The University Of Queensland
AUSTRALIA
CREATE CHANGE

Humans and AI Making Decisions Together: Risks and Opportunities

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, CSUR), Prepositions (CIKM14), Entity Cards (SIGIR19)
  Eval (ECIR16 BestP, CIKM17, SIGIR18, CIKM19, WWW22, TOIS23, ICTIR23 BestP, ECIR26)
- **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)
  Decision making (CHB-R25 , HBET26), Critical Thinking (IP&M25)
- **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 and Better Data** (since 2018)
  Gender (SIGIR18,ACIS24,WWW25), Management (CACM24), SES (WebSci22,ICWSM25)
  Impact on ML (CIKM22), Political (WWW25, ICWSM26),
  Noise (WWW19, ICWSM25), Data Workers (SIGIR20, TOIS, TKDE, WWW23)
  Know. Graphs (ISWC19), Unknown Unknowns (ECAI20, HCOMP21), Behaviors (CIKM20)
  Fairness (CIKM22, SIGIR23, FAccT24, KDD24, TMLR), Active Learning (AAAI24)

Thanks to:

Australian Government
Australian Research Council

European Commission

Swiss National Science Foundation

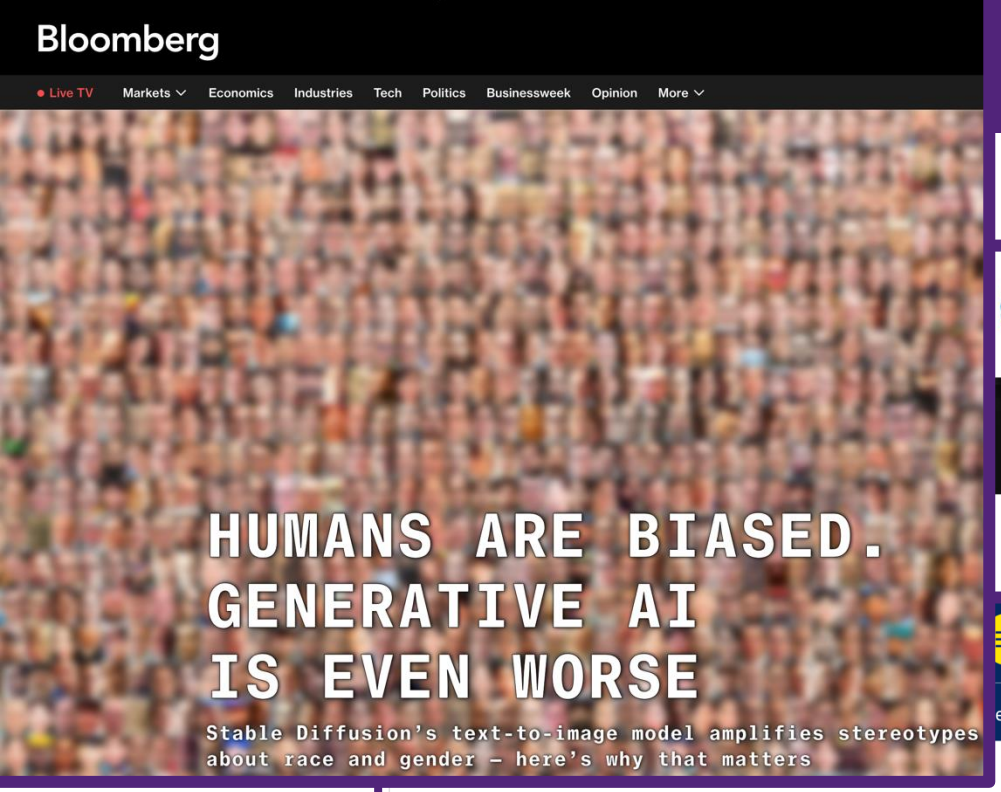
EPSRC
Engineering and Physical Sciences Research Council

