Trustworthy AI in Practice: How Bias and Fairness Challenges Intersect with Privacy and Data Constraints

Timo Spinde, DAAD postdoctoral researcher, NII Tokyo
Hanrui Wang, assistant professor, NII Tokyo

DAAD IFI Deutscher Akademischer Austauschdienst German Academic Exchange Service

MBG Media Bias Group

NII 国立情報学研究所 National Institute of Informatics

Agenda

01 **About us**
Echizen Laboratory and the Media Bias Group

02 **Privacy leakage and data constraints**
Face recognition, model inversion and computer-use AI

03 **Bias and fairness evaluation**
Data coverage, clear categories and distinct evaluation questions

04 **Shared definitions and tools**
Taxonomy development, Dagstuhl and BARI

05 **Discussion and further resources**
Implications for trustworthy AI in practice

About us


DAAD
IFI
Deutscher Akademischer Austauschdienst
German Academic Exchange Service


MBG
Media Bias Group


NII
国立情報学研究所
National Institute of Informatics

Echizen Laboratory

Multimedia security and privacy, cognitive security and Social AI



Prof. Isao Echizen

NII Professor / Director · Professor at UTokyo & SOKENDAI

30+ research staff and students

CURRENT FLAGSHIP PROJECTS

SYNTHETIQ X

JST K Program · AI Security

Defend against and prevent the spread of fake information across video, audio, text, and multimodal media.

Research platform: real/fake datasets + generation, defense & prevention models

CREST Social AI

JST CREST · Trust-Oriented Social AI Design

Design trustworthy human–AI collaboration to strengthen resilience in an uncertain, misinformation-rich society.

Echizen group: cognitive security & trustworthiness of AI in complex information environments

https://research.nii.ac.jp/~iechizen/official/index-e.html

Timo Spinde

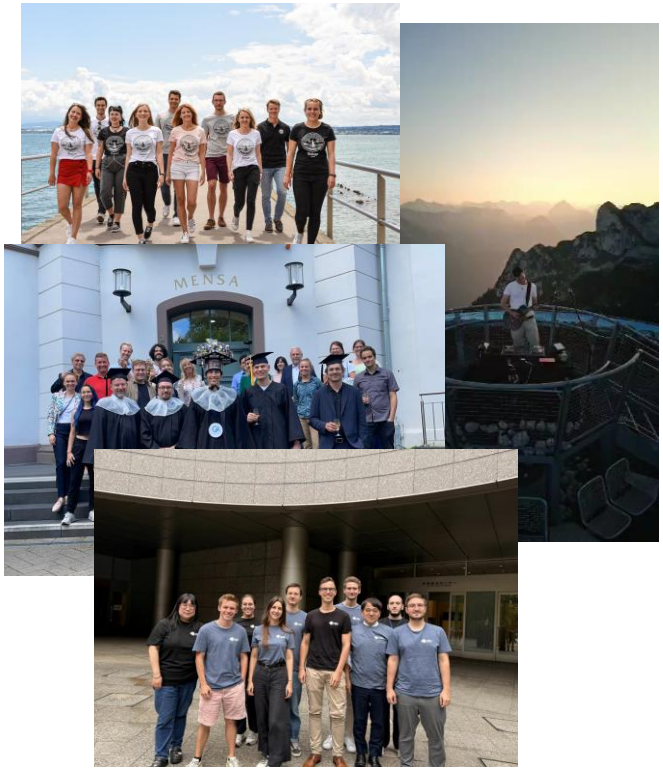
Main focus computational detection of media bias, AI safety, language processing

Studied Journalism and Computer Science

Founded multiple companies

Research experience across Germany, Australia, Israel and Japan

Ph.D. (2025) with distinction
~40 publications and ~€1.2 million in acquired funding



Hanrui Wang



Assistant Professor

Echizen Laboratory

National Institute of Informatics (NII)
Tokyo, Japan

RESEARCH HIGHLIGHTS · AI SECURITY & PRIVACY

01 Security & Privacy of Emerging AI Systems

AI Agents · Computer-Use Systems · World Models · 4D Models

Security and privacy risks in autonomous interaction and environmental understanding.

02 Adversarial Machine Learning

Adversarial Robustness · Attack Transferability
Robustness Evaluation · Model Inversion

Understanding, evaluating and exposing security and privacy risks.

