The University of Queensland
Australia
CREATE CHANGE

Bias in Data Annotations

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), Data Workers (SIGIR20, TOIS, TKDE, WWW23), Behaviors (CIKM20)
  Know. Graphs (ISWC19), Unknown Unknowns (ECAI20, HCOMP21)
  Fairness (CIKM22, SIGIR23, FAccT24, KDD24), Active Learning (AAAI24)

Thanks to:

Australian Government
Australian Research Council

European Commission

Swiss National Science Foundation

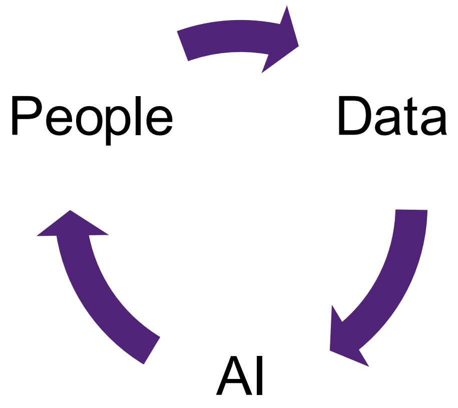
EPSRC
Engineering and Physical Sciences Research Council

Meta

WIKIMEDIA FOUNDATION

Google AI

Outline



Two examples of bias in data annotations

- Bias in crowdsourced fact-checking (ECIR 2020; SIGIR 2020)
- SES bias in humans and ML (WebSci 2022; ICWSM 2025)
- Human-AI annotations (CACM 2024; ICWSM 2024; CACM 2025)

Implications and solutions

- What happens when you train ML with biased labels (CIKM 2023)
- Bias Management (CACM Jan 2024)
- The BiasNavi tool (ACM TheWebConf 2025)

Crowdsourcing Truthfulness Judgements

~600 MTurk US workers

To assess truthfulness of

- US political statements (Politifact)
- non-US political statements (ABC)

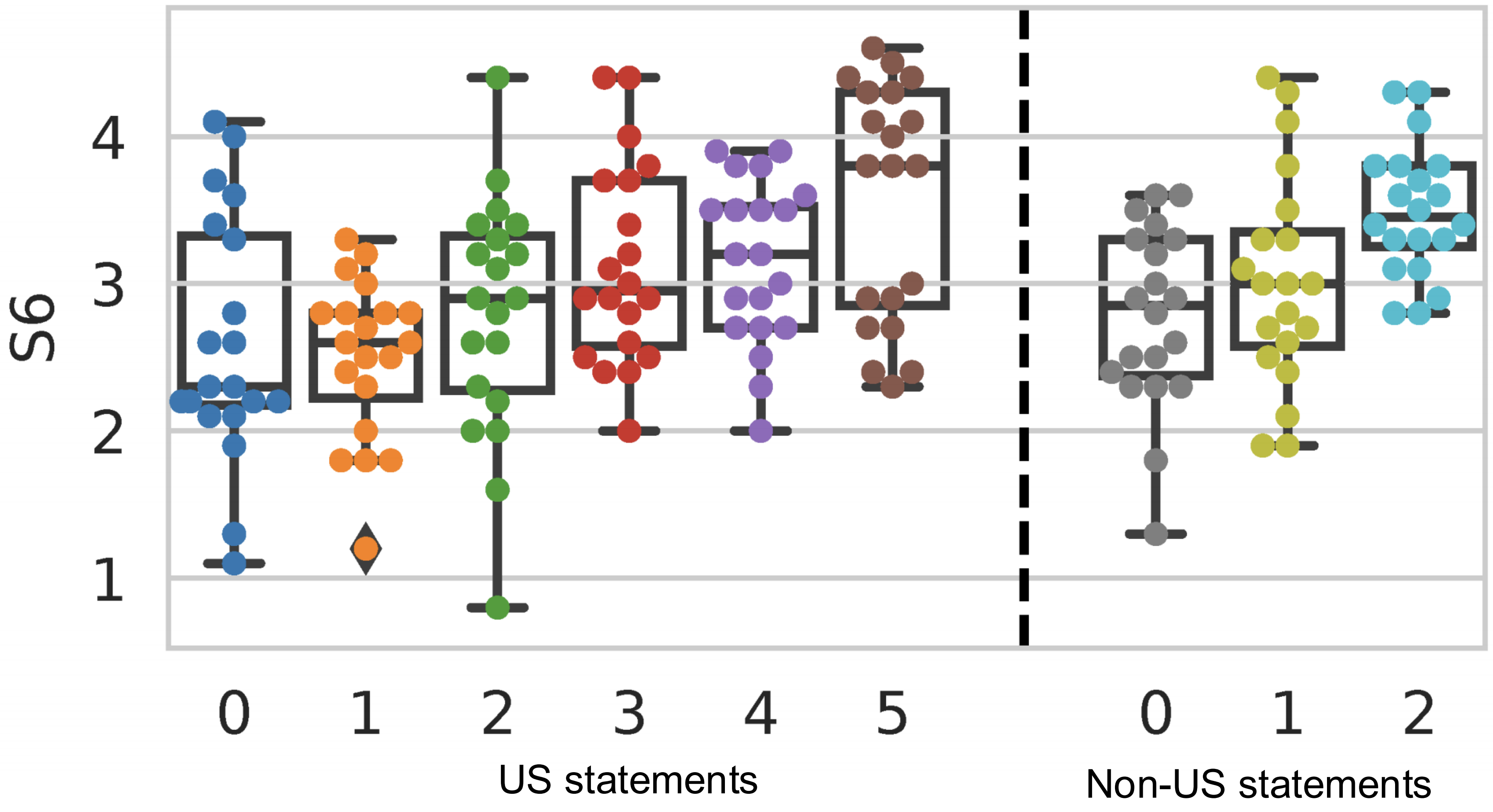
3 scales (3, 6, and 100 levels)

