THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Bias in Humans and AI – What To Do About It?

Prof. Gianluca Demartini
Data Science Discipline
School of Electrical Engineering and Computer Science

Research Interests

- **Information Access** (since 2005)
  Structured/Unstructured data (SIGIR12), Entity Types (ISWC13, WSemJ16)
  Entity Recognition (WWW14), Prepositions (CIKM14), Entity Cards (SIGIR19)
  Evaluation (ECIR16 Best P, CIKM17, SIGIR18, CIKM19, WWW22, TOIS23, ICTIR23 Best P)
- **Human-AI Systems** (since 2012)
  Entity Linking (WWW12,VLDBJ), CrowdQ (CIDR13), Learnersourcing (LAK21,LAK22,JCAL)
  LLM (COLING25, **CHI25**), Misinfo (ECIR20 Best SP, SIGIR20, CIKM20, IP&M, ICWSM24)
- **Better Crowdsourcing Platforms** (since 2013)
  Platforms (WWW15, CSCWJ18, **CACM25**), Experiments (CSCW21), Pricing (HCOMP14)
  Task Allocation (WWW13, WWW16, COR), Workers (CHI15, CSCW20 Hon. Mention)
  Metadata (IP&M), Attacks (HCOMP18 Best P, JAIR), Time (HCOMP16)
  Modus Operandi (UBICOMP17, HT19, WSDM20, TOIS24), Complexity (HCOMP16)
  Abandonment (WSDM19, TKDE, ACM TSC)
- **Data Bias** (since 2018)
  Gender (w/ Wiki; SIGIR18, ACIS24, **WWW25**), Management (**CACM24**, **WWW25**),
  Impact on ML (CIKM22), SES (**WebSci22**, **ICWSM25**), Political (**WWW25**)
- **Better Data** (since 2019)
  Noise (WWW19, **ICWSM25**), Data Workers (SIGIR20, TOIS, TKDE, WWW23)
  Know. Graphs (ISWC19), Unknown Unknowns (ECAI20, HCOMP21), Behaviors (CIKM20)
  Fairness (CIKM22, SIGIR23, FAccT24, KDD24), Active Learning (AAAI24)

Thanks to:

Australian Government
Australian Research Council

European Commission

Swiss National Science Foundation

EPSRC
Engineering and Physical Sciences Research Council

Meta

WIKIMEDIA FOUNDATION

Google AI

Outline

Bias in Humans – Gender Bias

- Gender completeness in Wikipedia (ACIS 2024)
- Gender exploration in Wikipedia (ACM TheWebConf 2025)

Bias in LLMs – SES and Political Bias

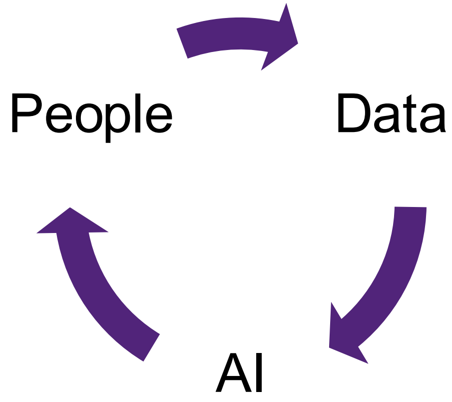
- SES bias in humans and ML (WebSci 2023; ICWSM 2025)
- Persona-based LLMs (ACM TheWebConf 2025)

Exploring Data Bias

- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)
- LLM-based Data Pipelines (ICWSM 2025)

Human-AI Collaboration and Trust

- LLM Agents to complete tasks for us (ACM CHI 2025)
- Provoking Critical Thinking (IP&M 2025)



How is gender represented in Wikipedia articles?

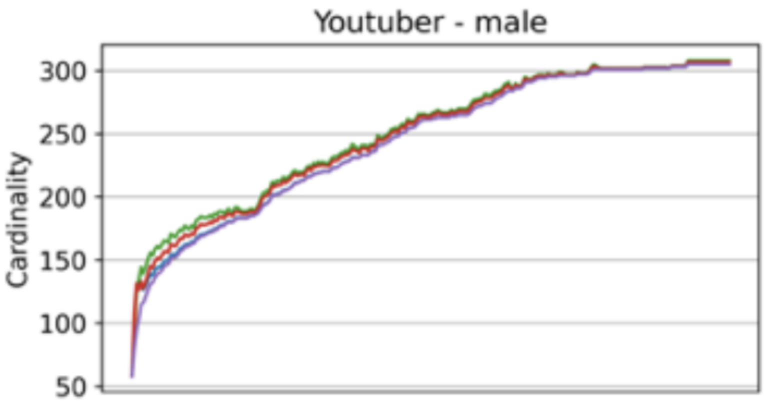
Using an automatic gender classifier based exclusively on the persons' name