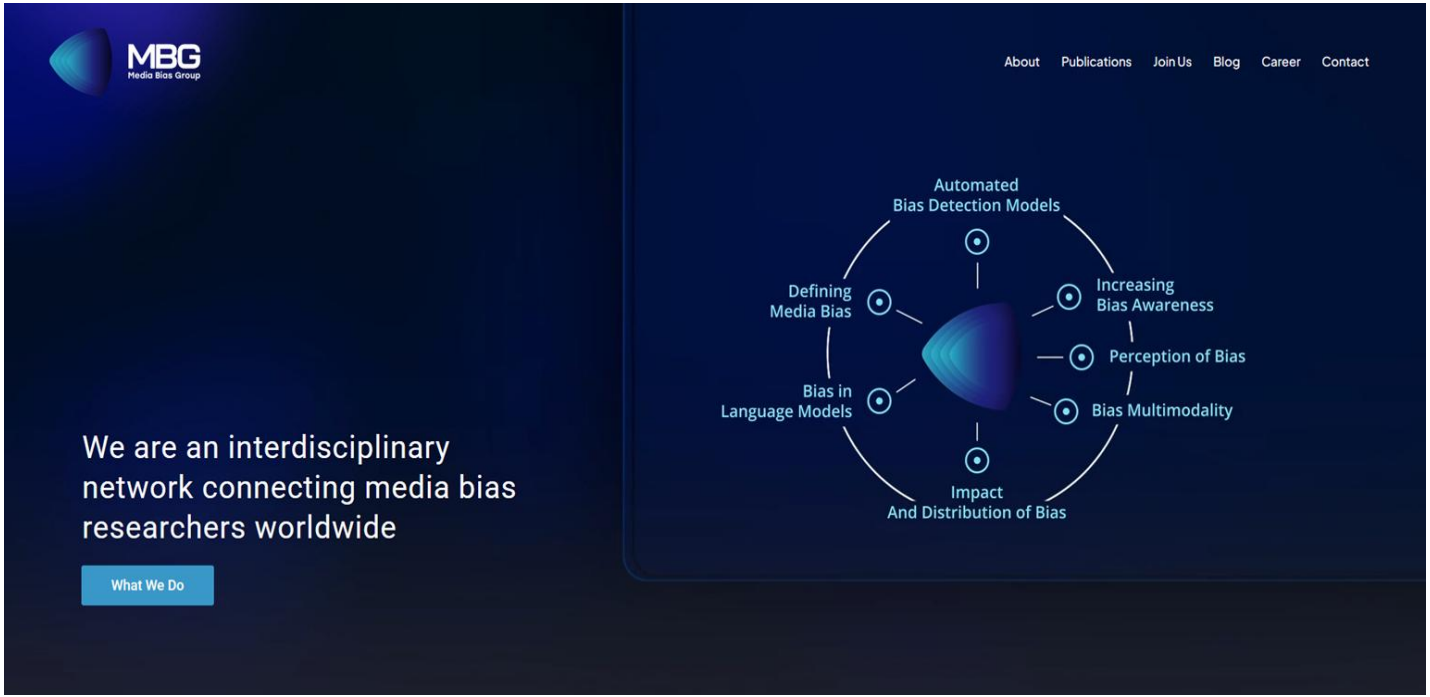
03 AI Evaluation & Governance

Scientific Peer Review · AI-Assisted Evaluation · AI Governance

Reliable and accountable AI in consequential evaluation and decisions.

Ph.D., Computer Science · Monash University | Former IT industry director

Media Bias Group



An interdisciplinary network connecting media bias researchers worldwide

Media Bias Group: goals



Bias theory and definitions



Bias Education and Awareness



Bias identification tools (BARI)

Privacy and data constraints

What models can reveal, and what we can protect







The privacy–bias dilemma

More personal data may improve coverage, but can also create privacy risks.

MORE PERSONAL DATA



Broader coverage

POTENTIAL BENEFIT

Better performance and group coverage.

RISK: more privacy exposure

LESS / RESTRICTED DATA



Missing groups?

POTENTIAL COST

Missing groups can increase bias.

BENEFIT: lower data exposure

Membership inference: was my face used?

