Ashwin Kumar

ashwinkumar@wustl.edu | kumar-ashwin.github.io | ashwinkumar97 | Google Scholar
St. Louis, MO, USA

Career Summary

- Computer Science Ph.D. candidate working on improving fairness and transparency in AI systems like Multi-Agent Reinforcement Learning and Large Language Models. Recently, my research has covered detecting and mitigating bias in RLHF pipelines, and learning long-term fairness in deep MARL with resource constraints.

Education

- **Washington University in St. Louis** *July 2025 (Expected)*
  Ph.D., Computer Science (Current GPA: 4.0) St. Louis, MO, USA
- **Washington University in St. Louis** *May 2022*
  Master of Science, Computer Science (GPA: 4.0) St. Louis, MO, USA
- **Shiv Nadar University** *May 2019*
  Bachelor of Engineering, Mechanical Engineering (CGPA 9.81, Gold Medalist) Dadri, India

Experience

- **Research Scientist Intern, Meta Platforms, Inc.** *Aug 2023 – Jan 2024*
  Metrics and Mitigations for Prefix Bias in LLM Reward Models Menlo Park, CA
  - Engineered techniques to detect and mitigate demographic bias in preference models used in **RLHF finetuning**.
  - Designed an attack to reveal prefix bias in preference models based on **various LLM architectures** (llama, vicuna, OPT, flan, GPTJ) trained on popular preference datasets. Developed metrics to detect and quantify this bias, showing **susceptibility to preference switching** in up to **98%** of the dataset using **short prefixes (<4 words**).
  - Designed a data augmentation technique to mitigate prefix bias, **reducing the bias** in model accuracy by over **85%**.
  - Published at FAccT 2025.
- **Research Scientist Intern, Meta Platforms, Inc.** *May 2022 – Aug 2022*
  Bias Bootstrapping and the Effects of Feedback in Self-Guided Dataset Sampling-based Models New York, NY
  - Designed methods for **detection and evaluation of bias feedback loops** in content safety classifiers which use Model Assisted Sampling to send data points for human review.
  - Quantified the propagation of bias to future models, showing up to **100% loss in accuracy** when using stratified sampling and devised sampling strategies to mitigate bias bootstrapping and accelerate recovery, **improving group detection by 57%**.
  - Worked with multiple product teams to identify susceptible sampling techniques and generalize experiments.

Research

- **Reducing Group Disparity in Next-location Prediction using Adaptive Sampling**
  May 2024 – Ongoing. Under submission
  - Analyzed large-scale mobility data with **5M users** and **500k unique locations** (businesses) for algorithmic bias, using census-derived group grounding to show distribution disparities. Also developed **Size-Aware K-Means**, a clustering algorithm using Lagrangian penalties to **enforce group size constraints** for grounded analysis.
  - Designed a fair data sampling algorithm with tunable tradeoffs for efficiency vs group disparity oriented sampling, **reducing total Demographic Parity violations** by **30%** while maintaining efficiency during iterative training.
- **Improving Fairness in Multi-Agent Resource Allocation**
  Jan 2021 – Dec 2024. AAMAS 25, AASG 25, RLSW 24, AASG 23, ICAPS 23, ATT 22, ICAPS 22
  - Used properties of Integer Linear Programs to design an efficient online method to improve fairness in **two-sided matching** systems like ridesharing and homelessness resource matching.
  - Demonstrated the ability of fairness incentives to improve utility in addition to improving fairness.
  - Developed algorithms for learning fair-efficient behavior by framing repeated resource allocation as a multi-agent RL problem under resource constraints.
  - Designed resource-constrained environments and implemented custom **multi-agent RL algorithms** based on DDQN and MAPPO which allow flexible tradeoffs in utility and fairness, **Pareto-dominating** existing fair multi-agent RL methods.
- **Explainable AI Planning and Human-AI Interaction**
  Mar 2020 – Dec 2023. KR 24, JAIR 22, ICAPS 22, XAIP 21, ICAPS 21
  - Demonstrated that visualization-based interfaces **improve explanation comprehension** in users by **11%** through a user study comparing text explanations to an abstraction-based visualization. Designed user studies with the visualization system for classical and hybrid planning with propositional and first-order logic-based explanations.
  - Designed and tested an **argumentation framework** for interactive **dialogue-based explanations** based on communication rules and differing mental models between the user and agent, simulating the **theory of mind**.