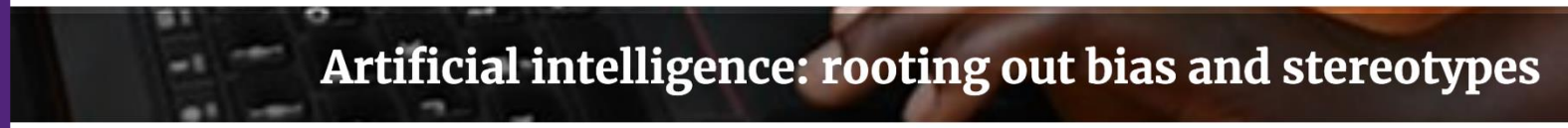
Meta

WIKIMEDIA FOUNDATION

Google AI

The Problem – Bias in AI





Use of AI could worsen racism and sexism in Australia, human rights commissioner warns

Outline

Bias in Humans and Bias in AI

- Gender completeness in Wikipedia
- Gender exploration in Wikipedia
- The impact of bias on ML models
- Bias Management

Humans Using AI – Trust and Reliance

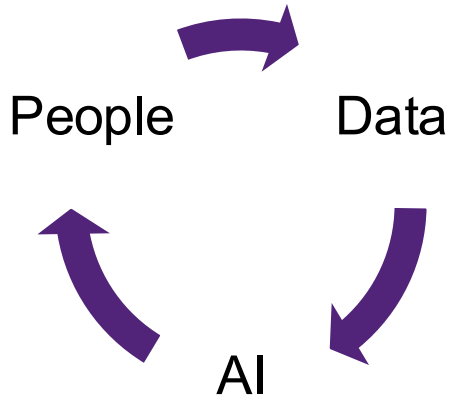
- Joint decision-making quality

Designing AI that Supports Human Decision Making

- Topics that require expertise
- Provoking critical thinking

Human-AI Collaboration Design

- LLM Agents to complete tasks for us
- A spectrum of Human-AI collaboration



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 %

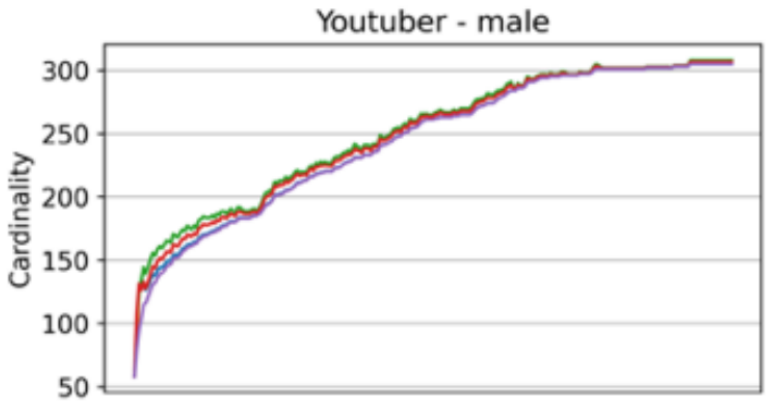
Hrishi Patel, Tianwa Chen, Ivano Bongiovanni, and Gianluca Demartini. **Estimating Gender Completeness in Wikipedia**. In: The Australasian Conference on Information Systems (ACIS 2024). Canberra, Australia, December 2024.