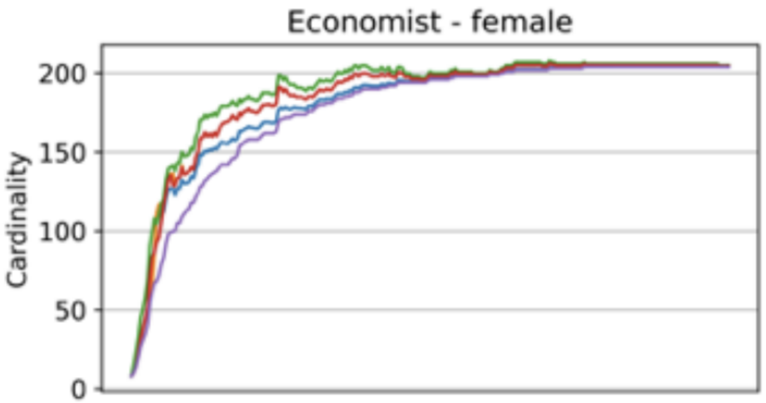
subclass	size	female	male	undefined
Coach	9975	4.21 %	81.36 %	14.43 %
Academic	9934	33.28 %	55.52 %	11.20 %
Artist	9907	20.31 %	63.34 %	16.35 %
Scientist	9880	17.14 %	75.12 %	7.74 %
MilitaryPerson	9816	2.04 %	87.87 %	10.10 %
Writer	9723	27.61 %	60.93 %	11.46 %
Politician	9578	15.88 %	68.99 %	15.13 %
Royalty	8841	27.89 %	41.38 %	30.73 %
Athlete	7979	16.47 %	64.41 %	19.13 %
Noble	7949	18.87 %	63.45 %	17.68 %
SportsManager	6324	1.36 %	89.34 %	9.30 %
Architect	5574	6.30 %	76.80 %	16.90 %
Religious	4832	7.99 %	66.02 %	25.99 %
Philosopher	2987	11.72 %	76.13 %	12.15 %
Model	2045	69.10 %	10.86 %	20.05 %
Journalist	1858	24.49 %	59.04 %	16.47 %
Economist	1720	11.86 %	79.36 %	8.78 %
Youtuber	900	14.33 %	34.56 %	51.11 %
Chef	897	24.86 %	59.31 %	15.83 %
Engineer	885	2.82 %	89.72 %	7.46 %
Astronaut	738	10.57 %	75.88 %	13.55 %
BusinessPerson	691	13.17 %	73.23 %	13.60 %
PoliceOfficer	413	5.33 %	72.88 %	21.79 %
HorseTrainer	355	2.25 %	73.52 %	24.23 %
Pilot	286	18.53 %	65.73 %	15.73 %
AmericanLeader	264	17.80 %	41.67 %	40.53 %

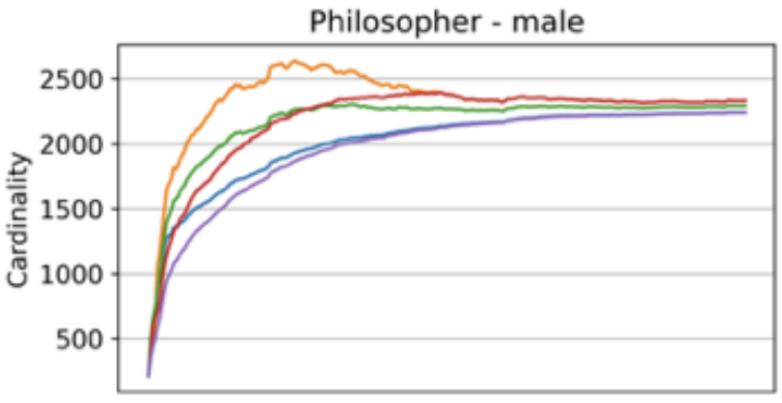
Hrishi Patel, Tianwa Chen, Ivano Bongiovanni, and Gianluca Demartini. **Estimating Gender Completeness in Wikipedia**. In: The Australasian Conference on Information Systems (ACIS 2024). Canberra, Australia, December 2024.

Estimating gender cardinality (and completeness)

Using statistical estimators based on how often articles are edited







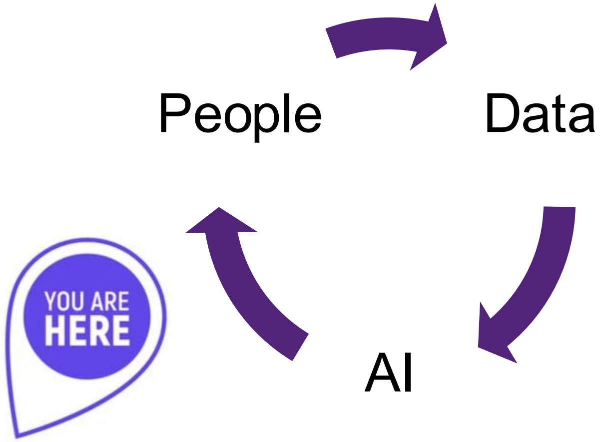
Estimating the completeness of each class/gender

Subclass	Gender	Entities	Est. (N1)	Conv. (N1)	% Compl. (N1)	Est. (J1)	Conv. (J1)	% Compl. (J1)
Academic	female	3300	3305	0.001000	99.850000	3384	0.009900	97.520000
Academic	male	5497	5505	0.001500	99.850000	5626	0.016300	97.710000
AmericanLeader	female	47	47	0.000000	100.000000	50	0.000000	94.000000
AmericanLeader	male	109	109	0.000700	100.000000	116	0.009100	93.970000
Architect	female	339	341	0.004300	99.410000	370	0.039700	91.620000
Architect	male	3946	3977	0.007800	99.220000	4274	0.064600	92.330000
Artist	female	2003	2006	0.001300	99.850000	2054	0.019300	97.520000
Artist	male	6198	6207	0.001500	99.860000	6386	0.024800	97.060000
Astronaut	female	76	76	0.000000	100.000000	81	0.000000	93.830000
Astronaut	male	554	555	0.001400	99.820000	583	0.030600	95.030000
Athlete	female	1297	1303	0.003900	99.540000	1369	0.040700	94.740000
Athlete	male	5037	5063	0.005200	99.490000	5417	0.064400	92.990000
BusinessPerson	female	86	88	0.009000	97.730000	96	0.057900	89.580000
BusinessPerson	male	483	491	0.017500	98.370000	533	0.097000	90.620000
Chef	female	220	221	0.000400	99.550000	229	0.005700	96.070000
Chef	male	527	528	0.001300	99.810000	541	0.020800	97.410000
Coach	female	415	416	0.001900	99.760000	435	0.024000	95.400000
Coach	male	8048	8073	0.003100	99.690000	8457	0.035300	95.160000
Economist	female	204	204	0.000400	100.000000	208	0.004900	98.080000
Economist	male	1353	1355	0.001700	99.850000	1423	0.021700	95.080000
Engineer	female	25	25	0.000000	100.000000	32	0.000000	78.120000
Engineer	male	785	790	0.005600	99.370000	871	0.047900	90.130000

Exploration of Gender in Wikipedia

Yahya Yunus, Tianwa Chen and Gianluca Demartini. **Exploring Wikipedia Gender Diversity Over Time - The Wikipedia Gender Dashboard (WGD)**. In: The 2025 ACM Web Conference (TheWebConf 2025) - Demo track. Sydney, Australia, April 2025.

