Manish Nagireddy

manish.nagireddy@gmail.com | mnagireddy | @mnagired

Background

Manish Nagireddy is a Research Software Engineer at IBM Research and the MIT-IBM Watson AI Lab. His current work focuses on use-case centered algorithmic auditing and evaluation, in the context of large language models. Some representative projects include:

- **developing efficient and provably capable guardrails for unsafe input and output detection**
- **deploying inclusive and participatory red-teaming sessions for targeted model improvement**

Education

Carnegie Mellon University *Pittsburgh, PA*
B.S. Statistics and Machine Learning, Additional Major in Computer Science *2019-2022*

Professional Experience

2023- **Research Software Engineer**, IBM Research, MIT-IBM Watson AI Lab
2022 **Undergraduate Research Intern**, IBM Research

Publications

Published

Inkit Padhi, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, **Manish Nagireddy**, Pierre Dognin, Kush R Varshney *Value Alignment from Unstructured Text.* Workshop on Pluralistic Alignment @ NeurIPS '24

Clara Higuera Cabañes, Ryo Iwaki, Beñat San Sebastian, Rosario Uceda Sosa, **Manish Nagireddy**, Hiroshi Kanayama, Mikio Takeuchi, Gakuto Kurata, Karthikeyan Natesan Ramamurthy *SocialStigmaQA Spanish and Japanese - Towards Multicultural Adaptation of Social Bias Benchmarks.* Socially Responsible Language Modelling Research (SoLaR) Workshop @ NeurIPS '24

Inkit Padhi, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, **Manish Nagireddy**, Pierre Dognin, Kush R Varshney *Value Alignment from Unstructured Text.* Conference on Empirical Methods in Natural Language Processing (EMNLP) Industry Track '24

Erik Miehling, **Manish Nagireddy**, Prasanna Sattigeri, Elizabeth M Daly, David Piorkowski, John T Richards *Language Models in Dialogue: Conversational Maxims for Human-AI Interactions.* Findings of Empirical Methods in Natural Language Processing (EMNLP) '24

Swapnaja Achintalwar, Ioana Baldini, Djallel Bouneffouf, Joan Byamugisha, Maria Chang, Pierre Dognin, Eitan Farchi, Ndivhuwo Makondo, Aleksandra Mojsilović, **Manish Nagireddy**, Karthikeyan Natesan Ramamurthy, Inkit Padhi, Orna Raz, Jesus Rios, Prasanna Sattigeri, Moninder Singh, Siphiwe Thwala, Rosario A Uceda-Sosa, Kush R Varshney *Alignment Studio: Aligning Large Language Models to Particular Contextual Regulations.* IEEE Internet Computing 2024

Manish Nagireddy, Bernat Guillén Pegueroles, Ioana Baldini *DARE to Diversify: DAta Driven and Diverse LLM REd Teaming.* Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) '24

Inkit Padhi, Pierre Dognin, Jesus Rios, Ronny Luss, Swapnaja Achintalwar, Matthew D Riemer, Miao Liu, Prasanna Sattigeri, **Manish Nagireddy**, Kush R Varshney, Djallel Bouneffouf *ComVas: Contextual Moral Values Alignment System* Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI) Demo Track. '24

Brandon Dominique, David Piorkowski, **Manish Nagireddy**, Ioana Baldini *Prompt Templates: A Methodology for Improving Manual Red Teaming Performance* ACM CHI Conference on Human Factors in Computing Systems '24, Workshop on Human-centered Evaluation and Auditing of Language Models

Manish Nagireddy, Lamogha Chiazor, Moninder Singh, Ioana Baldini *SocialStigmaQA: A Benchmark to Uncover Stigma Amplification in Generative Language Models* Proceedings of the 2024 AAAI Conference on Artificial Intelligence

Hussein Mozannar, Valerie Chen, Dennis Wei, Prasanna Sattigeri, **Manish Nagireddy**, Subhro Das, Ameet Talwalkar, David Sontag *Simulating Iterative Human-AI Interaction in Programming with LLMs* NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following

Soumya Ghosh, Prasanna Sattigeri, Inkit Padhi, **Manish Nagireddy**, Jie Chen *Influence Based Approaches to Algorithmic Fairness: A Closer Look* NeurIPS 2023 Workshop on XAI in Action: Past, Present, and Future Applications

Manish Nagireddy, Moninder Singh, Samuel C Hoffman, Evaline Ju, Karthikeyan Natesan Ramamurthy, Kush R Varshney *Function Composition in Trustworthy Machine Learning: Implementation Choices, Insights, and Questions.* Preprint.

Nil-Jana Akpinar, **Manish Nagireddy**, Logan Stapleton, Hao-Fei Cheng, Haiyi Zhu, Steven Wu, Hoda Heidari *A sandbox tool to bias (Stress)-test fairness algorithms.* ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO) '22 Poster

Wesley Hanwen Deng, **Manish Nagireddy**, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, Haiyi Zhu. *Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits.* 2022 ACM Conference on Fairness, Accountability, and Transparency

Wesley Hanwen Deng, **Manish Nagireddy**, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, Haiyi Zhu. *Fairness in Practice: Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits* Human Centered AI workshop @ NeurIPS 2021

Under Review

Swapnaja Achintalwar, Adriana Alvarado Garcia, Ateret Anaby-Tavor, Ioana Baldini, Sara E Berger, Bishwaranjan Bhattacharjee, Djallel Bouneffouf, Subhajit Chaudhury, Pin-Yu Chen, Lamogha Chiazor, Elizabeth M Daly, Rogério Abreu de Paula, Pierre Dognin, Eitan Farchi, Soumya Ghosh, Michael Hind, Raya Horesh, George Kour, Ja Young Lee, Erik Miehling, Keerthiram Murugesan, **Manish Nagireddy**, Inkit Padhi, David Piorkowski, Ambrish Rawat, Orna Raz, Prasanna Sattigeri, Hendrik Strobelt, Sarathkrishna Swaminathan, Christoph Tillmann, Aashka Trivedi, Kush R Varshney, Dennis Wei, Shalisha Witherspooon, Marcel Zalmanovici *Detectors for safe and reliable llms: Implementations, uses, and limitations.* Under Review

Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, **Manish Nagireddy**, Prasanna Sattigeri, Ameet Talwalkar, David Sontag *The RealHumanEval: Evaluating Large Language Models' Abilities to Support Programmers.* Under Review

Lucas Monteiro Paes, Dennis Wei, Hyo Jin Do, Hendrik Strobelt, Ronny Luss, Amit Dhurandhar, **Manish Nagireddy**, Karthikeyan Natesan Ramamurthy, Prasanna Sattigeri, Werner Geyer, Soumya Ghosh *Multi-Level Explanations for Generative Language Models.* Under Review

Pierre Dognin, Jesus Rios, Ronny Luss, Inkit Padhi, Matthew D Riemer, Miao Liu, Prasanna Sattigeri, **Manish Nagireddy**, Kush R Varshney, Djallel Bouneffouf *Contextual Moral Value Alignment Through Context-Based Aggregation.* Under Review

Manish Nagireddy, Inkit Padhi, Soumya Ghosh, Prasanna Sattigeri *When in Doubt, Cascade: Towards Building Efficient and Capable Guardrails.* Under Review

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, **Manish Nagireddy**, Amit Dhurandhar *Programming Refusal with Conditional Activation Steering.* Under Review

Awards, Fellowships, & Grants

2024	**Technical Collaboration Achievement Program (TCAP)**, IBM Research	*$ 2,000*
2020	**Siegel Public Interest Technology Summer Fellowship**, Princeton University	*$ 5,000*

Presentations

Invited Talks

Sept 2024. *Operationalizing Trustworthy AI*. Invited talk to Honors Seminar on Algorithmic Bias @ SUNY Plattsburgh. Virtual

Sept 2024. *Operationalizing Trustworthy AI*. Invited talk to online course on AI Governance. Virtual

August 2024. *Trustworthy LLMs: Detection and Red-Teaming*. Ethical AI Workshop @ KDD '24. Barcelona, Spain

June 2024. *Operationalizing Trustworthy AI*. Invited talk to online course on AI Governance. Virtual

June 2024. *Red-Teaming With Diverse Lived Experiences*. Presentation to IBM Accelerate cohort. Virtual

April 2024. *LLM Alignment and Governance*. Wells Fargo Innovation Showcase. San Francisco, California

April 2024. *Trustworthy (Enterprise) Foundation Models*. Workshop on Standards for Foundation Models hosted by Salesforce. San Francisco, California

April 2024. *Trustworthy (Enterprise) Foundation Models*. Invited talk to Georgetown Center for Security and Emerging Technology. Virtual

April 2024. *Trustworthy (Enterprise) Foundation Models*. Invited talk to LinkedIn Fairness Research team. Virtual

February 2024. *SocialStigmaQA: A Benchmark to Uncover Stigma Amplification in Generative Language Models*. Invited briefing, TrustML Workshop @ UBC. Vancouver, Canada

July 2022. *Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits*. Invited talk to IBM Trustworthy Machine Learning team. Virtual

March 2021. *Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits*. Invited talk at Fairlearn developer team meeting. Virtual

Conference Demos and Tutorials

Manish Nagireddy, Michael Feffer, Ioana Baldini. 2025. DAMAGeR: Deploying Automatic and Manual Approaches to GenAI Red-teaming. Tutorial @ KDD '25, Philadelphia, USA

Manish Nagireddy. 2024. EvaluLLM: LLM assisted evaluation of generative outputs @ AAAI '24, Vancouver, British Columbia, Canada

Manish Nagireddy, Bernat Guillen, Ioana Baldini. 2024. DARE to Diversify: DAta Driven and Diverse LLM REd Teaming. Hands-On Tutorial @ KDD '24, Barcelona, Spain

Manish Nagireddy, Anupama Murthi, Samuel Hoffman, Karthikeyan Natesan Ramamurthy. 2023. AI Fairness 360. Open Source Day @ Grace Hopper Conference '23, Virtual

Manish Nagireddy, Anupama Murthi, Samuel Hoffman, Karthikeyan Natesan Ramamurthy. 2022. AI Fairness 360. Open Source Day @ Grace Hopper Conference '22, Virtual

Mentoring

2024-	**Katherine He**, Yale University, Informal Mentorship
2024	**Bruce Lee**, University of Pennsylvania, IBM Research Internship
2024	**Kweku Kwegyir-Aggrey**, Brown University, IBM Research Internship
2024	**Nilton Cesar Rojas Vales**, University National of Engineering, AAAI Undergrad Consortium
2023	**Shrey Jain**, Rensselaer Polytechnic Institute, IBM Research Externship
2023	**Clare Arrington**, Rensselaer Polytechnic Institute, IBM Research Externship
2023	**Brandon Dominique**, Northeastern University, IBM Research Internship
2023	**Lucas Paes**, Harvard University, IBM Research Internship

Service

Peer Review

ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT), '23, '24
Association for the Advancement of Artificial Intelligence (AAAI), '23, '24, '25
ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), '24
Annual Conference on Neural Information Processing Systems (NeurIPS), '24

Professional Memberships

Association for Computing Machinery (ACM)
Association for the Advancement of Artificial Intelligence (AAAI)

Other

IBM Data Science and AI for Social Impact NYC Workshop 2023 Q1