

CAREER SUMMARY

- Computer Science Ph.D. candidate, working in fairness and explainable AI. My research focus is on improving fairness and transparency in AI systems including Reinforcement Learning and Large Language Models.
- Skilled in modeling complex problems and designing algorithms to learn from data.

EDUCATION

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

TECHNICAL SKILLS

- **Programming Languages/ Frameworks:** Python, C/C++, JavaScript, MATLAB, SQL, PyTorch, TensorFlow, Transformers, Gurobi.
- **Skills:** Deep Reinforcement Learning, Machine Learning, Natural Language Processing, Constraint Optimization
- **Relevant courses:** Data Structures and Algorithms, Advanced Algorithms, Machine learning, Bayesian methods in Machine Learning, Artificial Intelligence, Adversarial Methods in Machine Learning, Human-in-the-Loop Computation, Advanced Visualization

PROFESSIONAL AND ACADEMIC WORK EXPERIENCE

Research Scientist Intern, Meta Platforms, Inc., Menlo Park, CA **Aug 23 –Jan 24**

- 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.
- Designed a **data augmentation** technique to mitigate prefix bias, **reducing the bias** in model accuracy by over **85%**.

Research Scientist Intern, Meta Platforms, Inc., New York, NY **May 22 –Aug 22**

- Designed methods for **detection of bias feedback loops** in content 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 safety by 50%**.
- Worked with multiple product teams to identify issues with different sampling techniques and generalize experiments.

Assistant in Instruction, Department of Computer Science, Washington University in St. Louis **Jan 22 –May 22**

- Assisted over **130 students** with coursework and homework relating to AI concepts and algorithms.
- Graded assignments and exams and delivered guest lectures on Reinforcement Learning and conducted problem-solving sessions on Logic and MDPs.

ACADEMIC PROJECTS

Improving Fairness in Transformer-based Next-location Prediction **May 24 – Ongoing**

- Developed a transformer-based model to predict candidate future business visits for consumers based on visit history.
- Designed algorithms for fair sampling of new customer data for equitable prediction quality to minimize expected sampling regret.

Improving Fairness in Multi-Agent Resource Allocation **May 23 – Dec 24**

- Developed an algorithm for learning fair-efficient behavior in multi-agent Reinforcement Learning with resource constraints which allows flexible tradeoffs in utility and fairness.
- Designed environments and implemented custom RL algorithms based on popular paradigms like DDQN and MAPPO.

Passenger and Driver-side Fairness in Ridesharing **Jan 21 – Dec 22**

- Used properties of Mixed-Integer Programs to design an efficient and completely training-free method to improve fairness in ridesharing systems.
- Demonstrated the surprising ability of fairness incentives to improve utility in addition to improving fairness.

Using Machine Learning Models for Robot Control Prediction

Jan 22 – May 22

- Predicted control feature values using techniques like Gradient-Boosted Trees, Neural Network and Kernel Regression on a supervised learning problem.
- Performed feature engineering using correlation features, outlier removal, K-Means Clustering and Principle Component Analysis.

PUBLICATIONS

- **BIAS IN LLM-BASED REWARD MODELS.**
Ashwin Kumar, Yuzi He, Aram H. Markosyan, Bobbie Chern, and Imanol Arrieta-Ibarra. *Under review.*
- **DECAF: LEARNING TO BE FAIR IN MULTI AGENT RESOURCE ALLOCATION.**
Ashwin Kumar, and William Yeoh. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, (Extended abstract) 2025.
And In *Proceedings of the Reinforcement Learning Safety Workshop (RLSW)*, 2024.
- **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).
- **DIALECTICAL RECONCILIATION VIA STRUCTURED ARGUMENTATIVE DIALOGUES.**
Stylianios Loukas Vasileiou, Ashwin Kumar, William Yeoh, Tran Cao Son, and Francesca Toni. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, 2024.
- **USING SIMPLE INCENTIVES TO IMPROVE TWO-SIDED FAIRNESS IN RIDESHARING SYSTEMS.**
Ashwin Kumar, Yevgeniy Vorobeychik, and William Yeoh. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 2023.
- **FAIRNESS IN SCARCE SOCIETAL RESOURCE ALLOCATION: A CASE STUDY IN HOMELESSNESS APPLICATIONS.**
Ashwin Kumar, and William Yeoh. In *the 4th International Workshop on Autonomous Agents for Social Good (AASG)*, 2023.
- **IMPROVING ZONAL FAIRNESS WHILE MAINTAINING EFFICIENCY IN RIDESHARE MATCHING.**
Ashwin Kumar, Yevgeniy Vorobeychik, and William Yeoh. In *Proceedings of the Workshop on Agents in Traffic and Transportation (ATT-22)*, 2022.
- **VIZXP: A VISUALIZATION FRAMEWORK FOR CONVEYING EXPLANATIONS TO USERS IN MODEL RECONCILIATION PROBLEMS.**
Ashwin Kumar, Stylianios Loukas Vasileiou, Melanie Bancilhon, Alvitta Ottley, and William Yeoh. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 2022.
And In *Proceedings of the Workshop on Explainable Planning (XAIP)*, 2021.
- **A LOGIC-BASED EXPLANATION GENERATION FRAMEWORK FOR CLASSICAL AND HYBRID PLANNING PROBLEMS.**
Stylianios Loukas Vasileiou, William Yeoh, Tran Cao Son, Ashwin Kumar, Michael Cashmore, and Daniele Magazzeni. In *Journal of Artificial Intelligence Research (JAIR) Vol 73*, 2022.
- **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. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS) (System Demonstration)*, 2021