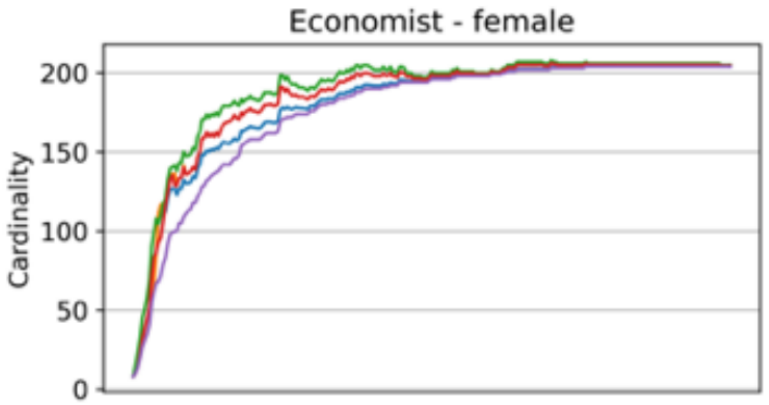
Exploration of Gender in Wikipedia

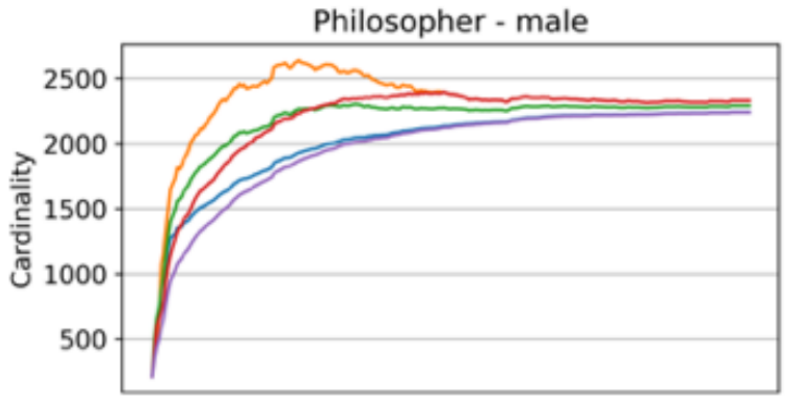
Yahya Yunus, Tianwa Chen and Gianluca Demartini. **Exploring Wikipedia Gender Diversity Over Time - The Wikipedia Gender Dashboard (WGD)**. In: The 2025 ACM Web Conference (TheWebConf 2025) - Demo track. Sydney, Australia, April 2025.

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

What happens when we train ML models with biased labels?

Online Demo at: https://recant.cyens.org.cy/

Periklis Perikleous, Andreas Kafkalias, Zenonas Theodosiou, Pınar Barlas, Evgenia Christoforou, Jahna Otterbacher, Gianluca Demartini, and Andreas Lanitis. **How Does the Crowd Impact the Model? A tool for raising awareness of social bias in crowdsourced training data**. In: The 31st ACM International Conference on Information and Knowledge Management (CIKM 2022). Atlanta, Georgia, USA, Oct 2022

1. Input image:

Click here to change the image

Current image: CFD-BF-003-003-N

2. Classification task:

Select a classification task.

Gender | Race | Attractiveness | Trustworthiness

The models try to predict the depicted person's Trustworthiness.

3. Results:

Click to show Results.

Execute

Nine different models were trained on the same images for each task, with different (sub)sets of crowd-worker annotations. The same input image (above) was passed through each of the nine models, resulting in the following outputs (possible outputs: Low, Medium, High):

Model	Model Description	Classification Decision
CFD Annotators	Model trained on the norming data provided with the CFD.	High
All Annotators	Model trained using all the annotations for all images.	Medium
Random	Model that simulates the case where annotators generate labels without considering the image content.	Medium
Men	Model trained using all the annotations provided by male crowdworkers.	Low
Women	Model trained using all the annotations provided by female crowdworkers.	Medium
Black	Model trained using all the annotations provided by Black crowdworkers.	Medium
Asian	Model trained using all the annotations provided by Asian crowdworkers.	Low
White	Model trained using all the annotations provided by White crowdworkers.	Medium
Latino	Model trained using all the annotations provided by Latino crowdworkers.	High

