

Anubrata Das, Ph.D.

anubrata.github.io • anubrata.das[at]utexas.edu

Research Interests

My research spans [interpretability](#), [explainability](#), and [Human-AI collaboration](#). I aim to (a) develop interpretability methods to help generative AI be more fair, transparent and accountable, (b) design post-training approaches that enable AI models to be effective collaborators for end users.

Employment

McCombs School of Business, University of Texas at Austin

Postdoctoral Scholar

Austin, Texas

04/2025 – Present

- Mentors: Maytal Saar-Tsechansky, Liu Leqi
- Project: Building Trustworthy Reasoning Models to aid High-Stakes Decision Makers

Education

University of Texas at Austin

Ph.D., School of Information

Austin, Texas

04/2025

- Dissertation: Towards Human-Centered and Trustworthy Natural Language Processing
- Co-advisors: Matt Lease (iSchool, Computer Science); Junyi Jessy Li (Linguistics)

Indian Institute of Engineering Science and Technology Shibpur

Bachelor of Engineering, Department of Computer Science and Technology

Kolkata, India

06/2015

🏆 Awards and Honors

- **Rising Star in Data Science** 2025
 - Organized by Stanford University, University of California, San Diego, and the University of Chicago
 - **Best Paper Honorable Mention Award (CSCW 2024) (top 4%)** 2024
 - **Diversity & Inclusion Best Student Paper Award** 2019
 - Awarded by the School of Information, University of Texas at Austin
 - **Spot Award - Mu Sigma Inc.** 2016
 - For developing an interactive visualization for stock market as a causal network
 - **Class of 1990 Excellence in Student Leadership Award** 2014
 - Awarded by the Global Alumni Association of BESU (now IIST)
-

Publications [Google Scholar]

* denotes equal contribution

Pre-print

1. [Linear Representation Transferability Hypothesis: Leveraging Small Models to Steer Large Models](#)
Femi Bello, **Anubrata Das**, Fanzhi Zeng, Fangcong Yin, Liu Leqi
preprint arXiv:2506.00653, 2025

Conference and Journal Publications

14. [On Localizing and Deleting Toxic Memories in Large Language Models](#)
Anubrata Das, Manoj Kumar, Ninareh Mehrabi, Anil Ramakrishna, Anna Rumshisky, Kai-Wei Chang, Aram Galstyan, Morteza Ziyadi, Rahul Gupta
NAACL Findings, 2025
13. [Finding Pareto trade-offs in fair and accurate detection of toxic speech](#)
Soumyajit Gupta, Venelin Kovatchev, **Anubrata Das**, Maria De-Arteaga, Matthew Lease
iConference, 2025
12. [Human-centered NLP Fact-checking: Co-Designing with Fact-checkers using Matchmaking for AI](#)
Anubrata Das*, Houjiang Liu*, Alexander Boltz*, Didi Zhou, Daisy Pinaroc, Matthew Lease, Min Kyung Lee
CSCW, 2024

🏆 **Best Paper Honorable Mention (Top 4%)**

11. [The State of Human-centered NLP Technology for Fact-checking](#)
Anubrata Das, Houjiang Liu, Venelin Kovatchev, Matthew Lease
Information Processing and Management (IPM Journal), 2023 (Impact Factor: 6.9)
10. [True or false? Cognitive load when reading COVID-19 news headlines: an eye-tracking study](#)
Li Shi, Nilavra Bhattacharya, **Anubrata Das**, Jacek Gwizdka
ACM SIGIR CHIIR, 2023
9. [ProtoTex: Explaining Model Decisions with Prototype Tensors](#)
Anubrata Das*, Chitrang Gupta*, Venelin Kovatchev, Matthew Lease, Junyi Jessy Li
ACL Main, 2022
(Acceptance rate: 701 of 3378 submissions, 20.8%)
8. [The Need for Human-centered Design in Fact-checking Research](#)
Anubrata Das, Houjiang Liu, Venelin Kovatchev, Matthew Lease
IPM Conference, 2022
7. [Fairness in Information Access Systems](#)
Michael D Ekstrand, **Anubrata Das**, Robin Burke, Fernando Diaz
Foundations and Trends in Information Retrieval (FnTIR), 2022
(100 page monograph; only student co-author)
6. [The Effects of Interactive AI Design on User Behavior: An Eye-tracking Study of Fact-checking COVID-19 Claims](#)
Li Shi, Nilavra Bhattacharya, **Anubrata Das**, Matt Lease, Jacek Gwizdka
ACM SIGIR CHIIR, 2022
5. [Pareto Solutions vs Dataset Optima: Concepts and Methods for Optimizing Competing Objectives with Constraints in Retrieval](#)
Soumyajit Gupta, Gurpreet Singh, **Anubrata Das**, Matthew Lease
ACM SIGIR ICTIR, 2021
4. [Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content](#)