Table 1: Example of statements in the `PolitiFact` and `ABC` datasets.

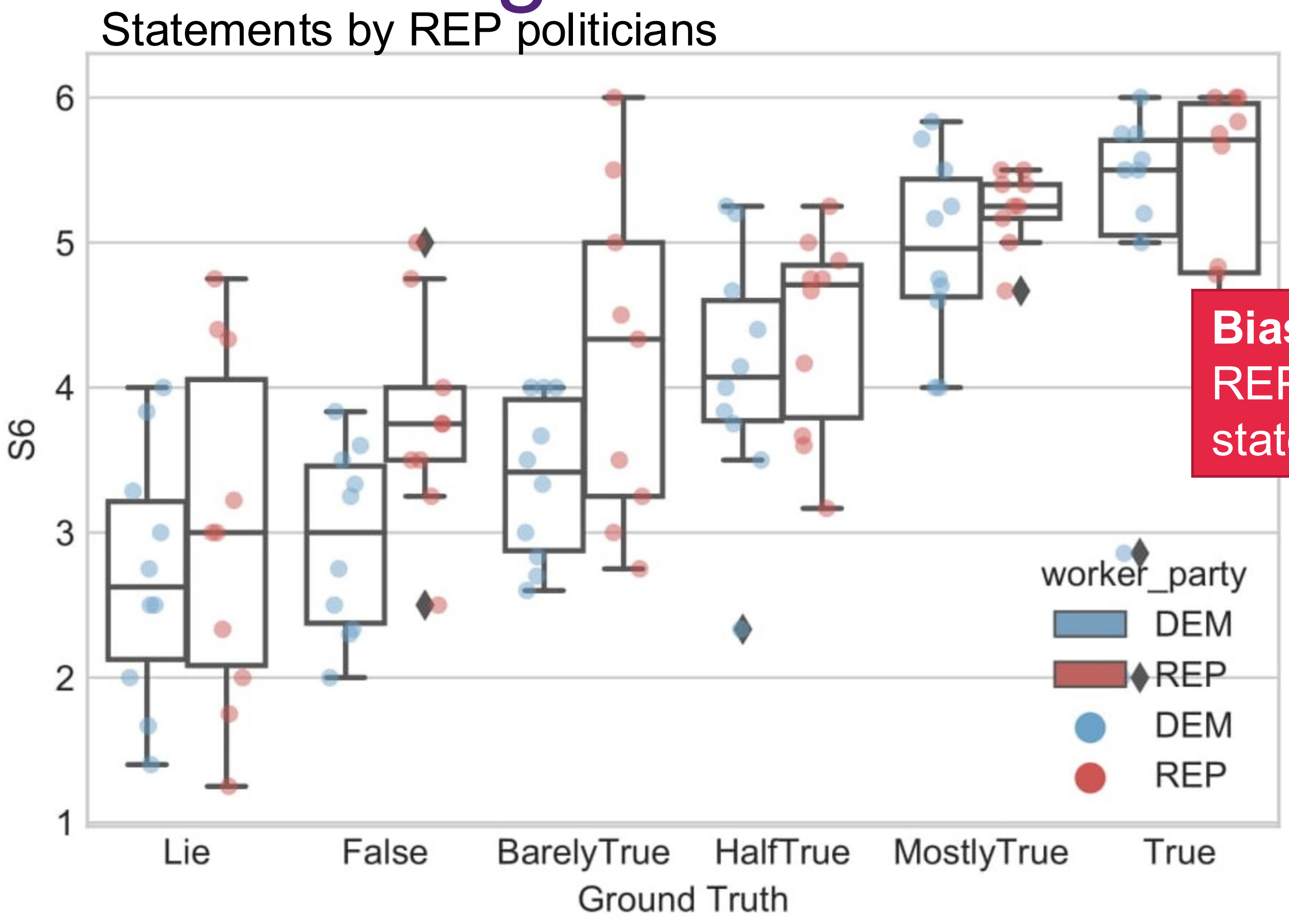
	Statement	Speaker, Year
`PolitiFact` Label: `mostly-true`	"Florida ranks first in the nation for access to free prekindergarten."	Rick Scott, 2014
`ABC` Label: `in-between`	"Scrapping the carbon tax means every household will be $550 a year better off."	Tony Abbott, 2014

