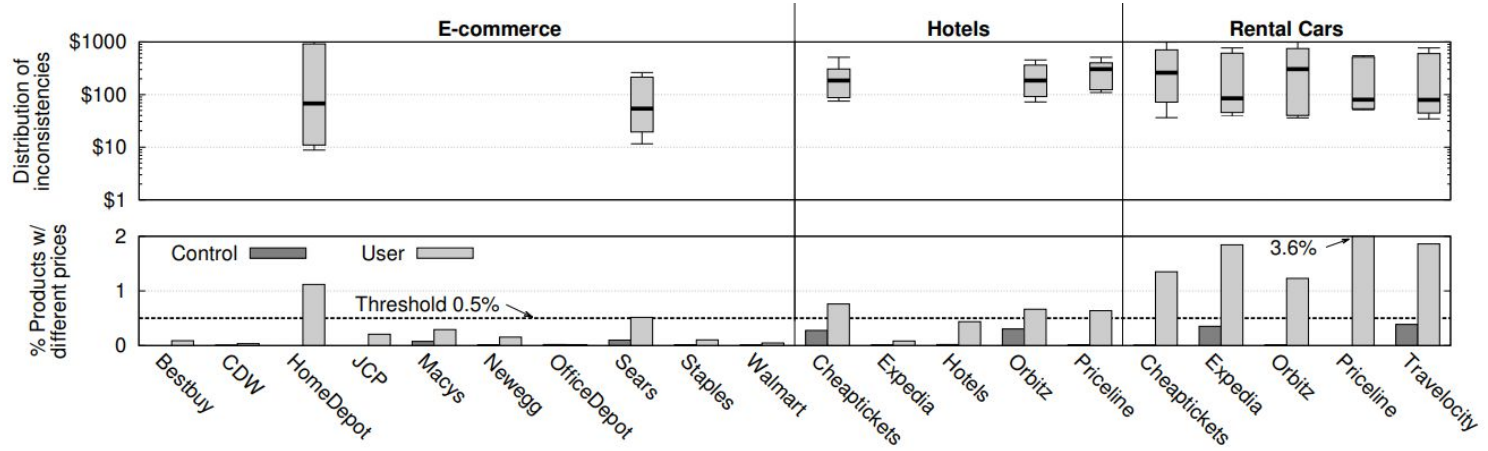


Figure 3: Percent of products with inconsistent prices (bottom), and the distribution of price differences for sites with ≥0.5% of products showing differences (top), across all users and searches for each web site. The top plot shows the mean (thick line), 25th and 75th percentile (box), and 5th and 95th percentile (whisker).

https://personalization.ccs.neu.edu

Hannak, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014, November). Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference* (pp. 305-318).

Types of Algorithm Auditing Methods

Taxonomy by Sandvig et al.:

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry (2014).

- Code audits
  - access to code and system design
- Noninvasive user audits
  - surveys
- Scraping audits
  - send repeated queries to test behavior of system under variety of conditions
- Sock puppet audits
  - researchers generate fake accounts to study system behavior for different user characteristics or patterns of behavior
- Crowdsourced/collaborative audits
  - researchers hire crowdworkers as testers

Types of Algorithm Auditing Methods

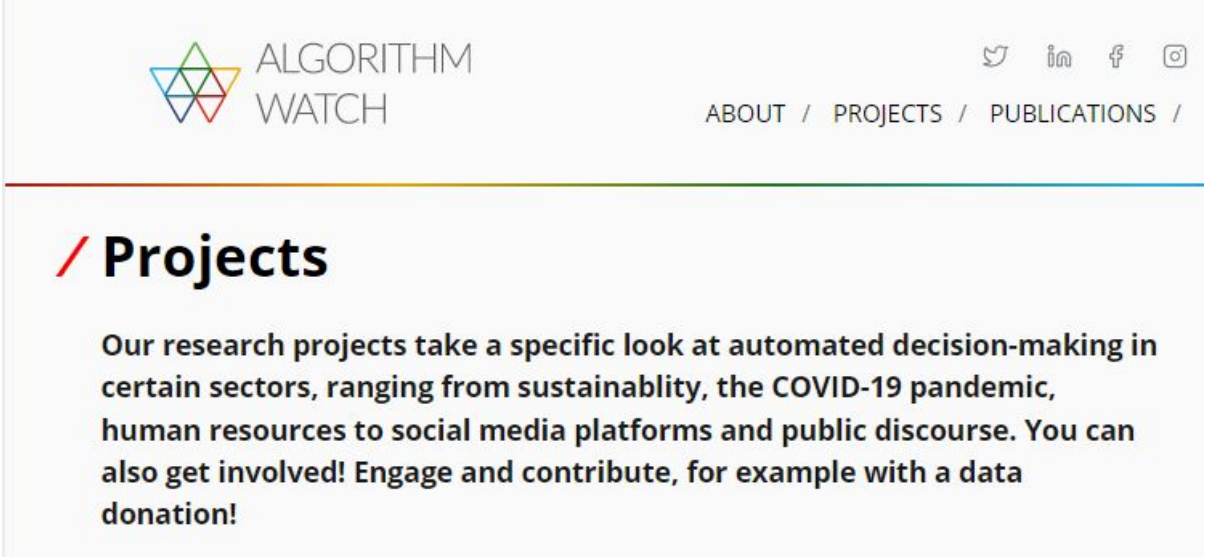
Taxonomy by Sandvig et al.:

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and Discrimination: Converting Critical Concerns into Productive Inquiry (2014).

- Code audits
  - access to code and system design
- Noninvasive user audits
  - surveys
- Scraping audits
  - send repeated queries to test behavior of system under variety of conditions
- Sock puppet audits
  - researchers generate fake accounts to study system behavior for different user characteristics or patterns of behavior
- Crowdsourced/collaborative audits
  - researchers hire crowdworkers as testers

Limits of Algorithm Auditing Methods

- Auditing requires technical expertise that might not always be available
  - Frequently: NGOs like AlgorithmWatch doing audits



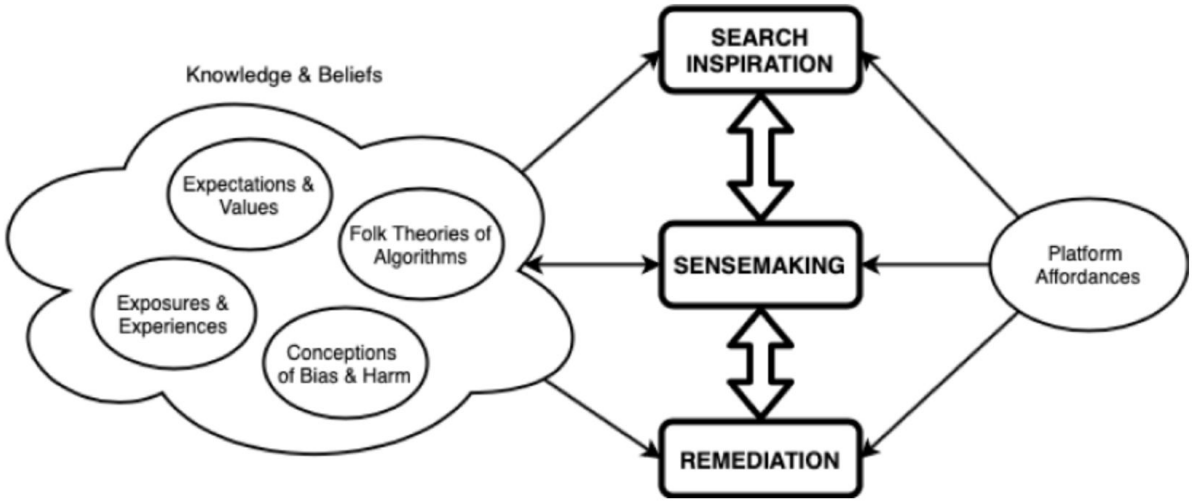
https://algorithmwatch.org

Limits of Algorithm Auditing Methods

- Many harmful algorithmic behaviors are hard to detect outside situated contexts
  - bias happens in specific social / cultural dynamics
  - challenging to anticipate real-world contexts
- Crowdworkers may not represent demographics of investigated system
  - biases might still be undetected
- Expert-driven audits might miss harmful behavior!