A person tests whether their own photo was part of the model's training data.

1 TRAIN



Many training identities

2 QUERY



My own selfie

Trained face model

Prediction confidence

3 INFER

Was my face in training?

YES? **NO?**

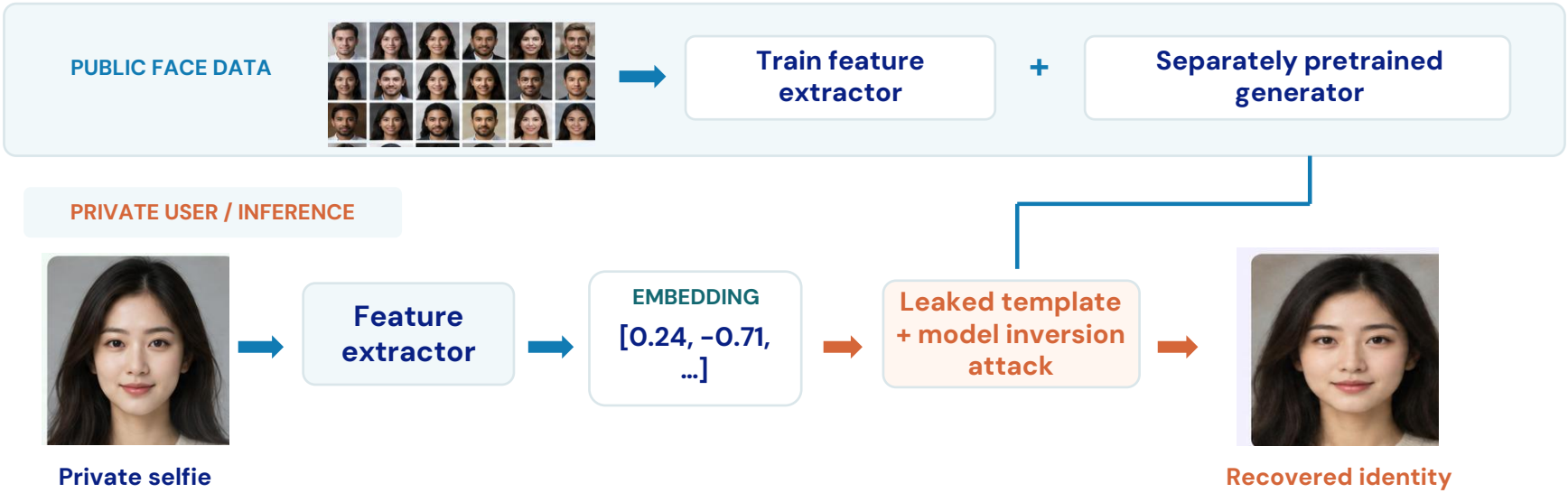
An estimate, not proof

PRIVACY RISK **The model may reveal whether a person's face was used in training.**

Shokri et al., "Membership Inference Attacks Against Machine Learning Models," IEEE S&P, 2017.