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro and Gianluca Demartini. **Can The Crowd Identify Misinformation Objectively? The Effects of Judgments Scale and Assessor's Bias.** In: The 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020)

Crowd Performance VS Expert Ground Truth



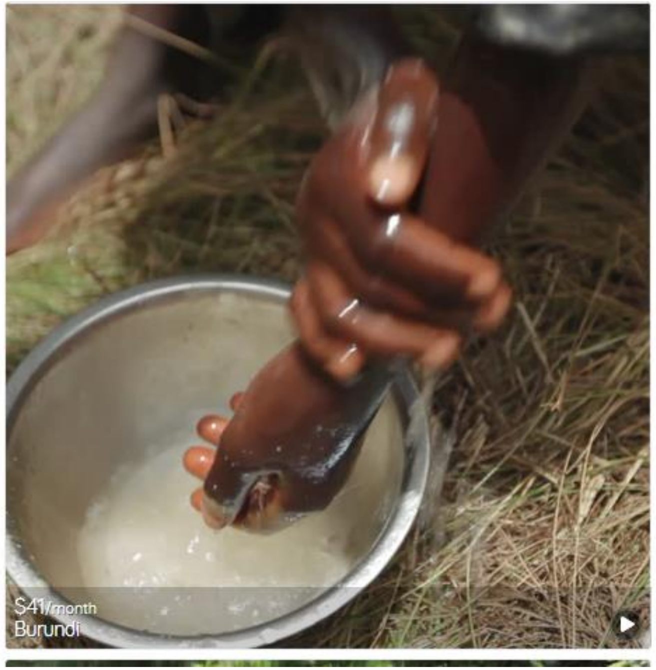
Fake News labelling - Political bias



Bias: Non-expert people who vote REP are more likely to believe to statements by REP politicians.

David La Barbera, Kevin Roitero, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. **Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias**. In: The 42nd European Conference on Information Retrieval (ECIR 2020 - Best paper award).

