Anubrata Das, Ph.D.

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

Research Interests

My research integrates Artificial Intelligence (specifically Natural Language Processing) and Human–Computer Interaction with the primary goal to develop trustworthy AI that augments human experts. I develop model-centric approaches to interpretability, explainability, and fairness, and employ HCI methods to engage with stakeholders at different stages of the AI development pipeline.

Education

University of Texas at Austin — **Austin, Texas**
Ph.D., School of Information — *05/2025*

- 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 — **Kolkata, India**
Bachelor of Engineering, Department of Computer Science and Technology — *06/2015*

Employment

McCombs School of Business, University of Texas at Austin — **Austin, Texas**
Postdoctoral Scholar — *06/2025 – Present*
Graduate Research Assistant — *09/2024 – 05/2025*

- Mentors: Maytal Saar-Tsechansky, Liu Leqi

AI & HCC Lab, University of Texas at Austin — **Austin, Texas**
Graduate Research Assistant — *08/2018 – 05/2024*

- Mentors: Matthew Lease, Junyi Jessy Li

Cisco Research, Responsible AI — **New York City, NY**
Research Intern — *09/2023 – 12/2023*

- Mentors: Ali Payani, Jayanth Srinivasa

Amazon Alexa Responsible AI — **New York City, NY**
Research Intern — *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 — **Saarbrücken, Germany**
Research Intern, Databases and Information Systems Group — *06/2019 – 08/2019*

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

Microsoft — **Hyderabad, India**
Software Engineer — *01/2016 – 07/2018*

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

Mu Sigma   **Bangalore, India**
Decision Scientist   *08/2015 – 10/2016*

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

Indian Institute of Management Calcutta   **Kolkata, India**
Undergraduate Research Assistant, Management Information Systems Group   *10/2012 – 09/2015*

– Mentor: Somprakash Bandyopadhyay

Indian Institute of Technology Kharagpur   **West Bengal, India**
Research Intern, Complex Networks and Research Group   *05/2013 – 07/2013*

– Mentor: Saptarshi Ghosh

Publications [Google Scholar]

* denotes equal contribution

Work in progress

2. Femi Bello, **Das, Anubrata**, Fanzhi Zeng, Fangcong Yin, and Liu Leqi. Linear representation transferability hypothesis: Leveraging small models to steer large models. *preprint arXiv:2506.00653*, 2025. Under review at NeurIPS 2025

1. **Das, Anubrata**, Liu Leqi, and Maytal Saar-Tsechansky. Information Assurance in Clinical LLMs with Unlearning. *Manuscript under preperation for MISQ Special Issue*, 2025

Conferences / Journals

13. **Das, Anubrata**, Manoj Kumar, Ninareh Mehrabi, Anil Ramakrishna, Anna Rumshisky, Kai-Wei Chang, Aram Galstyan, Morteza Ziyadi, and Rahul Gupta. On Localizing and Deleting Toxic Memories in Large Language Models. *Findings of the Association for Computational Linguistics*, (**NAACL Findings**), 2025

12. Soumyajit Gupta, Venelin Kovatchev, **Das, Anubrata**, Maria De-Arteaga, and Matthew Lease. Finding pareto trade-offs in fair and accurate detection of toxic speech. *Information Research an international electronic journal*, (**iConference**), 2025

11. **Das, Anubrata***, Houjiang Liu*, Alexander Boltz*, Didi Zhou, Daisy Pinaroc, Matthew Lease, and Min Kyung Lee. Human-centered NLP Fact-checking: Co-Designing with Fact-checkers using Matchmaking for AI. *Proceedings of the ACM on Human-Computer Interaction*, (**CSCW**), 2024. 🏆 Best Paper Honorable Mention (Top 3%)

10. **Das, Anubrata**, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. The state of human-centered nlp technology for fact-checking. *Information Processing & Management, Special Issue on Machine and Human Factors in Misinformation Management*, (**IPM**), 2023. (Impact Factor: 6.222)

9. Li Shi, Nilavra Bhattacharya, **Das, Anubrata**, and Jacek Gwizdka. True or false? cognitive load when reading covid-19 news headlines: an eye-tracking study. *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, (**ACM SIGIR CHIIR**), 2023

8. Li Shi, Nilavra Bhattacharya, **Das, Anubrata**, Matt Lease, and Jacek Gwizdka. The effects of interactive ai design on user behavior: An eye-tracking study of fact-checking covid-19 claims. *ACM SIGIR Conference on Human Information Interaction and Retrieval*, (**ACM SIGIR CHIIR**), 2022

7. **Das, Anubrata**, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. The need for human-centered design in fact-checking research. In *Information Processing & Management Conference* (**IPM Conference**), 2022

6. Michael D Ekstrand, **Das, Anubrata**, Robin Burke, and Fernando Diaz. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, (**FnTIR**), 2022. (**100 page monograph; only student co-author**)