Outline



Bias in Humans – Gender Bias

- Gender completeness in Wikipedia (ACIS 2024)
- Gender exploration in Wikipedia (ACM TheWebConf 2025)

Bias in LLMs – SES and Political Bias

- SES bias in humans and ML (WebSci 2023; ICWSM 2025)
- Persona-based LLMs (ACM TheWebConf 2025)

Exploring Data Bias

- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)
- LLM-based Data Pipelines (ICWSM 2025)

Human-AI Collaboration and Trust

- LLM Agents to complete tasks for us (ACM CHI 2025)
- Provoking Critical Thinking (IP&M 2025)

Video of people washing hands across different socio-economic statuses

- 4 regions: Africa, Asia, Europe, the Americas; 4 different income level for each region (4*4*7=112)
- Average video duration：13.7 seconds (SD = 9.14 seconds)

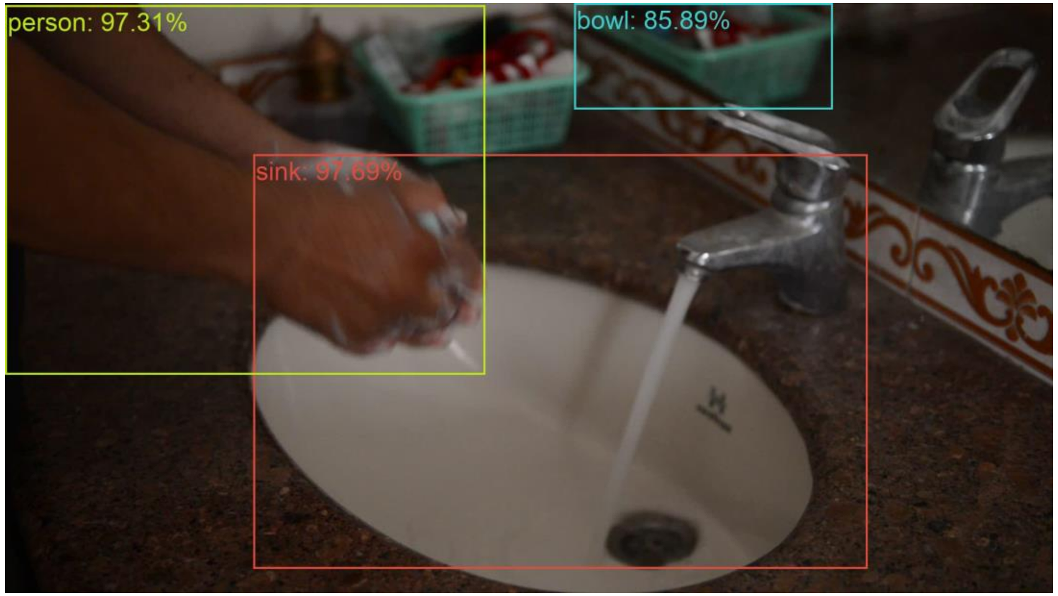
Bias in the annotation of SES-diverse content

- **Less accurate** in guessing families' income levels for **African videos**.
- Videos depicting **low-income** households were more likely to receive **negative** annotations
- Videos with **higher-income** families received more **positive** annotations.
- **Negative** annotations were more prevalent for videos shot in **Africa** than in **Asia**.
- Video from **higher income** groups **more appropriate** for inclusion in search results and public service announcements

Bias: Being used to see high-SES content on social media means that SES-diverse content gets critical views (confirmation bias)

Shaoyang Fan, Pınar Barlas, Evgenia Christoforou, Jahna Otterbacher, Shazia Sadiq, and Gianluca Demartini. **Socio-Economic Diversity in Human Annotations**. In: The 14th ACM Web Science Conference 2022, Barcelona, Spain. June 2022.

Human vs ML annotations



AI can label images too! We do not need humans

Nardiena A. Pratama, Shaoyang Fan, and Gianluca Demartini. **Perception of Visual Content: Differences between Humans and Foundation Models**. In: 19th International AAAI Conference on Web and Social Media (ICWSM 2025). Copenhagen, Denmark, June 2025.

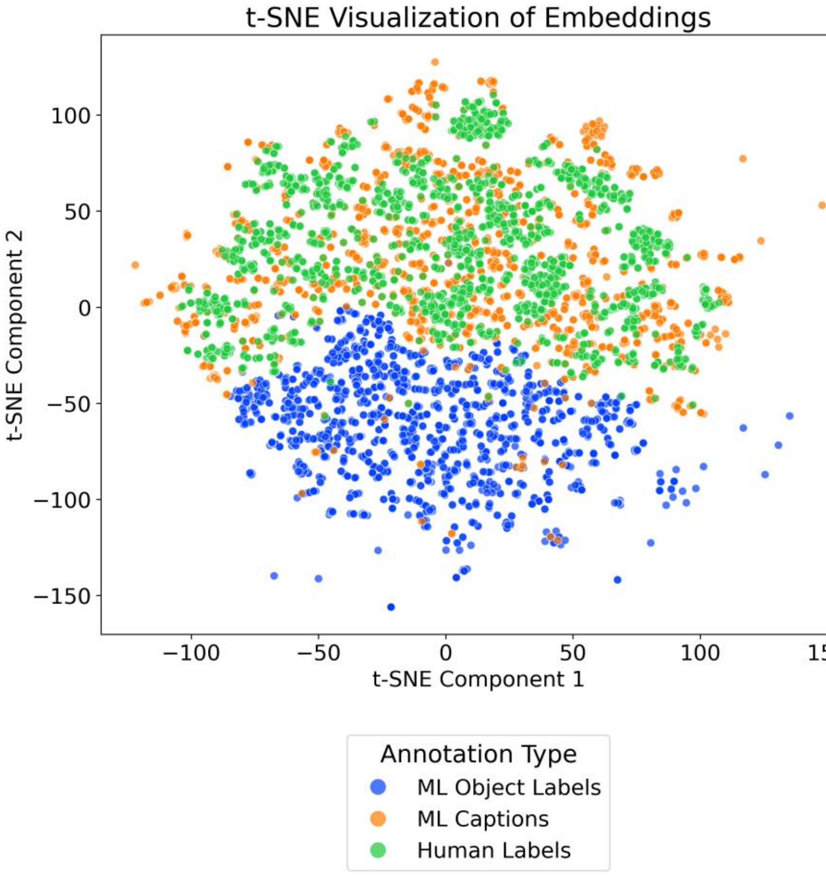
Research Questions

RQ1 How similar are human-generated and ML-generated annotations?