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)

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

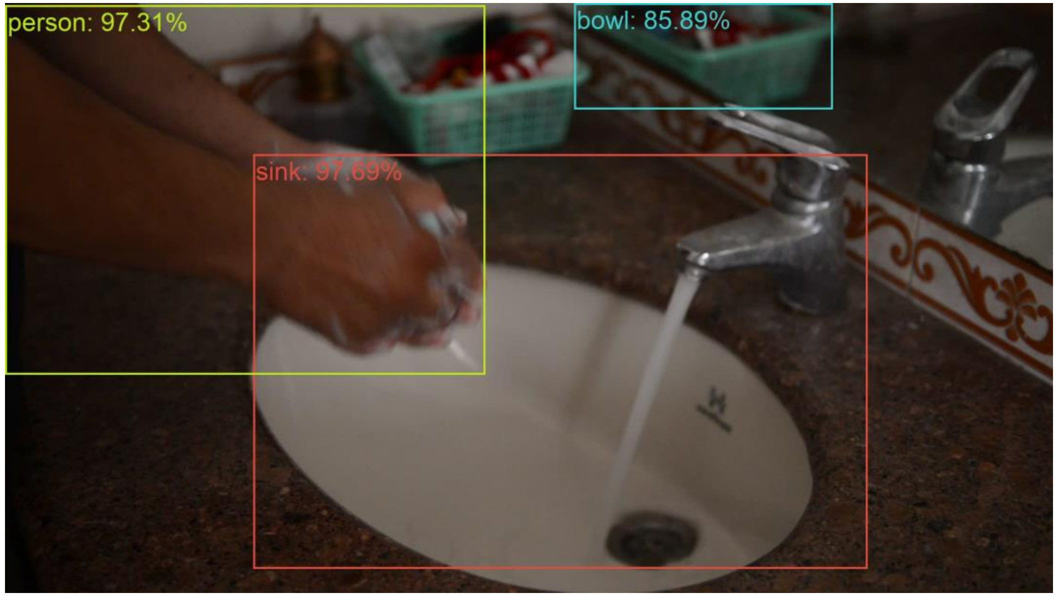
Bias in the annotation of SES-diverse content

- **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.

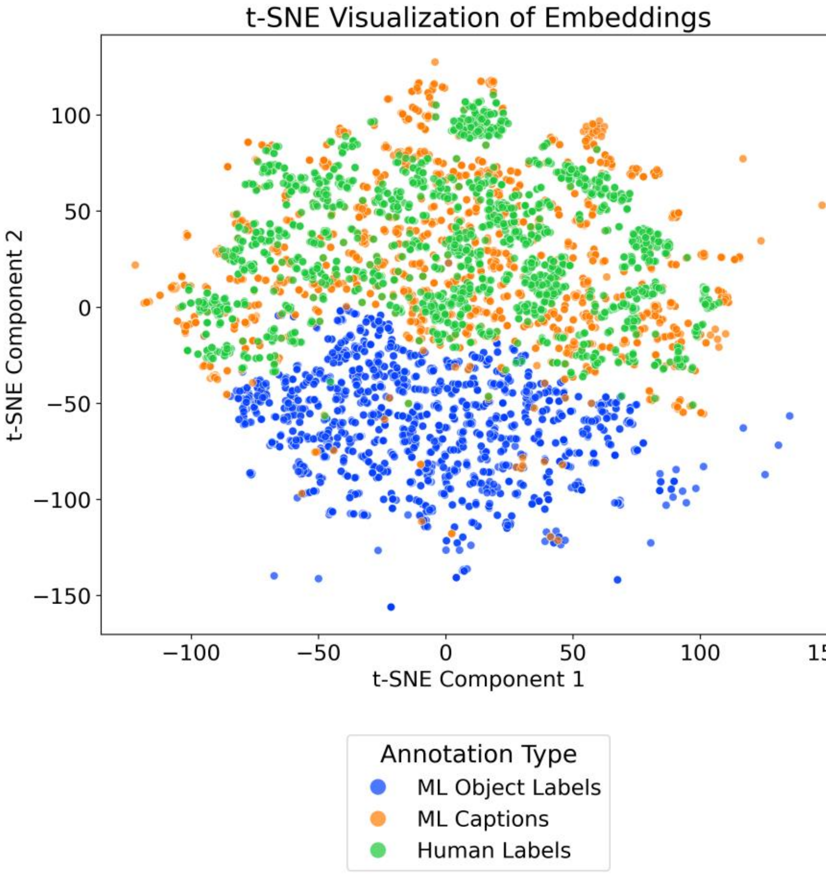
Research Questions

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**

Bias in LLMs?
The role of Humans

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.
$\cdot n$	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.