Numbers can still reveal your face

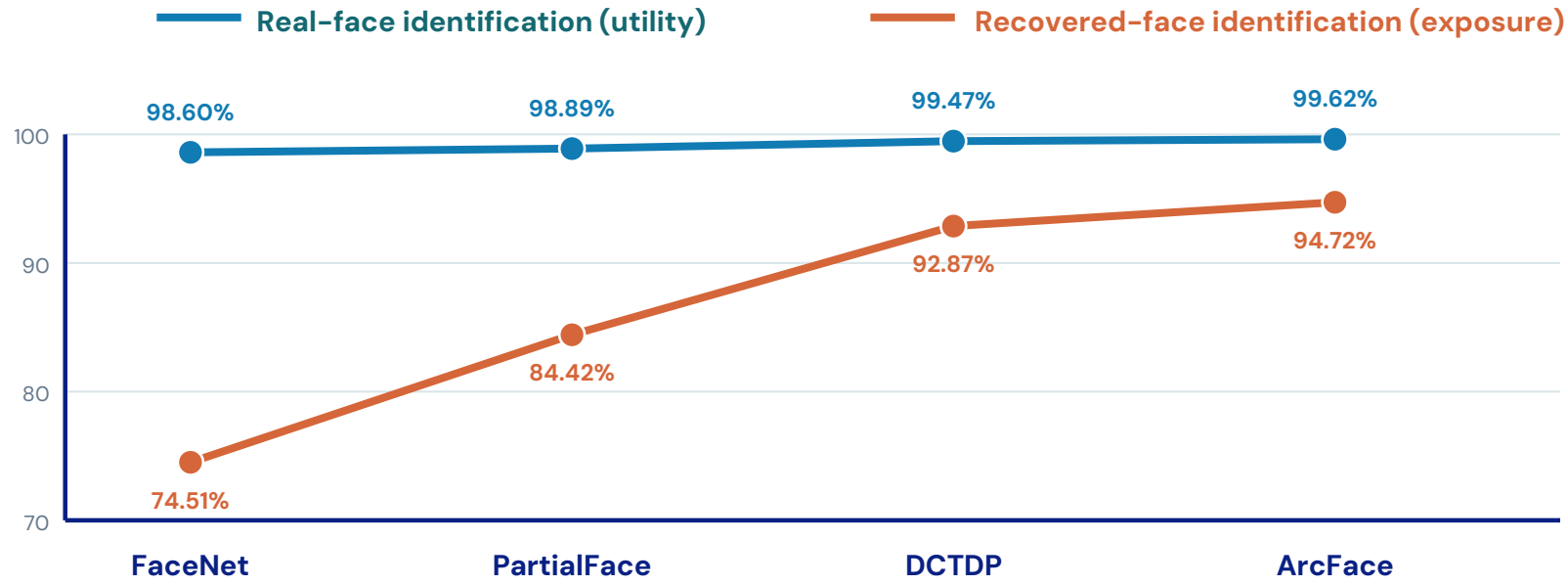
A public-data-trained face model turns a private selfie into numbers—but leaked numbers can expose identity.



Wang et al., "DiffMI: Breaking Face Recognition Privacy via Diffusion-Driven Training-Free Model Inversion," IEEE TIFS, 2026.

Better recognition, easier identity recovery?

Four face models: reported real-face and reconstructed-face matching rates.
This comparison does not establish causality.



In this comparison, higher recognition accompanies higher recovery success.

Wang et al., "DiffMI: Breaking Face Recognition Privacy via Diffusion-Driven Training-Free Model Inversion," IEEE TIFS, 2026.

How can we protect people?

Three practical security questions for future face recognition systems.

01 Biometric Template Protection



Face → embedding → key-encrypted template

Keep matching possible while making leaked embeddings harder to invert.

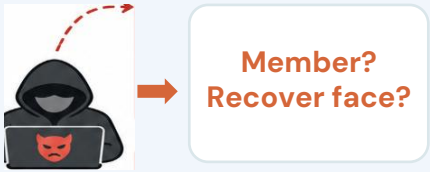
02 Model Unlearning



A person requests their face be removed.

Can a trained model forget one person without losing others?

03 Defend Against Privacy Attacks



Membership inference + model inversion

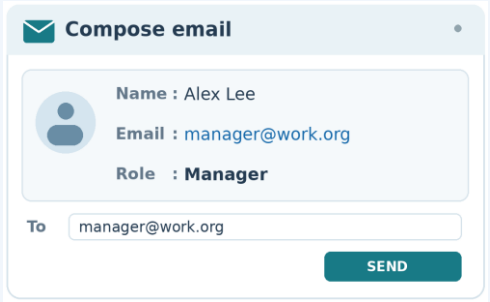
Detect or prevent training-membership and face-reconstruction leaks.

Reason Without Seeing

How much personal information must a computer-use AI see to complete a task?

EXAMPLE TASK: Send an email to my manager.