5. **Das, Anubrata***, Gupta, Chitrank*, Venelin Kovatchev, Matthew Lease, and Junyi Jessy Li. ProtoTEx: Explaining Model Decisions with Prototype Tensors. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, (**ACL Main**), 2022

4. Soumyajit Gupta, Gurpreet Singh, **Das, Anubrata**, and Matthew Lease. Pareto solutions vs dataset optima: Concepts and methods for optimizing competing objectives with constraints in retrieval. *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, (**ACM SIGIR ICTIR**), 2021

3. **Das, Anubrata**, Brandon Dang, and Matthew Lease. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, (**AAAI HCOMP**), 2020

2. **Das, Anubrata**, Samreen Anjum, and Danna Gurari. Dataset bias: A case study for visual question answering. *Proceedings of the Association for Information Science and Technology*, (**ASIS&T**), 2019. (**Diversity and Inclusion Student Best Paper Award** by the School of Information, UT Austin)

1. **Das, Anubrata**, Moumita Roy, Soumi Dutta, Saptarshi Ghosh, and Asit Kumar Das. Predicting trends in the twitter social network: a machine learning approach. *International Conference on Swarm, Evolutionary, and Memetic Computing*, (**Springer SEMCCO**), 2014

Workshops / Book Chapters

6. Trista Cao, **Das, Anubrata**, Tharindu Kumarage, Yixin Wan, Satyapriya Krishna, Ninareh Mehrabi, Jwala Dhamala, Anil Ramakrishna, Aram Galystan, Anoop Kumar, Rahul Gupta, and Kai-Wei Chang, editors. *Proceedings of the 5th Workshop on Trustworthy NLP* (*TrustNLP 2025*). Association for Computational Linguistics, 2025. (**NAACL Workshop Report**)

5. Michael D Ekstrand, **Das, Anubrata**, Robin Burke, and Fernando Diaz. Fairness in recommender systems. In *Recommender systems handbook*, pages 679–707. Springer, 2022

4. **Das, Anubrata**, Kunjan Mehta, and Matthew Lease. 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*, (**SIGIR FACTS-IR Workshop**), 2019

3. Venelin Kovatchev, Trina Chatterjee, Venkata S Govindarajan, Jifan Chen, Eunsol Choi, Gabriella Chronis, **Das, Anubrata**, Katrin Erk, Matthew Lease, Junyi Jessy Li, et al. longhorns at dadc 2022: How many linguists does it take to fool a question answering model? a systematic approach to adversarial attacks. In *Proceedings of the First Workshop on Dynamic Adversarial Data Collection*, 2022

2. Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, ... **Das, Anubrata**, et al. Facts-ir: fairness, accountability, confidentiality, transparency, and safety in information retrieval. *ACM SIGIR Forum*, (**SIGIR FACTS-IR Workshop Report**), 2021

1. **Das, Anubrata**, Neeratyoy Mallik, Somprakash Bandyopadhyay, Sipra Das Bit, and Jayanta Basak. Interactive information crowdsourcing for disaster management using sms and twitter: A research prototype. *2016 IEEE International Conference on Pervasive Computing and Communication Workshops*, (**PerCom Workshops**), 2016

Technical Reports

2. Prakhar Singh, **Das, Anubrata**, Junyi Jessy Li, and Matthew Lease. The case for claim difficulty assessment in automatic fact checking. *arXiv preprint arXiv:2109.09689*, 2021

1. **Das, Anubrata** and Matthew Lease. A conceptual framework for evaluating fairness in search. *arXiv preprint arXiv:1907.09328*, 2019

Presentations

Invited Talks

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 InformationRetrieval (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

- 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).
- **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.

🏆 Awards and Honors

- **Best Paper Honorable Mention Award (CSCW 2024) (top 3%)**   2024
- Annual Diversity & Inclusion Best Student Paper Award   2019
    - Das et al., ASIS&T 2019 (see the pulication section above)
    - Awarded by the School of Information, University of Texas at Austin
- Spot Award - Mu Sigma Inc.   2016
    - Awarded by the Innovation and Development Team
    - Interactive visualization for Stock Market as a network
- Class of 1990 Award: Excellence in Leadership   02/2014
    - Awarded by the Global Alumni Association of BESU (now IIEST)

Service

Workshop Organization

- TrustNLP 2025 at NAACL 2025

Program Committees and Reviewing

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

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

---

Skills

Research Methodologies: Experimental Design, User Study, Crowdsourcing,
  Natural Language Processing, Machine Learning, Inferential Statistics
Programming Languages: Python, R, JavaScript, SQL
Technologies: Flask, Pytorch, Scikit-Learn, NLTK, SciPy, NumPy, Git
Survey Tools: Qualtrics
Crowdsourcing: Amazon Mechnaical Turk, AWS Sagemaker Ground Truth, Prolific
Languages: Fluent in English and Bengali, Knowledge of Hindi