Strong use of chatGPT
Especially on Amazon MTurk

Generative AI in Crowdwork

	ALL	USA	India	UK	EU
Prolific	13.1% 13.4%	19.0% 14.0%	- -	9.0 % 10.0%	9.0% 14.5%
MTurk	80.3% 73.2%	94.3% 86.2%	66.3% 59.4%	- -	- -
Clickworker	20.7% 15.0%	27.9% 20.6%	- -	16.9% 11.0%	15.3% 12.6%

We asked crowd workers regarding their use of GenAI tools.

Table 4: Workers reporting self-initiated use of AI chatbots in tasks, by platform, country and T1/T2 [top/bottom].

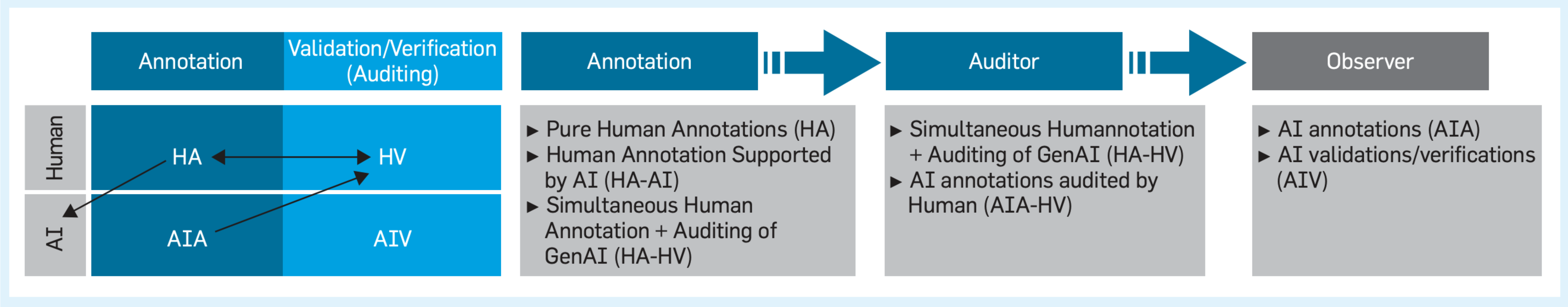
Prolific, Mturk, Clickworker; May 2023, and Dec 2023

- Workers' self-reported use of GenAI
  - did not change over time
  - was strongly correlated to the platform they use.
- **MTurk workers use GenAI on their own volition** significantly more often than those operating at Clickworker or Prolific.
- Many expressed concerns that GenAI would reduce the number of opportunities for surveys, as requesters are looking for authentic human responses.

Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. **Generative AI in Crowdwork for Web and Social Media Research: A Survey of Workers at Three Platforms**. In: The 18th International AAAI Conference on Web and Social Media (ICWSM 2024).

Crowd-Sourcing or AI-Sourcing?

There will always be a role for humans in AI pipelines, although GenAI is disrupting the crowdsourcing environment as we know it.



Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. **Crowd-Sourcing or AI-Sourcing? - The Impact of GenAI on Data Annotation Tasks**. In: Communications of the ACM (CACM), Vol. 68, No. 4 April 2025.

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

What happens when we train ML models with biased labels?

Live 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.

