Vanshaj Khattar

☎ +1 540-514-3029 | ✉ vanshajk@vt.edu | ⌂ vanshajkhattar.github.io | *Google Scholar*

Summary

I am a Ph.D. candidate with research experience in reinforcement learning, optimization, and large language models. I am interested in how we can develop AI agents that are not only smart but also safe, interpretable, and can continually adapt to changing conditions.

Education

Virginia Polytechnic Institute and State University (Virginia Tech) *Blacksburg, VA*

Ph.D. Electrical Engineering *August 2021 - Present*

- Advisor: Dr. Ming Jin
- GPA: 3.64/4.0
- Thesis: Towards Trustworthy Reinforcement Learning Agents.

Virginia Polytechnic Institute and State University (Virginia Tech) *Blacksburg, VA*

MS Electrical Engineering *August 2019 - May 2021*

- Advisor: Dr. Azim Eskandarian
- Thesis: Controller design for crash avoidance in autonomous vehicles.

Delhi Technological University *New Delhi, India*

B.Tech Electrical and Electronics Engineering *August 2014 - May 2018*

- CGPA: 8.09/10.0

Industry Experience

Mitsubishi Electric Research Labs (MERL) *Cambridge, Boston*

Research Scientist Intern in Trustworthy and General AI *May 2025 - August 2025*

- **Mentors:** Ye Wang, Jing Liu, and Toshiaki Koike-Akino
- **Project:** Investigated the vulnerabilities of the current test-time training methods in large reasoning models (LRMs); developed novel jailbreaks exploiting vulnerabilities and proposed a safe-RL method to mitigate these vulnerabilities.
- **Achievements:** One workshop paper accepted at **AAAI 2026**, longer version currently under review.

National Renewable Energy Lab (NREL) *Golden, Colorado*

Graduate Summer Intern in ML for Power Systems *June 2024 - August 2024*

- **Mentors:** Yiyun Yao and Fei Ding
- **Project:** Developed a hierarchical graph-reinforcement learning-based solution for distribution grid critical load restoration under uncertain topology changes.
- **Achievements.** Conference paper accepted at **PES-GM 2025**, and one journal paper under preparation **(Preprint)**

Publications

Conference and Workshop Publications

Khattar, V., Choudhury, M., Rashid, M., Liu, J., Koike-AKino, T., Jin, M., Wang, Y., “Amplification Effects in Test-Time Reinforcement Learning: Safety and Reasoning Vulnerabilities”. **AAAI 2026-Trustworthy Agentic AI Workshop**

Khattar, V., Yao, Y., Ding, F., Jin, “Distribution Grid Critical Load Restoration under Uncertain Topology Changes via a Hierarchical Multi-Agent Reinforcement Learning Approach”. **IEEE PES-GM 2025**

Sel, B., Al-Tawaha, A., **Khattar, V.**, Jia, R. and Jin, M., “Algorithm of thoughts: Enhancing exploration of ideas in large language models”. **(ICML 2024)**

Khattar, V.*, Lin, T.*, Huang. Y*, Jia, R., Hong, J., Liu C, Vincentelli, A and Jin, M., “CausalPrompt: Enhancing LLMs with Weakly Supervised Causal Reasoning for Non-Language Applications”. **(ICLR 2024 Workshop Paper)**

Khattar, V. and Jin, M., “Optimization Solution Functions as Deterministic Policies for Offline Reinforcement Learning”. (American Control Conference) **(ACC 2024)**

Manzoor, F., **Khattar, V.**, Liu, C., and Jin, M., “Zero-day Attack Detection in Digital Substations using In-Context Learning”. **(SmartGridComm 2024)**

Khattar, V., Ding, Y., Sel, B., Lavaei, J. and Jin, M., “A CMDP-within-online framework for Meta-Safe Reinforcement Learning”. In The Eleventh International Conference on Learning Representations (**ICLR 2023 Spotlight**) .

Khattar, V. and Jin, M., “Winning the CityLearn challenge: adaptive optimization with evolutionary search under trajectory-based guidance”. In Proceedings of the (**AAAI 2023**).

Jin, M., **Khattar, V.**, Kaushik, H., Sel, B. and Jia, R., “On solution functions of optimization: universal approximation and covering number bounds”. In Proceedings of the (**AAAI 2023**).

Meimand, M., **Khattar, V.**, Yazdani, Z., Jazizadeh, F., Jin, M., “TUNEOPT: An Evolutionary Reinforcement Learning HVAC System Controller For Tuning Energy-Comfort Optimization Formulations”. (**BuildSys 2023**).

Khattar, V. and Eskandarian, A., “Stochastic predictive control for crash avoidance in autonomous vehicles based on stochastic reachable set threat assessment”. **(IMECE 2021)**.

Khattar, V. and Eskandarian, A., “Reactive online motion re-planning for crash mitigation in autonomous vehicles using bezier curve optimization”. ASME (**IMECE 2020**).

Valluru, S.K., Singh, M., Singh, M. and **Khattar, V.,**, “Experimental validation of PID and LQR control techniques for stabilization of cart inverted pendulum system”. In IEEE International Conference on (**RTEICT 2018**).

Journal Publications

Khattar, V. and Eskandarian, A., “Stochastic reachable set threat assessment for autonomous vehicles using trust-based driver behavior prediction”. SAE International Journal of Connected and Automated Vehicles. Paper link.

Technical Skills

Programming languages. Python, C, MATLAB, HTML

Frameworks. PyTorch, Tensorflow, cvxpy, NumPy, Pandas, Scikit-learn, Hugging Face, OpenAI Playground

Awards & Scholarships

2023	**AAAI 2023 travel scholarship.**, AAAI	*$ 750*
2022	**Member of the winning team ROLEVT at CityLearn challenge 2021.(ROLEVT team)**,	*$ 1500*
2021	**Second position in 2021 Torgersen Graduate Student Research Excellence Award for MS Oral presentation. (Link)**, Virginia Tech	*$ 500*

Outreach and Service

Conference reviewer: *1)* AISTATS 2022, 2023, 2024, 2025, 2026; *2)* ICLR 2025, 2026; *3)* ICML 2025, 2026; *4)* AAAI 2026

Workshops: Organized Trustworthy Interactive Decision-Making with Foundation Models workshop at IJCAI 2024 (Link)

Tutorials: Safe RL for Smart Grids tutorial at SmartGridComm 2024 conference. (Link)

Selected Talks and Presentations

Fall, 2024. *Tu*. PEC Conference at Virginia Tech. Spring, 2023. *Offline Actor-Critic with Optimization Policies for Demand Response and Urban Energy Management*. PEC Conference at Virginia Tech.

Fall, 2022. *Trustworthy Reinforcement Learning.* Presented to 150+ undergraduates in the undergraduate engineering research seminar, Fall 2022

Fall 2021. *Zeroth-Order Implicit Reinforcement Learning for Distributed Control Systems*. Southeast Control Conference 2021, Virginia Tech.