Everyday Algorithm Auditing

- Idea: everyday users detect problematic system behavior via day-to-day interactions with system
- Recent work looked at what strategies users apply in such user-driven audits



DeVos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022, April). Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Examples: Everyday Algorithm Auditing

https://arxiv.org/pdf/2105.02980.pdf

Domains	Cases	Descriptions
Search	Google Image Search [65]	Researcher Noble searched "black girls" on Google and found out the results were primarily associated with pornography.
Rating/review	**Yelp advertising bias [29]**	Many small business owners on Yelp came together to investigate Yelp's potential bias against businesses that do not advertise with Yelp.
	Booking.com quality bias [28]	A group of users on Booking.com scrutinized its rating algorithm after realizing the ratings appeared mismatched with their expectations.
Recommendation systems	YouTube LGBTQ+ demonetization [73]	A group of YouTubers found that the YouTube recommendation algorithm demonetizes LGBTQ+ content, resulting in a huge loss of advertising revenue for LGBTQ+ content creators.
	Google Maps [34]	A group of users reported that when they searched for the N-word on Google Maps, it directed them to the Capitol building, the White House, and Howard University, a historically Black institution. Other users joined the effort and uncovered other errors.
	TikTok recommendation algorithm [54, 82]	A group of users found that TikTok's "For You Page" algorithm suppresses content created by people of certain social identities, including LGBTQ+ users and people of color. As a result, they worked together to amplify the suppressed content.

Outline

- Motivation & EU regulations
- Categories of Transparency
- Explainability
- Traceability and Auditability
- **Documentation**

Datasheets for Datasets

- Aim: transparency on datasets used to train and evaluate ML models
  - dataset creation process, possible sources of bias
- Questions: motivation, composition, collection, pre-processing, labeling, intended uses, distribution, and maintenance.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Datasheets for Datasets

This template contains a set of questions covering the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions.

The questions are grouped into seven sections that roughly match the key stages of the dataset creation, maintenance, and distribution process. By grouping the questions in this way, we encourage dataset creators to reflect on the process of creating, distributing, and maintaining datasets, and even to modify this process in response to that reflection. We recommend that dataset creators read through the questions in all sections prior to any data collection so as to flag potential issues early on, and then provide answers to the questions in each section during the relevant stage of the process.

We emphasize that the questions are intended to be used as a starting point for dataset creators to customize. Not all questions will be applicable to all datasets, and dataset creators will likely need to add, revise, or remove questions to better fit their specific circumstances and needs.

To prompt dataset creators to provide sufficient information, all questions are worded so as to discourage yes/no answers. The questions are not intended to serve as a checklist, and dataset creators must be as transparent and forthcoming as possible for datasheets to be useful to dataset consumers.

Questions

Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
- **Any other comments?**

Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
- **How many instances are there in total (of each type, if appropriate)?**

For more information about these questions and about datasheets for datasets in general, please see T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. *Datasheets for Datasets.* The latest version of this paper can be found online at https://arxiv.org/abs/1803.09010

Model Cards

- Aim: transparent model reporting
  - performance characteristics of trained ML model
- Idea: release model cards in addition to datasets
- Contains:
  - model details, intended use, metrics, training data, evaluation data, ethical considerations

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

slido

Have you used datasheets / model cards in your work or have you created such documentation?

ⓘ Start presenting to display the poll results on this slide.

slido

Audience Q&A Session

ⓘ Start presenting to display the audience questions on this slide.

Part 5:
Open Challenges

Open Challenges (Bias and Fairness)

- Which technological foundation do we need to debias data and algorithms in state-of-the-art ranking systems, such as IR and RS?
- How should requirements and aims of various stakeholders (e.g., content creators and consumers, platform providers, policymakers) be accounted for?
- Do computational bias metrics really capture how users perceive fairness?
- What are economic and social consequences of biases resulting from IR and RS technology adopted in high-risk areas (e.g., in recruitment, healthcare)?
- What are the legal implications of unfair or intransparent algorithms?

Open Challenges (Transparency)

- What level of transparency is useful for the needs of different stakeholders and how can transparency be adjusted depending on varying needs?
- What is the relation between explanations and perceived fairness?
- What are effective explanation types for different retrieval and recommendation domains?
- What do explanations tell us about the user? What ethical and privacy implications can arise?

Open Challenges (Social Impact)

- How to collectively set targets and indicators for social impact, e.g. diversity, impact on jobs, environment?
- Which methodologies should be put in place to assess the short-term and long-term social impact of algorithms, to be able to maximize opportunities while avoiding risks?
- Which data may researchers need from real-world scenarios to carry out evaluations on different aspects, e.g. fairness, transparency or social and environmental impact?

slido

Which are the most important open challenges that research should address, in your opinion?

ⓘ Start presenting to display the poll results on this slide.

Thank You!

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

Emilia Gómez
Joint Research Centre, European Commission, Spain
Universitat Pompeu Fabra, Spain
emilia.gomez-gutierrez@ec.europa.eu | https://emiliagomez.com

Elisabeth Lex
Graz University of Technology, Austria
elisabeth.lex@tugraz.at | https://elisabethlex.info