- Consistent similarity and dissimilarity of annotations across regions implies that **their level of bias is comparable**

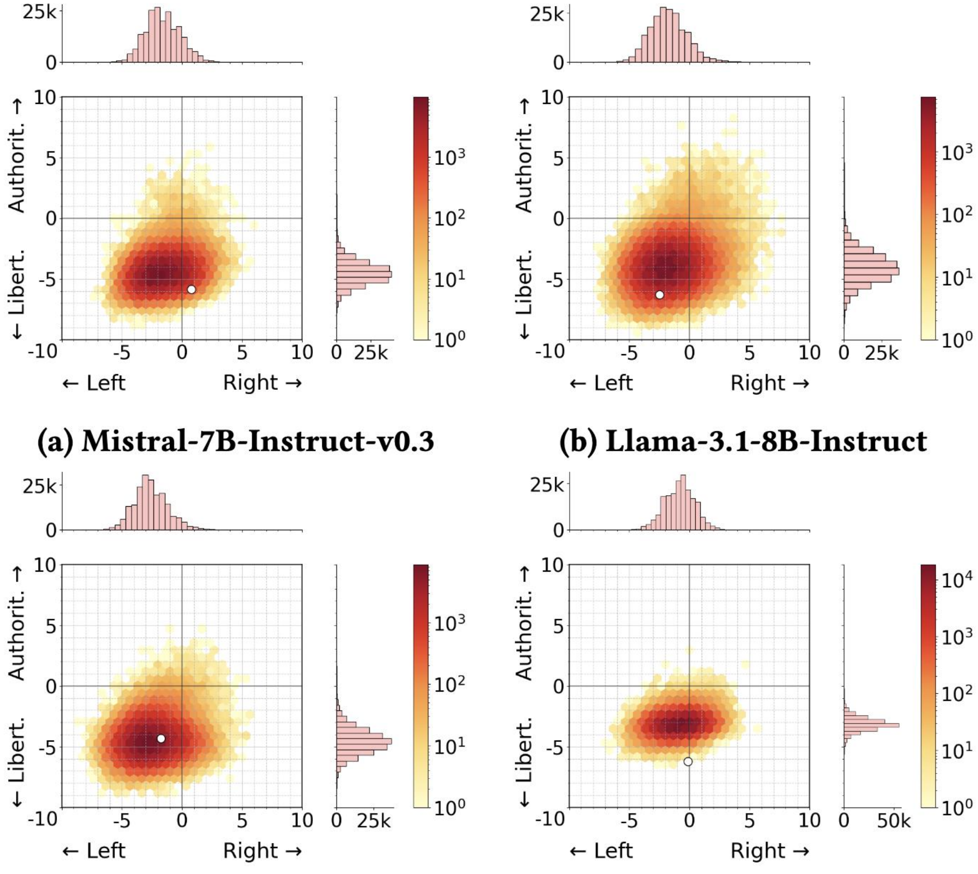
RQ2 How do different combinations of annotations affect fairness in ML predictive models?”

- Certain annotation types (human vs machine) work better for certain geographical areas and income levels



All annotations are important, and machine-generated annotations **cannot just replace human-generated ones**

Persona-based LLMs



(a) Mistral-7B-Instruct-v0.3 (b) Llama-3.1-8B-Instruct

(c) Qwen2.5-7B-Instruct (d) Zephyr-7b-beta

- We make LLMs answer the Political Compass Test

- We then make them impersonate 200,000 personas and answer the PCT again

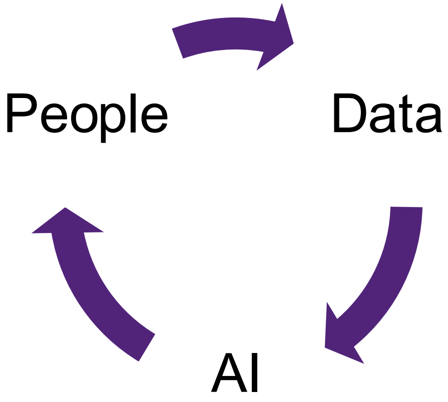
- This shows how we can measure and control the political bias of LLMs.

- It also highlights embedded stereotypes like

"A business developer trying to bring new investments to the region, regardless of environmental cost" being authoritarian right

Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. **Mapping and Influencing the Political Ideology of Large Language Models using Synthetic Personas**. In: The 2025 ACM Web Conference (TheWebConf 2025) - Short paper track. Sydney, Australia, April 2025.

Outline



Bias in Humans – Gender Bias

- Gender completeness in Wikipedia (ACIS 2024)
- Gender exploration in Wikipedia (ACM TheWebConf 2025)

Bias in LLMs – SES and Political Bias

- SES bias in humans and ML (WebSci 2023; ICWSM 2025)
- Persona-based LLMs (ACM TheWebConf 2025)

Exploring Data Bias

- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)
- LLM-based Data Pipelines (ICWSM 2025)

Human-AI Collaboration and Trust

- LLM Agents to complete tasks for us (ACM CHI 2025)
- Provoking Critical Thinking (IP&M 2025)