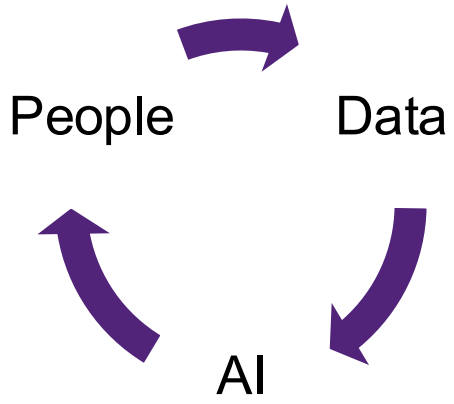
Bias Management, not bias removal

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*

Outline



Bias in Humans and AI

- Gender completeness in Wikipedia
- Gender exploration in Wikipedia
- The impact of bias on ML models
- Bias Management

Humans Using AI – Trust and Reliance

- Joint decision-making quality

Designing AI that Supports Human Decision Making

- Topics that require expertise
- Provoking critical thinking

Human-AI Collaboration Design

- LLM Agents to complete tasks for us
- A spectrum of Human-AI collaboration

Making Decisions with AI

Social Media platforms have a scale challenge when it comes to **content moderation**.

Task: **Online hate speech detection**; Humans making the final decision, *aided* by AI.

AI: gpt-5-2025-08-07

Humans: crowd workers from the Prolific platform

Step 1: Humans are shown a tweet
and asked to make a decision and report confidence level

Step 2: Humans are presented with AI decision, explanation, and confidence

Step 3: Humans make a final decision

Example Hateful Speech Detection Tast:

"Crooked Hillary you didn't lose because you couldn't portray yourself as an 'agent of change.' Your loss because you are a socialist agent of destruction. Shut up Hillary and stop whining ! "

Is it a hateful speech?

a. Not a hateful speech
b. Implicit hateful speech
c. Explicit hateful speech

What's your confidence level?

a. Very High (80-100)
b. High (60-80%)
c. Medium (40-60%
d. Low 20-40%)
e. Very Low (0-20%)

Findings – Confidence in the decision

What happens to human confidence once they get the AI recommendation?

Blue: The AI agrees with the original human decision

Red: The AI disagrees with the human

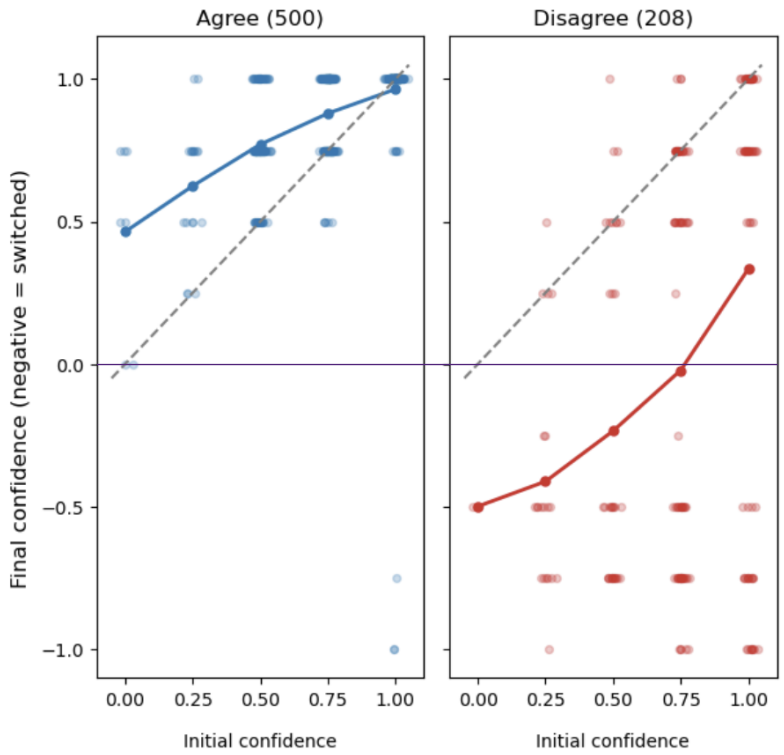


Figure 2: Post-advice confidence as a function of pre-advice confidence, divided by consensus (agree vs. disagree). Numbers in parenthesis indicate the number of data points plotted.