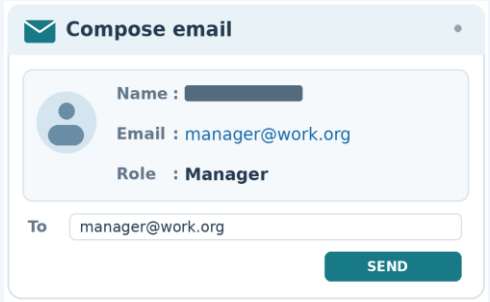
FULL SCREEN



Task **SUCCESS**

Privacy exposure **FULL**

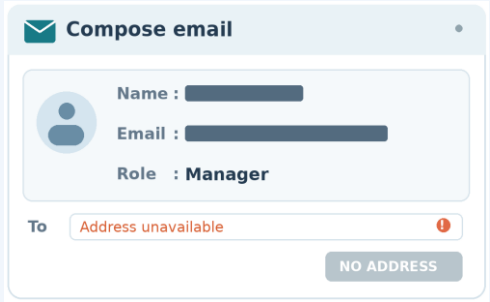
TASK-RELEVANT VISIBLE



Task **SUCCESS**

Privacy exposure **PARTIAL**

MORE DATA HIDDEN



Task **FAILED**

Privacy exposure **LEAST**

RESEARCH QUESTION **How much can we hide while still finishing the task?**

Privacy, fairness and bias evaluation

Privacy choices change what data we can use and what we can measure.

Privacy leakage	What can a model reveal about an individual?
Fairness	How do errors and outcomes differ across groups?
Media bias	How does content select, frame or portray people and events?

Media bias illustrates why explicit categories are essential for evaluation.

Understanding bias

Clear categories make evaluation more meaningful

DAAD IFI Deutscher Akademischer Austauschdienst German Academic Exchange Service
MBG Media Bias Group
NII 国立情報学研究所 National Institute of Informatics

Components of media bias



Word choice

How does wording influence the portrayal of an event or a person?

Components of media bias

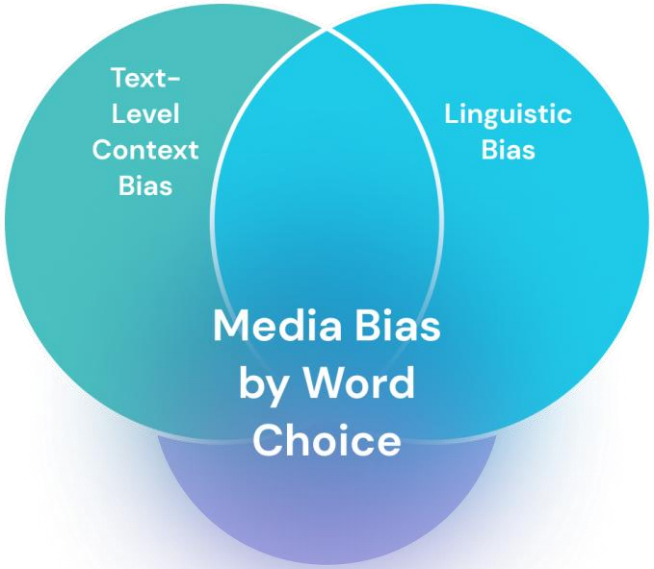


Linguistic bias

Language can carry associations and stereotypes that models may learn.

Caliskan et al., Science (2017). Garg et al., PNAS (2018).

Components of media bias

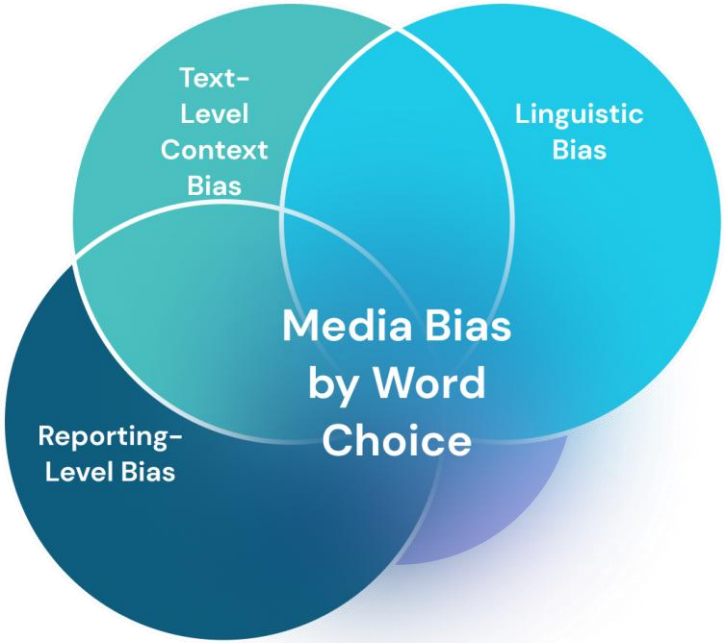


Text-level context

Framing, emphasis and omitted context shape what a text communicates.