Bias Management, not bias removal

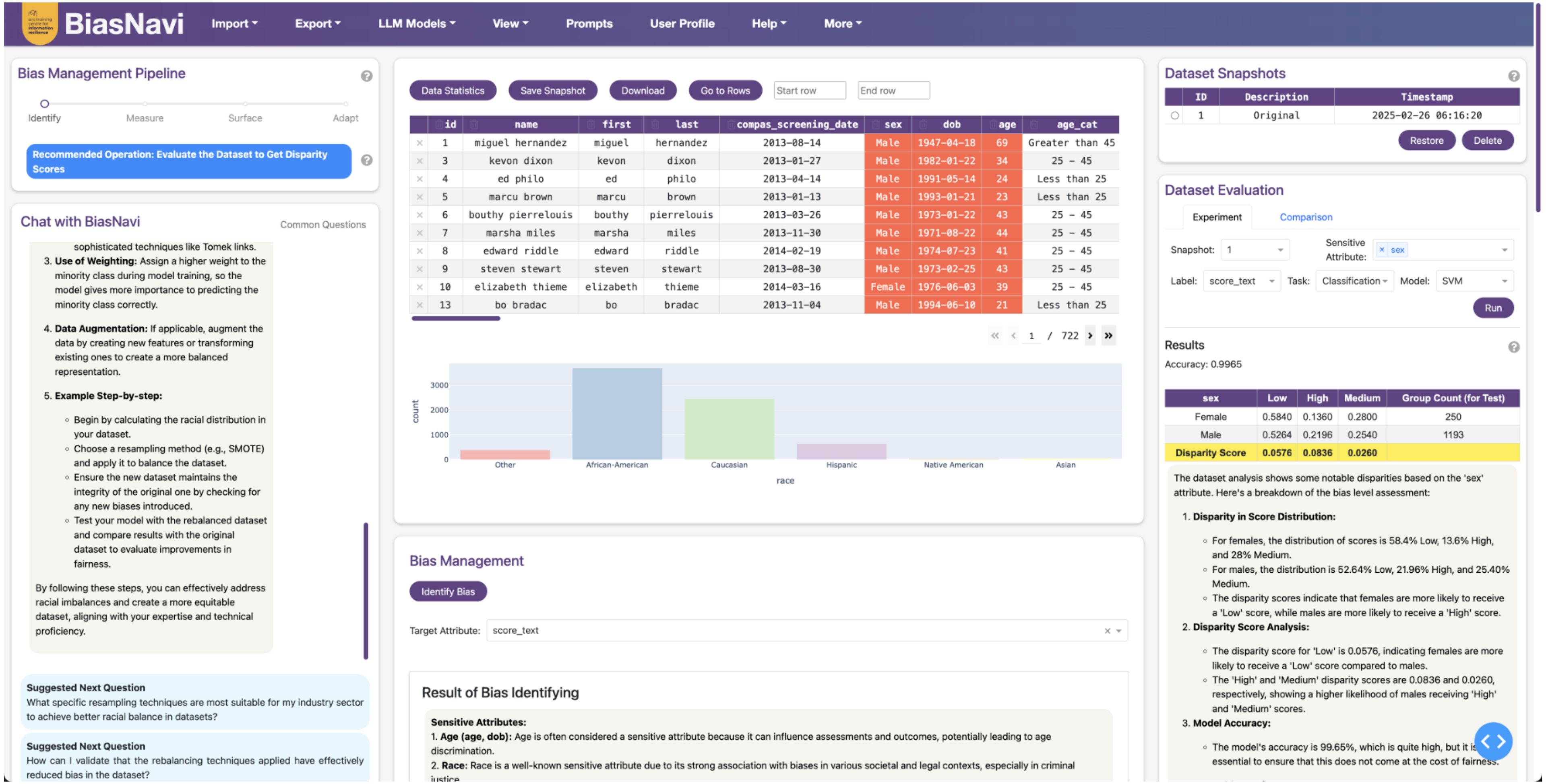
Employing an explicit and not transparent bias removal intervention might be potentially harmful to the user

Figure 2. The five steps of bias management.

Demartini, Roitero, Mizzaro. “**Data Bias Management**”, in *Communications of the ACM, Vol. 67, No. 1, Jan 2024*

BiasNavi

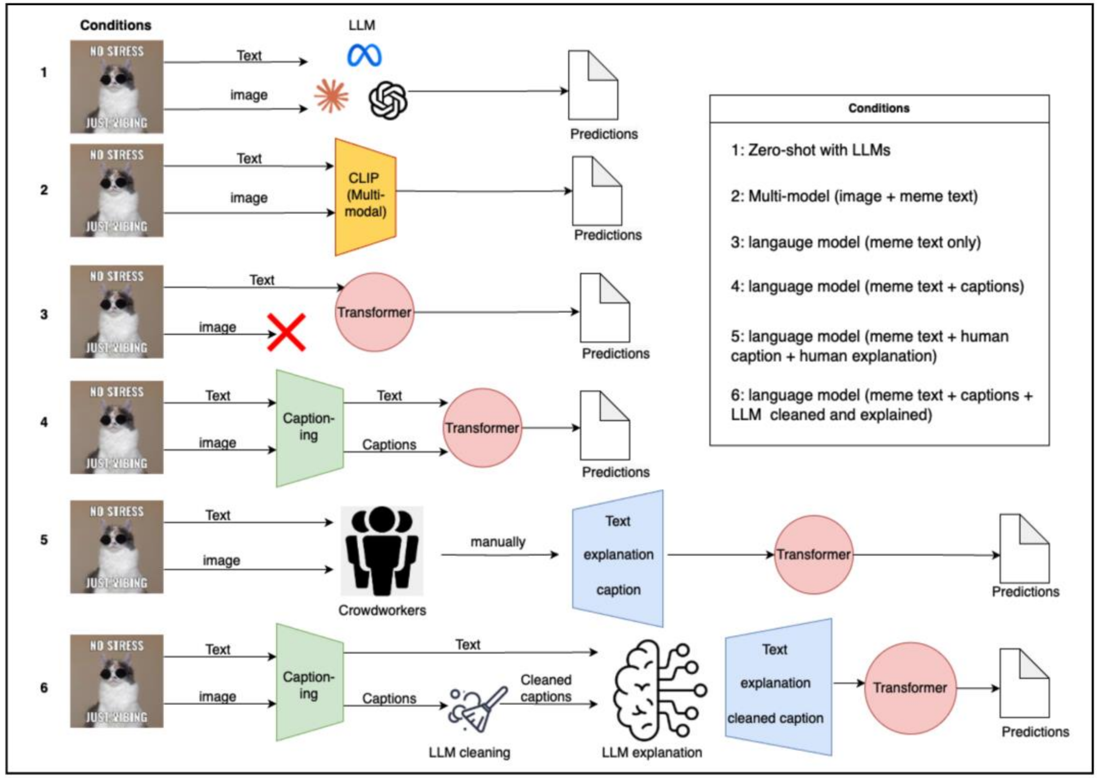
https://github.com/CIRES-Hub/BiasNavi/



Junliang Yu, Jay Thai Duong Huynh, Shaoyang Fan, Gianluca Demartini, Tong Chen, Hongzhi Yin, and Shazia Sadiq. **BiasNavi: LLM-Empowered Data Bias Management**. In: The 2025 ACM Web Conference (TheWebConf 2025) - Demo track. Sydney, Australia, April 2025