Findings – Changing decision if the AI disagrees?

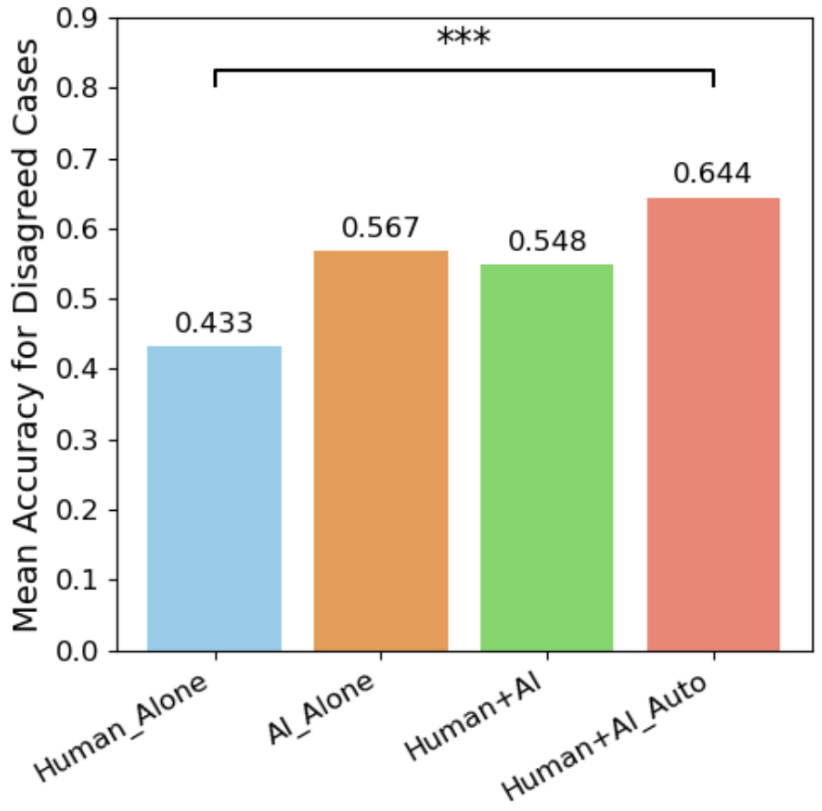
What heuristics do humans employ to **handle disagreement** under different relative confidence levels between human and AI, and how do these heuristics affect complementary team performance?

Confidence	Switch Rate	Acc. (HM)	Acc. (AI)	Acc. (Team)	CTP
HM High, AI High	0.76	0.53	0.47	0.54	True
HM High, AI Low	0.04	0.43	0.57	0.43	False
HM Low, AI High	0.86	0.18	0.82	0.75	False
HM Low, AI Low	0.34	0.29	0.71	0.44	False

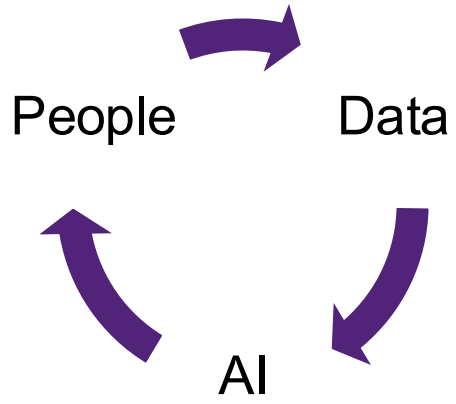
Table 5: Humans' switch rate and performance under different confidence combinations for disagreed cases.

Findings – Humans+AI better than Humans or AI?

Can complementary team performance be achieved between humans and AI, and if so, under what conditions and through what mechanisms?