Publications

1. **Detecting Prefix Bias in LLM-based Reward Models.**
   Ashwin Kumar, Yuzi He, Aram H. Markosyan, Bobbie Chern, and Imanol Arrieta-Ibarra. FAccT 2025.
2. **Remember, but also, Forget: Bridging Myopic and Perfect Recall Fairness with Past-Discounting**
   Ashwin Kumar, and William Yeoh. Autonomous Agents for Social Good (AASG) 2025.
3. **DECAF: Learning to be Fair in Multi-Agent Resource Allocation.**
   Ashwin Kumar, and William Yeoh. AAMAS 2025 (Extended Abstract); RL Safety Workshop 2024.
4. **Disproportionate Energy Disruptions Afflicted Rural Hispanic Households During Winter Storm URI.**
   Ashwin Kumar, Tyler H. Ruggles, and Edgar Virgüez. In *Environmental Research: Energy* (Perspective), 2024(3).
5. **Dialectical Reconciliation via Structured Argumentative Dialogues.**
   Stylianos L. Vasileiou, **Ashwin Kumar**, William Yeoh, Tran Cao Son, and Francesca Toni. KR 2024.
6. **Using Simple Incentives to Improve Two-Sided Fairness in Ridesharing Systems.**
   Ashwin Kumar, Yevgeniy Vorobeychik, and William Yeoh. ICAPS 2023.
7. **Fairness in Scarce Societal Resource Allocation: A Case Study in Homelessness Applications.**
   Ashwin Kumar, and William Yeoh. Autonomous Agents for Social Good (AASG) 2023.
8. **Improving Zonal Fairness While Maintaining Efficiency in Rideshare Matching.**
   Ashwin Kumar, Yevgeniy Vorobeychik, and William Yeoh. ATT-22 2022.
9. **A Logic-based Explanation Generation Framework for Classical and Hybrid Planning Problems.**
   Stylianos L. Vasileiou, William Yeoh, Tran Cao Son, **Ashwin Kumar**, Michael Cashmore, and Daniele Magazzeni. *JAIR*, Vol. 73, 2022.
10. **VizXP: A Visualization Framework for Conveying Explanations to Users in Model Reconciliation Problems.**
    Ashwin Kumar, Stylianos L. Vasileiou, Melanie Bancilhon, Alvitta Ottley, and William Yeoh. ICAPS 2022; XAIP 2021.
11. **FairVizARD: A Visualization System for Assessing Fairness of Ride-Sharing Matching Algorithms.**
    Ashwin Kumar, Sanket Shah, Meghna Lowalekar, Pradeep Varakantham, Alvitta Ottley, and William Yeoh. ICAPS 2021 (Demo).

Technical Skills

- **Programming Languages:** Python, C++, JavaScript
- **Libraries and Frameworks:** PyTorch, Transformers, Langchain
- **Skills:** Deep Reinforcement Learning, Generative AI, RLHF, Machine Learning, Natural Language Processing, Constraint Optimization
- **Relevant Courses:** Data Structures and Algorithms; Advanced Algorithms; Machine Learning; Bayesian Methods in ML; Artificial Intelligence; Adversarial Methods in ML; Human-in-the-Loop Computation; Advanced Visualization

Awards

- Recipient of the Dean's List award for academic excellence (2017, 2018).
- Recipient of the university-wide gold medal for highest academic performance (2019)

Teaching

- Assistant in Instruction, Introduction to AI (Spring 2022)
- Guest Lecturer, Introduction to AI (Spring 2022)
- Student Teacher (Data Structures and Intro to CS) at SNU under the Learning and Academic Support Centre (LASC) program. (2017-18)

Program Committee Member and Reviewer

AAMAS 2025; AIES 2025; HAXP Workshop 2021, 2023, 2024; ICAPS 2024, 2025; ICML 2025; IJCAI 2024; NeurIPS 2024; XAI 2023;