LLM-based Data Pipelines to Detect Harmful Content

Facebook Hateful Meme Challenge: classify content as hateful or non-hateful



1. Zero-Shot with Meme Image and Text
2. Image + Text (Multimodal Model; CLIP)
3. Meme Text Only (Language Model; DistilBERT, RoBERTa)
4. Text + Captions (Language Model)
5. Human Captions and Explanations (Language Model)
6. Meme Text + Cleaned Captions + LLM Explanations (Language Model):

Elyas Meguellati, Assaad Zeghina, Shazia Sadiq, and Gianluca Demartini. **LLM-based Semantic Augmentation for Harmful Content Detection**. In: 19th International AAAI Conference on Web and Social Media (ICWSM 2025). Copenhagen, Denmark, June 2025.

Findings

- **Does LLM-Based Caption Cleaning Work? (RQ1a)**
  - GPT-4o-cleaned captions showed significant improvements over the uncleaned captions for the classifier ($p = 0.0157$)
- **Does Adding Context Improve Performance? (RQ1b)**
  - Leveraging LLMs to augment each meme with a short, explanatory context yields performance gains
  - Including meme text, caption and LLM-generated explanation yields strongest performance
- **Generalizability Across Related Domains (RQ2)**
  - The approach generalizes well across social media tasks (Jigsaw Toxic Comments and Facebook Hateful Memes) with differing data modalities (text vs multimodal)

A fundamental distinction between LLM explanatory capabilities and predictive performance

Observations

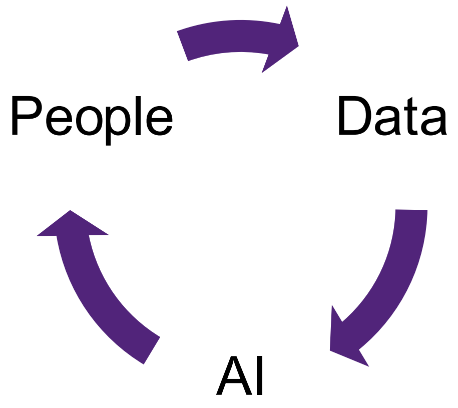
- LLM are not good harmful content detector if used as zero-shot classifiers
- LLM are good at segmenting, explaining, and providing more context for downstream harmful content classification
- LLM-based semantic augmentation is effective for context-dependent tasks
  - Reduced manual annotation costs

- **Safeguard mechanisms** embedded in LLMs limit performance on harmful content
  - Managing LLM safeguards by asking to preserve triggers
  - Important in domains where capturing explicit terms

or themes is critical for model training

Explanation: "This meme implies that interacting with anything connected to Islam is dangerous or undesirable, feeding into a narrative that paints Muslims as inherently threatening or alien. By humorously suggesting that no one dares pull the doll's string, the meme mocks and perpetuates fears of Islam."
Triggers: Islamophobia, Stereotyping, Muslim doll, what the fuck, no one has the guts.

Outline



Bias in Humans – Gender Bias

- Gender completeness in Wikipedia (ACIS 2024)
- Gender exploration in Wikipedia (ACM TheWebConf 2025)

Bias in LLMs – SES and Political Bias

- SES bias in humans and ML (WebSci 2023; ICWSM 2025)
- Persona-based LLMs (ACM TheWebConf 2025)

Exploring Data Bias

- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)
- LLM-based Data Pipelines (ICWSM 2025)

Human-AI Collaboration and Trust

- LLM Agents to complete tasks for us (ACM CHI 2025)
- Provoking Critical Thinking (IP&M 2025)

Trust in AI Agent

Agents as daily assistants
Tasks with different levels of risk
LLM agents used in a plan-then-execute manner

A double-edged sword
(1) they can work well with a high-quality plan and necessary user involvement
(2) users can easily mistrust the LLM agents with plans that seem plausible