Anubrata Das, Brandon Dang, Matthew Lease
AAAI HCOMP, 2020

3. [Dataset bias: A case study for visual question answering](#)
Anubrata Das, Samreen Anjum, Danna Gurari
ASIS&T, 2019
(**Diversity and Inclusion Student Best Paper Award** by the School of Information, UT Austin)
2. [Interactive information crowdsourcing for disaster management using SMS and Twitter: A research prototype](#)
Anubrata Das, Neeratyoy Mallik, Somprakash Bandyopadhyay, Sipra Das Bit, Jayanta Basak
PerCom Workshops, 2016
1. [Predicting trends in the twitter social network: a machine learning approach](#)
Anubrata Das, Moumita Roy, Soumi Dutta, Saptarshi Ghosh, Asit Kumar Das
Springer SEMCCO, 2014

Workshops / Book Chapters

6. [Proceedings of the 5th Workshop on Trustworthy NLP \(TrustNLP 2025\)](#)
Trista Cao, **Anubrata Das**, Tharindu Kumarage, Yixin Wan, Satyapriya Krishna, Ninareh Mehrabi, Jwala Dhamala, Anil Ramakrishna, Aram Galystan, Anoop Kumar, Rahul Gupta, Kai-Wei Chang
NAACL Workshop Report, 2025
5. [Fairness in recommender systems](#)
Michael D Ekstrand, **Anubrata Das**, Robin Burke, Fernando Diaz
Recommender systems handbook, Springer, 2022
4. [Longhorns at DADC 2022: How many linguists does it take to fool a Question Answering model? A systematic approach to adversarial attacks](#)
Venelin Kovatchev, Trina Chatterjee, Venkata S Govindarajan, Jifan Chen, Eunsol Choi, Gabriella Chronis, **Anubrata Das**, Katrin Erk, Matthew Lease, Junyi Jessy Li, and others
Proceedings of the First Workshop on Dynamic Adversarial Data Collection, 2022
3. [FACTS-IR: fairness, accountability, confidentiality, transparency, and safety in information retrieval](#)
Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, ... **Anubrata Das**, ... and others
SIGIR FACTS-IR Workshop Report, 2021
2. [ExFacto: An Explainable Fact-Checking Tool](#)
Anubrata Das, Sooyong Lee, An Thanh Nguyen, Abhilash Kharosekar, Shankar Krishnan, Saurav Krishnan, Eric Tate, Byron C Wallace, Matthew Lease, and others
Knight Research Network Tool Demonstration Day, 2021
1. [CobWeb: A Research Prototype for Exploring User Bias in Political Fact-Checking](#)
Anubrata Das, Kunjan Mehta, Matthew Lease
SIGIR FACTS-IR Workshop, 2019

Technical Reports

2. [The Case for Claim Difficulty Assessment in Automatic Fact Checking](#)
Prakhar Singh, **Anubrata Das**, Junyi Jessy Li, Matthew Lease
arXiv preprint arXiv:2109.09689, 2021
1. [A Conceptual Framework for Evaluating Fairness in Search](#)
Anubrata Das, Matthew Lease
arXiv preprint arXiv:1907.09328, 2019

Presentations

Invited Talks

Information Assurance in Clinical LLMs through Unlearning

- MISQ Workshop on *Artificial Intelligence-Information Assurance Nexus: The Future of Information Systems Security, Privacy, and Quality*, 07/11/2025

Developing Language Technologies to Complement Human Capabilities

- Microsoft Research FATE Group, New York City, 02/16/2024
- McCombs School of Business, University of Texas at Austin, 02/12/2024

ProtoTex: Explaining Model Decisions with Prototype Tensors

- Research Colloquium, UT Austin, iSchool, 09/20/2022
- iSchools European Doctoral Seminar Series, 09/16/2022
- Amazon Science Clarify Team, 05/17/2022
- NEC Laboratories Europe, 06/09/2022

Commercial Content Moderation and Psychological Well-Being

- TxHCI - A seminar organized by HCI Researchers across Universities in Texas, 10/02/2020
- Amazon AWS Science, 10/14/2020
- Amazon Human-in-the-loop (HILL) services team, 10/23/2020
- ACM SIGCHI Mumbai Chapter, 26th Meet, 08/28/2021

Conference Presentations

On Localizing and Deleting Toxic Memories in Large Language Models. NAACL. 2025. Albuquerque, New Mexico.

Finding pareto trade-offs in fair and accurate detection of toxic speech. iConference. 2025. Bloomington, Indiana.

ProtoTex: Explaining Model Decisions with Prototype Tensors. ACL. May 2022. Dublin, Ireland.

You are what you tweet: Profiling users by past tweets to improve hate speech detection. iConference. March 2022. Virtual Conference.

Exfacto: An explainable fact-checking tool. Knight Research Network Tool Demonstration Day, 2021. Virtual Conference.

Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. AAAI HCOMP 2020. Virtual Conference.

Dataset bias: A case study for visual question answering. ASIS&T 2019. Melbourne, Australia.

CobWeb: A Research Prototype for Exploring User Bias in Political Fact-Checking. ACM SIGIR Workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval (FACTS-IR), 2019. Paris, France.

Local Presentations

ProtoTex: Explaining Model Decisions with Prototype Layers. Research Colloquium, School of Information, University of Texas at Austin. November 2021. Lightning Talk.

ProtoBART: Explaining Model Decisions with Prototype Layers. TACCSTER: TACC Symposium for Texas Researchers. September 2021. Lightning Talk.

Funding

- Student Professional Development Award for Attending iConference 2025. Award Amount \$750.
 - Evaluating Example-based Explainable Models in Large Language Models. **Amazon AWS Cloud Credit for Research**. Funding period: 11/30/2022 - 11/30/2023. **26,000 USD** (AWS Service Credits).
 - Student Professional Development Award for Attending ACL 2022. Award Amount \$1000.
 - **UT Good Systems Grand Challenge** — Graduate Student Grant Proposal. **Anubrata Das**, Chenyan Jia, Shivam Garg. Supervisor: Min Kyung Lee. *Designing algorithmic nudge to reduce inadvertent COVID-19 misinformation sharing on social media*. Awarded - USD 7000.
-

Service

Workshop Organization

- TrustNLP 2025 at NAACL 2025

Program Committees and Reviewing

- ICLR 2026; CHI 2026; SIGIR Algorithmic Bias Workshop 2025; CoLM 2024, 2025; ACM FAccT 2025; ACL Rolling Review 2022, 2023, 2024; AAI AIES 2022; BlackboxNLP Workshop 2022; CHI 2021, 2022; CSCW 2021, 2022, 2023; The Web Conference 2021; Annual Meeting of the Association for Information Science and Technology: 2019, 2020; Information Processing and Management Journal

Conference Volunteer

- NAACL 2025; ACL 2022; CSCW 2019

University Committees

- Assistant Professor Hiring Committee 2020-2021
 - Doctoral Studies Committee, School of Information, 2019-2020
-

Teaching and Mentoring

- Teaching Assistant *Fall 2020*
 - INF385T.3 / CS395T: Human Computation and Crowdsourcing by Dr. Matt Lease
 - Three tutorials on Amazon Sagemaker Ground Truth for collecting data annotations
 - Co-Supervising student research with Dr. Matt Lease *01/2022 - 06/2022*
 - Undergraduate thesis on Active Learning with Natural Language Rationales
 - Featured in UT Austin, College of Natural Sciences News
 - Co-Supervising undergraduate research group with Dr. Matt Lease *06/2020 - 08/2021*
 - A group of ten students
 - Working on fact-checking using NLP and Human-computation methods
-

Research Internships

Cisco Research, Responsible AI
Research Intern

New York City, NY
09/2023 – 12/2023

- Mentors: Ali Payani, Jayanth Srinivasa

Amazon Nova Responsible AI
Research Intern

New York City, NY
06/2023 – 09/2023

- Mentors: Kai-Wei Chang, Anna Rumshisky, Aram Galstyan, Manoj Kumar, Ninareh Mehrabi, Anil Ramakrishna, Rahul Gupta

Max Planck Institute of Informatics
Research Intern, Databases and Information Systems Group

Saarbrücken, Germany
06/2019 – 08/2019

- Mentor: Gerhard Weikum
- Project: *Systematic discovery of bias: A case study on Airbnb Listings*

Indian Institute of Management Calcutta
Undergraduate Research Assistant, Management Information Systems Group

Kolkata, India
10/2012 – 09/2015

- Mentor: Somprakash Bandyopadhyay

Indian Institute of Technology Kharagpur
Research Intern, Complex Networks and Research Group

West Bengal, India
05/2013 – 07/2013

- Mentor: Saptarshi Ghosh

Industry Experience

Microsoft
Software Engineer

Hyderabad, India
11/2016 – 07/2018

- Built and maintained a marketing management tool for Microsoft Universal Store

Mu Sigma
Decision Scientist

Bangalore, India
08/2015 – 10/2016

- Design and build research prototypes for algorithmic trading using machine learning

Skills

Research Methodologies: Deep Learning, Large Language Models (LLMs), Foundation Models, Post-training of LLMs, Interpretability, Co-design, Human-AI Interaction

Programming Languages: Python, Javascript

Technologies: Pytorch, Huggingface Transformers, Scikit-Learn, NLTK, SciPy, NumPy, Git

Survey Tools: Qualtrics

Crowdsourcing: Amazon Mechanical Turk, AWS Sagemaker Ground Truth, Prolific

Languages: Fluent in English and Bengali, Knowledge of Hindi