Outline



Bias in Humans and AI

- Gender completeness in Wikipedia
- Gender exploration in Wikipedia
- The impact of bias on ML models
- Bias Management

Humans Using AI – Trust and Reliance

- Joint decision-making quality

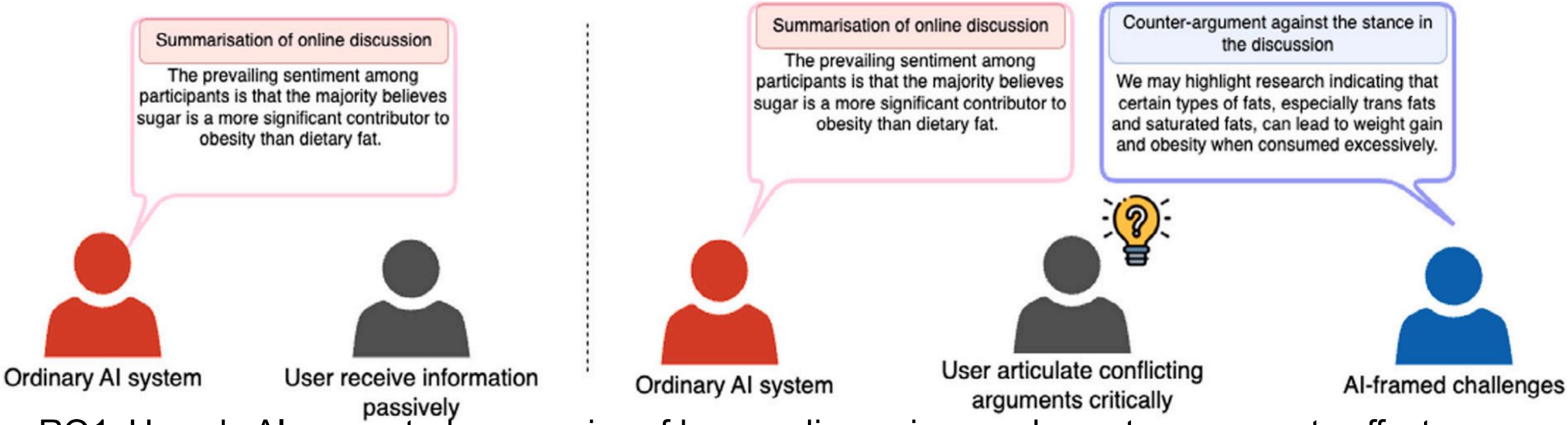
Designing AI that Supports Human Decision Making