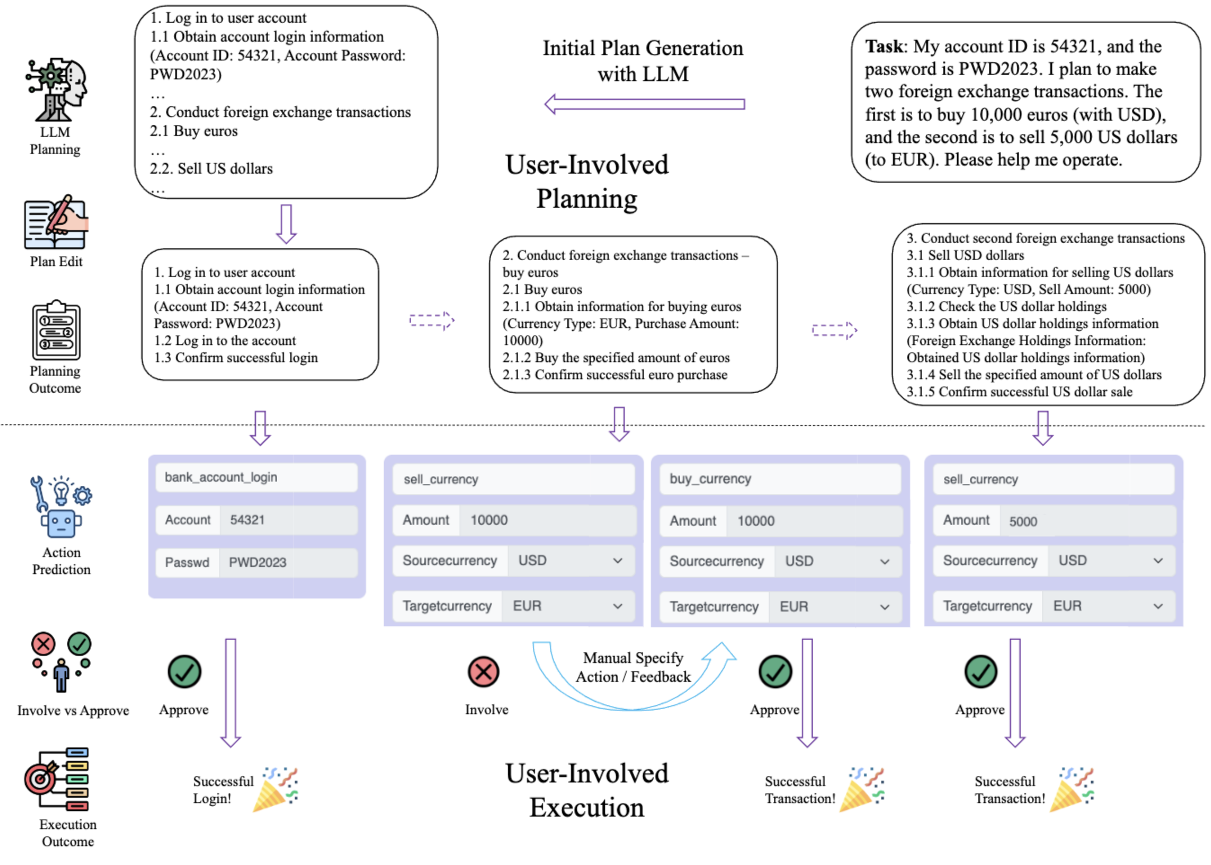


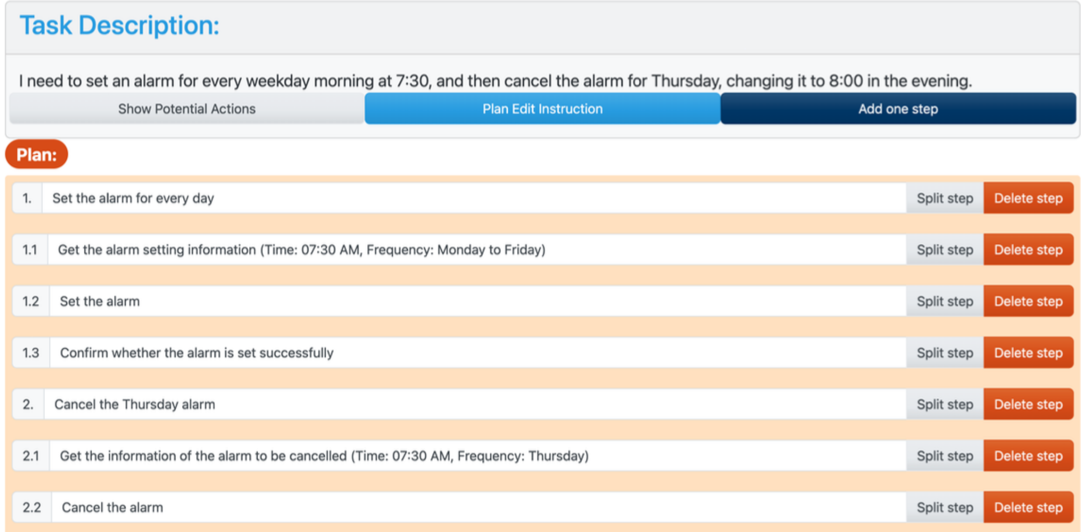
Figure 1: Illustration of the human-AI collaboration with plan-then-execute LLM agents.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. **Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant**. In: ACM CHI 2025 Conference on Human Factors in Computing Systems (CHI 2025). Yokohama, Japan, April 2025.

Key Findings

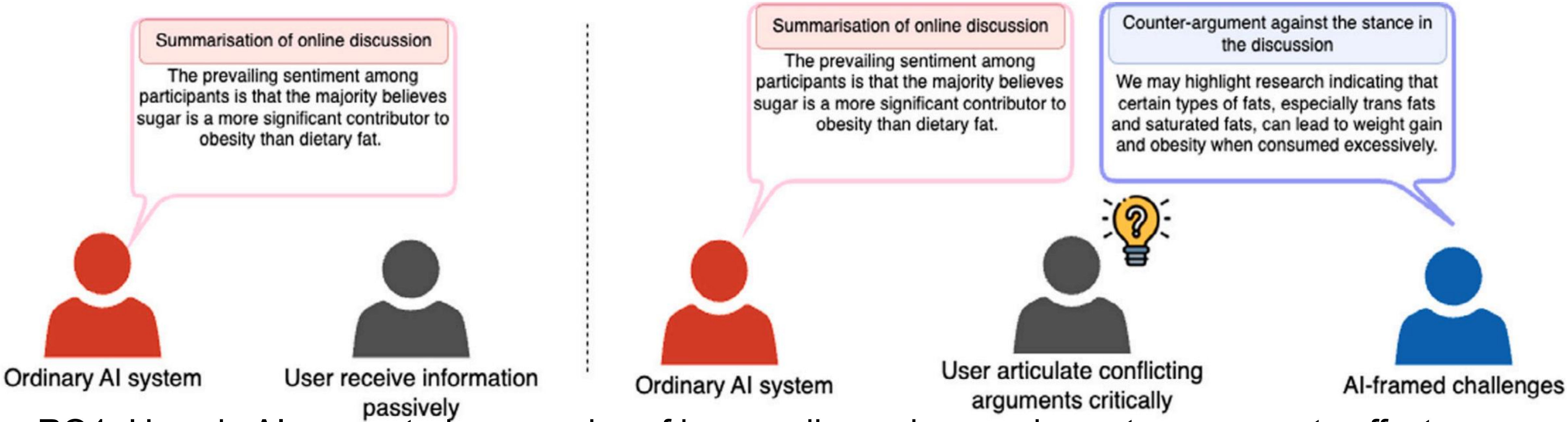
- User involvement does not significantly impact user trust and calibrated trust
- User involvement in planning can harm plan quality in tasks with a high-quality plan
- Plan quality has a significant positive correlation with calibrated trust
- User involvement in planning can help address imperfect plans

- Recommended approaches:
    - Iterative LLM agent simulation to decide when users should be involved
    - Users may need to articulate or manually override the agent action, posing a high cognitive load



Gaole He, Gianluca Demartini, and Ujwal Gadiraju. **Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant**. In: ACM CHI 2025 Conference on Human Factors in Computing Systems (CHI 2025). Yokohama, Japan, April 2025.