Nakov et al., EACL (2017). Fan et al., EMNLP (2020).

Components of media bias

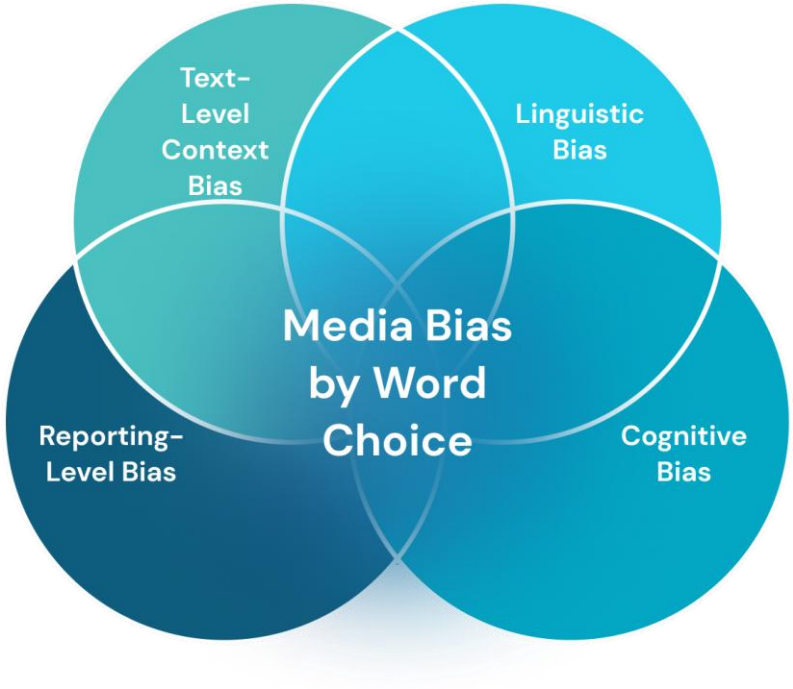


Reporting-level bias

Selection and prominence affect which people and events receive attention.

Gentzkow & Shapiro, Journal of Political Economy (2006). Puglisi & Snyder, JEEA (2015).

Components of media bias

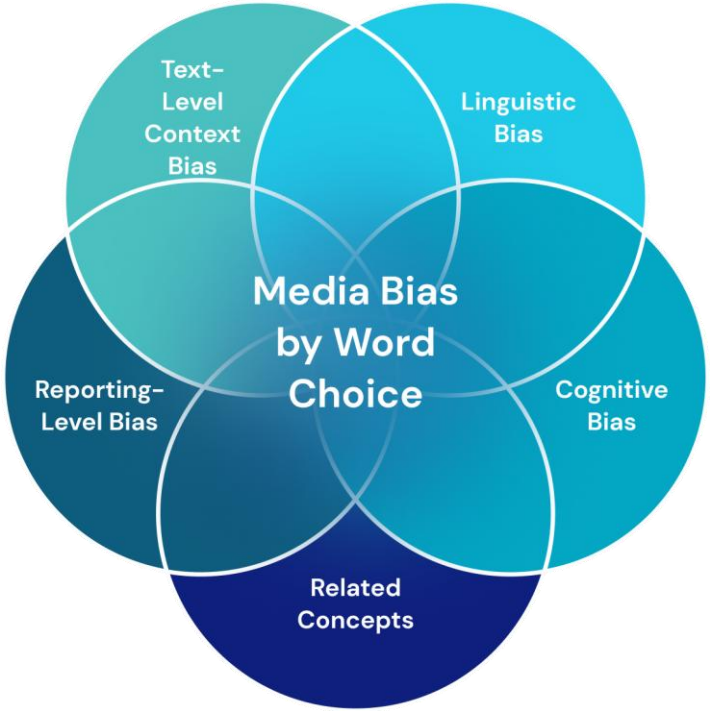


Cognitive bias

Prior beliefs and selective interpretation influence how people receive information.

Lord et al., JPSP (1979). Nickerson, Review of General Psychology (1998).

Components of media bias



Related concepts

Overlapping concepts need explicit definitions before we can compare model outputs.