- Topics that require expertise
- Provoking critical thinking

Human-AI Collaboration Design

- LLM Agents to complete tasks for us
- A spectrum of Human-AI collaboration

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

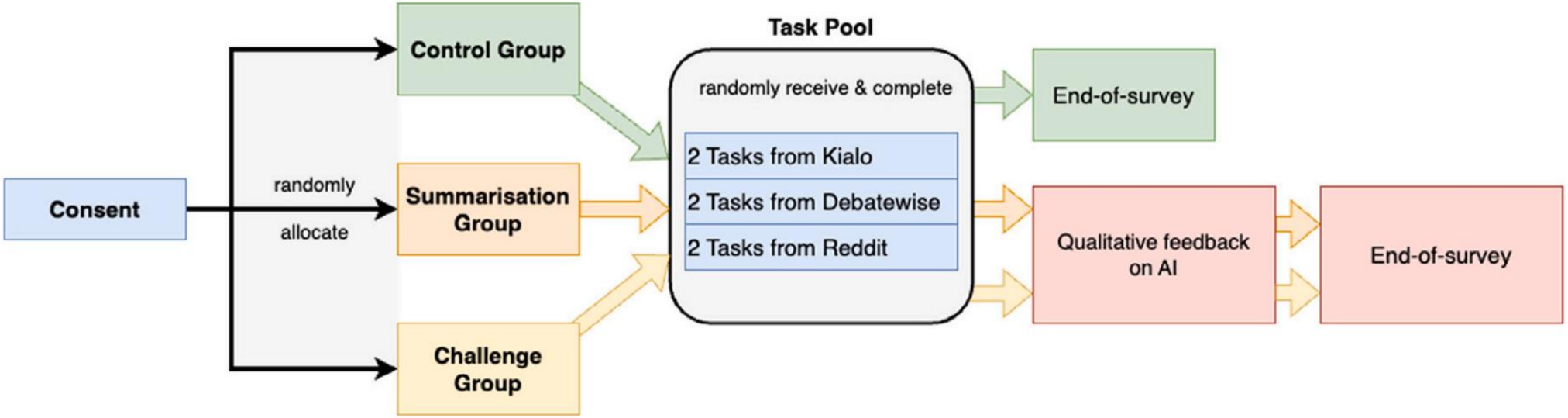
Example

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

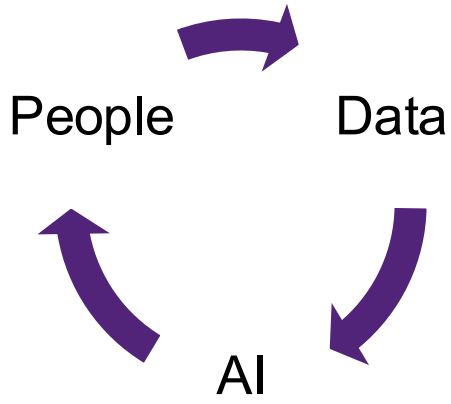
While there is debate about Musk's motivations, one aspect that deserves consideration is the potential impact of his ownership of Twitter on free speech and public discourse. With Musk wielding considerable influence and making decisions about checkmarks, bans, and platform rules, there's a risk that 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 of the user -> bias

Outline



Bias in Humans and AI

- Gender completeness in Wikipedia
- Gender exploration in Wikipedia
- The impact of bias on ML models
- Bias Management

Humans Using AI – Trust and Reliance

- Joint decision-making quality

Designing AI that Supports Human Decision Making

- Topics that require expertise
- Provoking critical thinking

Human-AI Collaboration Design

- LLM Agents to complete tasks for us
- A spectrum of Human-AI collaboration

Trust in AI Agent

Agents as daily assistants
Tasks with different levels of risk
AI 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 inappropriately trust the AI agents with plans that seem plausible