AI for Emotion Regulation and Opinion Formation



RQ1: How do AI-generated summaries of human discussions and counter arguments affect people's decision-making?

RQ2: How does AI summarisation affect user-perceived fairness and bias?

Shangqian Li, Lei Han, and Gianluca Demartini. **Provoking Critical Thinking: Using Counter-arguments in Online Discussion Summarisation**. In: Information Processing & Management (IP&M), Elsevier, June 2025

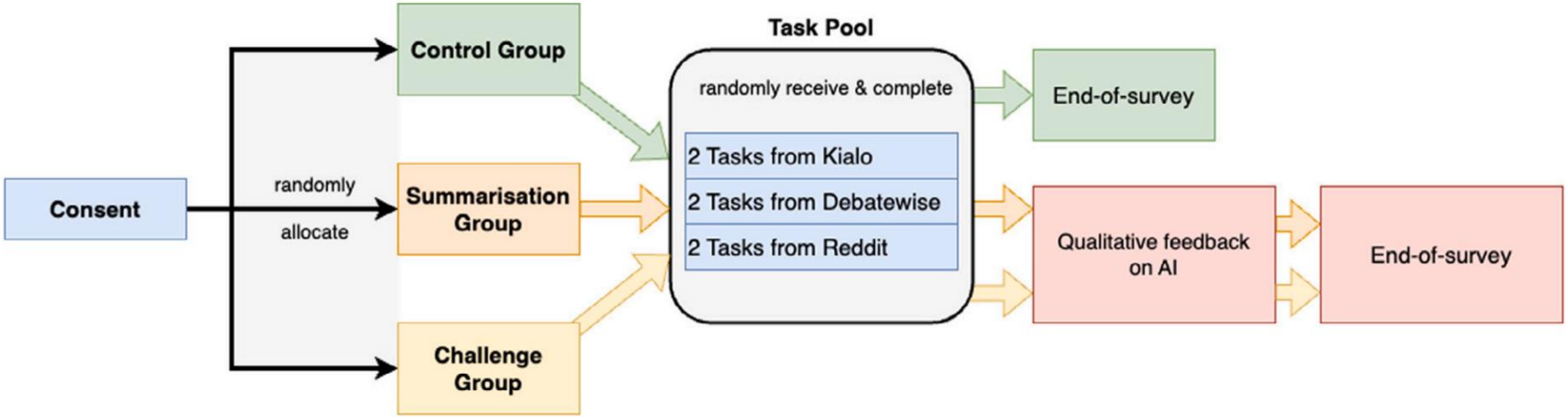
Example

AI Summarisation

The discussion revolves around Elon Musk's ownership of Twitter and the motivations behind it. The original assertion is that Musk is using Twitter for power, not money, with the argument that he can exert significant influence on world affairs through the platform. Others counter this by suggesting that Musk's ownership is more about ego and spite, particularly in dealing with critics and winning arguments. It's also mentioned that Musk was essentially forced to buy Twitter due to legal pressure.

AI Challenge

While there is debate about Musk's motivations, one aspect that deserves consideration is the potential impact of his ownership of Twitter on free speech and public discourse. With Musk wielding considerable influence and making decisions about checkmarks, bans, and platform rules, there's a risk that Twitter's policies and practices could become increasingly aligned with his personal views and preferences. This may raise concerns about whether Twitter remains a platform for open dialogue and diverse opinions or becomes a tool for furthering a specific agenda, regardless of whether it's driven by power, ego, or other motives. Ensuring the continued neutrality and inclusivity of Twitter could be a critical issue as Musk's influence on the platform grows.



Observations.

- Users are more likely to shift opinions after receiving (perceived) ‘unbiased’ AI assistance
- The system triggers reflective thinking by posing fresh arguments: The ‘challenge’ system assumed a starting stance -> bias

gianlucademartini.net
demartini@acm.org
@eglu81

Lessons learned and what to do

- Bias is present in human-generated data and is propagated in data pipelines
- LLMs carry human bias, but can help in data pipelines
- How do we make humans and LLM work together effectively?

- Track and profile data bias across the AI pipelines
- Select and diversify the sources of the labels (i.e., human annotators, LLMs)
- **Bias management** instead of bias removal

Demartini et al. "**Data Bias Management**", in *Communications of the ACM, Vol. 67, No. 1, Jan 2024*

To be continued …

DOI:10.1145/3611641
Gianluca Demartini, Kevin Roitero, and Stefano Mizzaro

Opinion

Data Bias Management

Envisioning a unique approach toward bias and fairness research.

THE PRESENCE OF bias in data has led to a lot of research include work looking at how to remove bias from learned word embeddings. increase fairness across groups when doing data augmentation,[17] feature

Visiting PhD Students Scheme

arc training centre for information resilience

Visit us in Brisbane, Australia!

2 or 3 months visits for PhD students to work on a joint paper

Funding and application instructions: https://cires.org.au/engagement/visitors/

Application deadlines in 2025:

~~March 22~~; June 22; September 22

Since 2023, we hosted 10 PhD students based in 7 countries

(CH, NL, DE, NO, BE, CN, IT); Example outputs:

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. **Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant**. In: ACM CHI 2025 Conference on Human Factors in Computing Systems (**CHI 2025**). Yokohama, Japan, April 2025.

Mads Skipanes, Tollef Emil Jørgensen, Kyle Porter, Gianluca Demartini, and Sule Yildirim Yayilgan. **Enhancing Criminal Investigation Analysis with Summarization and Memory-based Retrieval-Augmented Generation: A Comprehensive Evaluation of Real Case Data**. In: The 31st International Conference on Computational Linguistics (**COLING 2025**).