Baum & Groeling, Political Communication (2008). Iyengar & Hahn, Journal of Communication (2009).

Shared definitions and tools

Taxonomy development, Dagstuhl and BARI







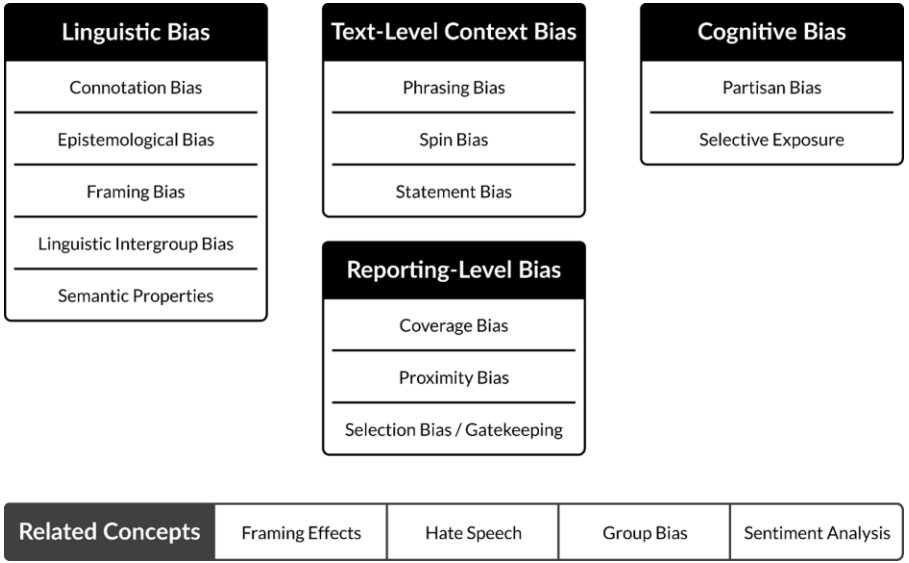
Defining bias: building a shared taxonomy

Existing taxonomies from systematic reviews need broader coverage.

We collect papers on media bias.

We build an automated definition extraction and summarization pipeline.

We hand-annotated ~2,400 papers for relevance and definitions.