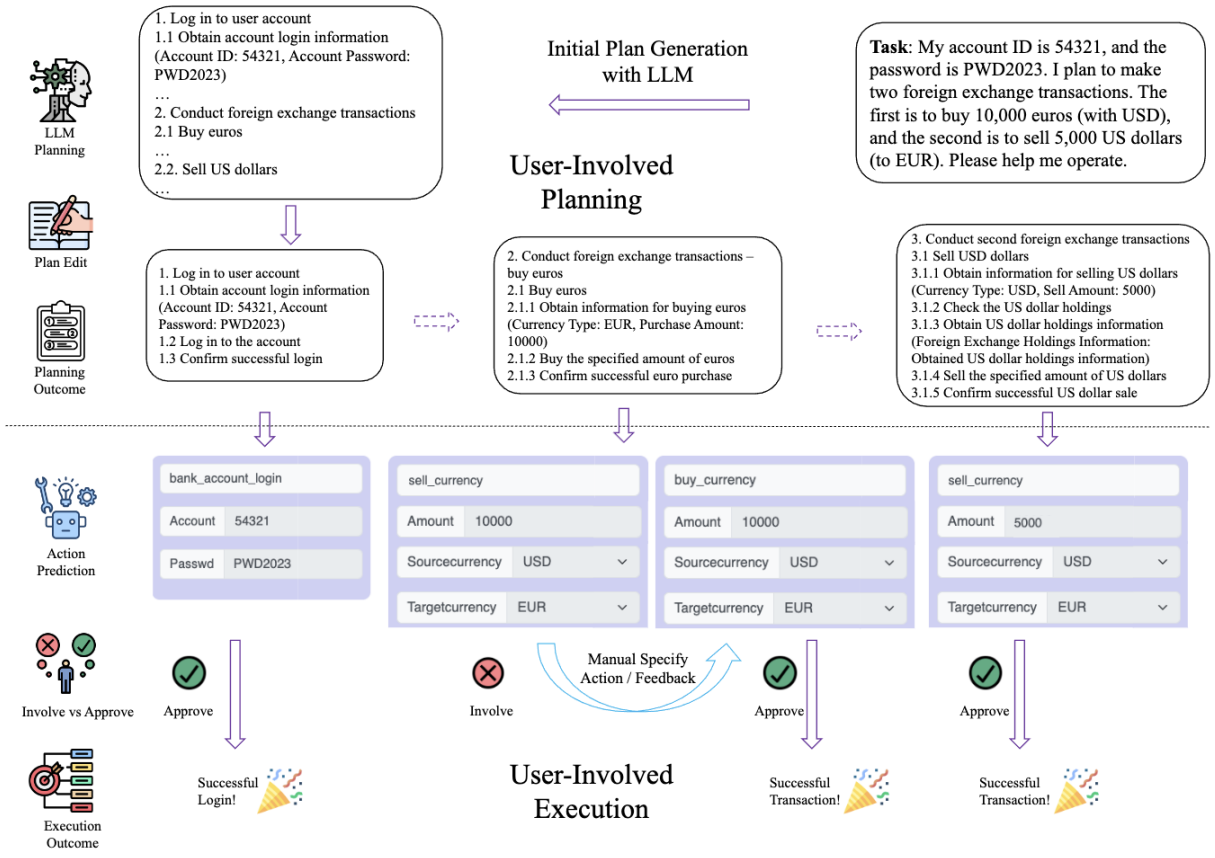


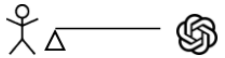
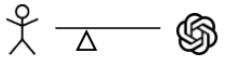
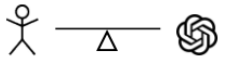
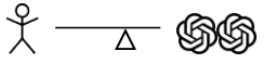
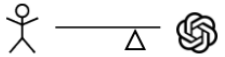
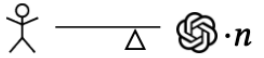
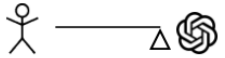
Figure 1: Illustration of the human-AI collaboration with plan-then-execute LLM agents.

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.

Human-AI Collaborative Spectrum

Humans used to annotate data
LLMs can replace humans in data annotation tasks
Microsoft Bing now uses GPT-4 for relevance judgments!

“Who is better?”
versus
“How can they work together?”

Collaboration Balance	Task Allocation
Human Judgment	
	Humans manually decide (about relevance) without any kind of AI support.
	Humans have full control of deciding but are supported by machine-based text highlighting, data clustering, etc.
Model In The Loop	
	Humans decide based on LLM-generated summaries needed for the decision.
	Balanced competence partitioning. Humans and LLMs focus on decisions they are good at.
Human In The Loop	
	Two (or more) LLMs each generate a decision, and a human selects the better one.
	An LLM makes a decision (and an explanation for it) that a human can accept / reject.
	LLMs are considered crowdworkers—varied by specific characteristics—, aggregated and controlled by a human.
Fully Automated	
	Fully automatic decision without humans.

Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth.
Who determines what is relevant? Humans or AI? Why not both!
In: Communications of the ACM (CACM). 2024.

demartini@acm.org
linkedin.com/in/gianlucademartini

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 decision making
- How do we make humans and LLM work together effectively?

- Increasing transparency, (e.g., track and profile data bias across AI pipelines)
- Select and diversify the sources of data (i.e., diverse human annotators, LLMs)
- **Bias management** instead of bias removal

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

Opinion

Data Bias Management

Envisioning a unique approach toward bias and fairness research.

To be continued …