Bias: Depending on who the human annotators are, the ML classifiers will make different decisions

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

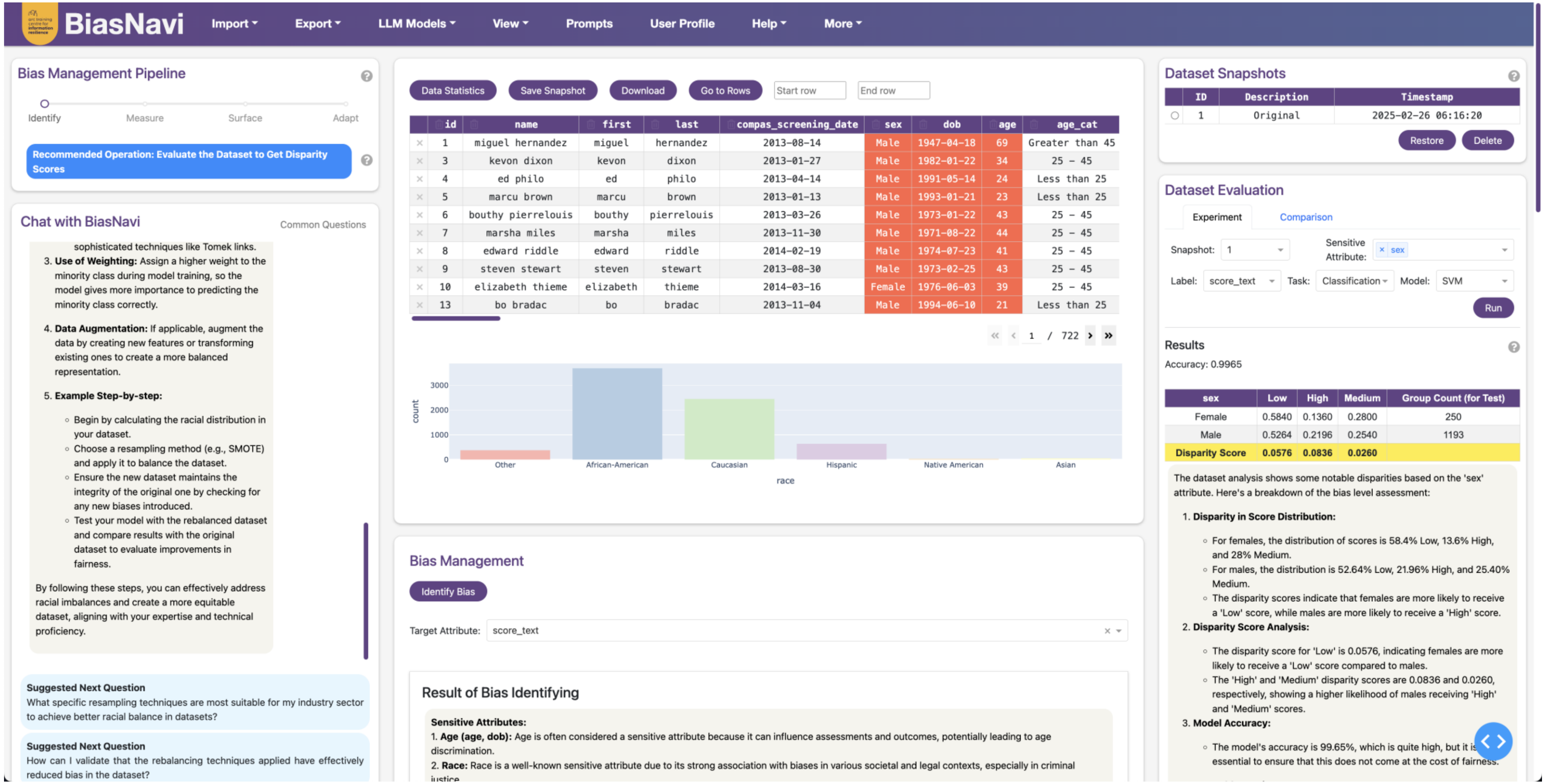
Employing an explicit and not transparent bias removal intervention might be potentially harmful to the user

Figure 2. The five steps of bias management.

https://doi.org/10.1145/3611641

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

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
- Bias comes from human annotators as much as system design choices

- 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

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

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)

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**).