Validating the taxonomy at Dagstuhl

Standardizing Media Bias Research

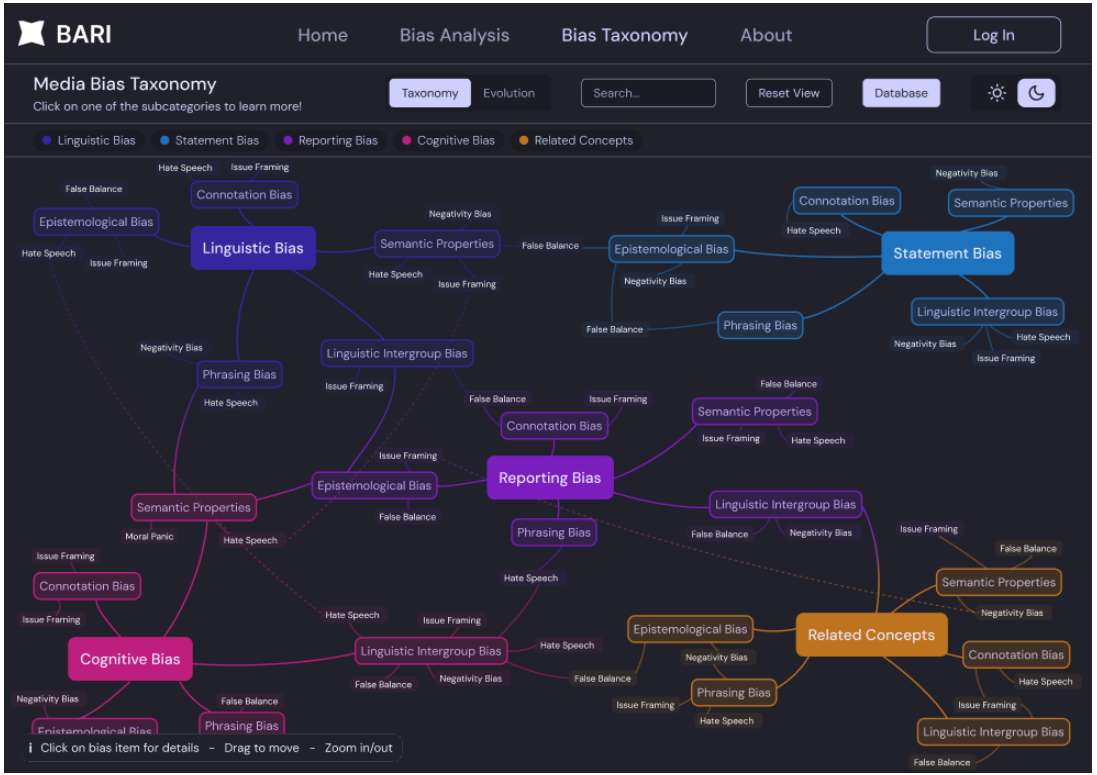
Accepted Dagstuhl seminar

Validate the taxonomy and discuss the state of the art in bias research.

The definition-extraction work provides a starting point for shared terminology.



BARI: exploring the taxonomy

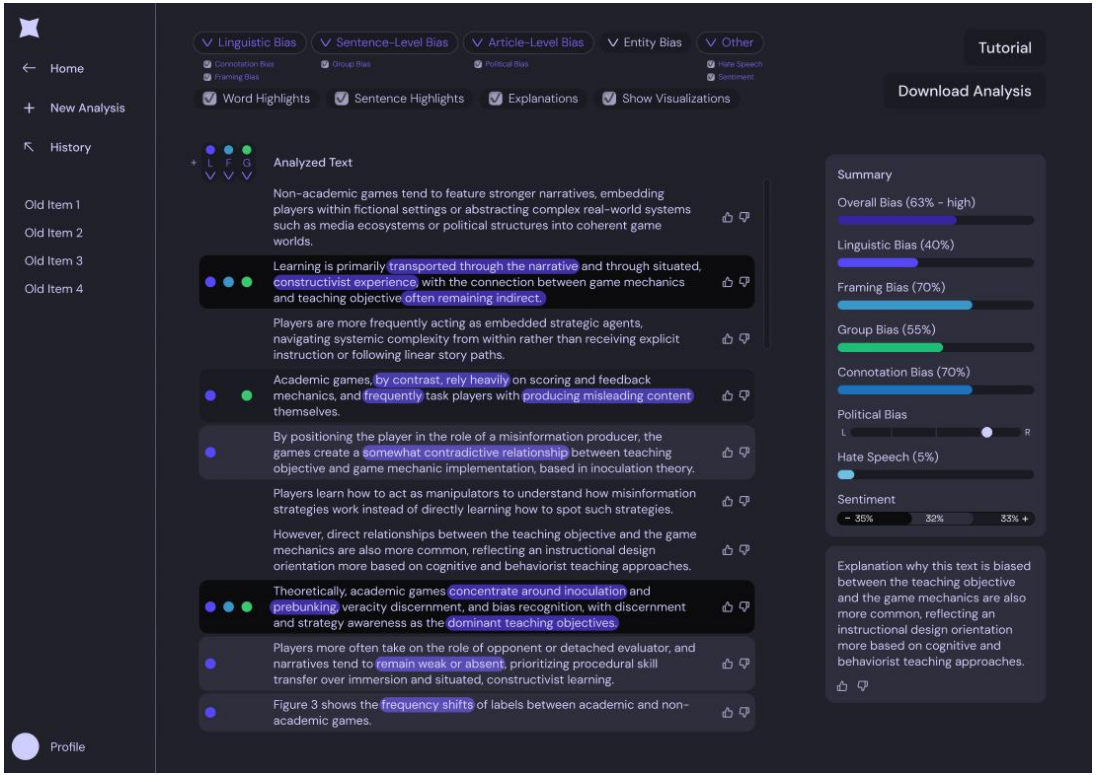


Bias Analysis Research Infrastructure

Explore bias categories and their relationships.

Connect definitions to analysis.

BARI: examining bias in context



Analysis interface

Inspect category-specific highlights and explanations in context.

Interpret outputs using explicit categories and evidence.

Media Bias Group: an interdisciplinary network



Further resources

MBG website



Echizen Lab



Timo’s website



Hanrui’s website

