Rakshit S. Trivedi

(678)-488-7228 triver@mit.edu www.rtrivedi.me

Research Interests

Cooperative AI, Alignment, Multi-Agent Safety and Evaluations, Game theory, Reinforcement Learning, Generative Agents, Imitation Learning

Professional Positions

2023–present **Postdoctoral Associate**, *Massachusetts Institute of Technology*, Computer Science and Artificial Intelligence Laboratory, Host: Dylan Hadfield-Menell

2021–2023 **Postdoctoral Fellow**, *Harvard University*, Economics and Computer Science Research Group, Host: David Parkes

Summer 2019 **Research Intern**, *X and Google Brain*, Robotics Group, Host: Yunfei Bai

Summer 2018 **Research Scientist Intern**, *Amazon*, Machine Learning Group, Host: Xin Luna Dong

Summer 2017 **Applied Scientist Intern**, *Amazon*, Product Graph Group, Host: Xin Luna Dong

2013–2014 **Research Engineer**, *Teradata Corporation*, Teradata Labs

2010–2011 **Research Assistant**, *Indian Institute of Science*, DB Systems Lab, Host: Jayant Haritsa

Education

2015–2020 **Ph.D. in Computer Science (Focus Area: Machine Learning)**, *Georgia Institute of Technology*, Advised by Hongyuan Zha, Thesis: Learning Dynamic Processes over Graphs.
Minor area: Optimization and Decision Processes (School of Industrial & Systems Engineering).
Committee: Hongyuan Zha, Peter Battaglia, Xin Luna Dong, Duen Horng Chau, Umit Catalyurek.

2011–2012 **M.S. in Computer Science**, *Georgia Institute of Technology*
Specialization: Machine Learning

2005–2009 **B.Tech. in Computer Science**, *Nirma University*, Summa Cum Laude

Teaching

Fall 2018 **CSE 6740: Computational Data Analysis / Machine Learning**, *Georgia Tech*, Graduate Student Instructor and Head Teaching Assistant (Class of 200 students)
Designed course structure and materials and co-instructed the class.

Spring 2018 **CSE 6240: Topics in Temporal Point Process, Optimal Transport and GANs**, *Georgia Tech*, Teaching Assistant for class of 40 PhD students
Helped design assignments and directed class projects which were submitted to NeurIPS 2018.

Fall 2016 **CSE 6740: Computational Data Analysis / Machine Learning**, *Georgia Tech*, Graduate Student Instructor (Class of 150 students)
Conducted lectures on Sequential Models, Support Vector Machines, Dimensionality Reduction for Manifold data, Neural Networks and Deep learning. Helped develop and grade assignments/exams.

Advising and Mentoring

Massachusetts Institute of Technology

2023–2024 **Ensueo Choi**, *S.M. in Computer Science and Technology Policy*
Social Learning

2023 **Mehul Damani**, *Ph.D. candidate in Computer Science*
Multi-agent Reinforcement Learning

2024 **Timothy Quian**, *Undergraduate Research Assistant*
Multi-agent Learning and Generative Agents

Cooperative AI Foundation

2025 **Chandler Smith**, *Research Engineer*
Multi-agent Learning and Generative Agents

University of Toronto

2024 **Nikhil Chandak**, *Research Intern at Vector Institute*
Cooperation and AI Alignment

2024 **Andrei Muresanu**, *AI Research Scientist at Vector Institute*
Cooperation and AI Alignment

Harvard University

2021–2022 **Zhou Fan**, *Ph.D. candidate in Computer Science*
Imitation Learning

Georgia Institute of Technology

2022–present **Kartik Sharma**, *Ph.D. candidate in Machine Learning*
Graph Machine Learning and Generative Agents – PhD co-advisor

2020 **Rahul Duggal**, *Ph.D. candidate in Computational Science and Engineering*
Neural Architecture Search

2019–2020 **Jiachen Yang**, *Ph.D. candidate in Machine Learning*
Deep (multi-agent) Reinforcement Learning for graph-structured environments

2018 **Prasenjeet Biswal**, *MS Student in Computational Science and Engineering*
Temporal Graph Representation Learning

2016–2017 **Jenna Kwon**, *Undergraduate Researcher in Computer Science*
Independent Research on Multi-Armed Bandits for Recommendation Systems

Awards and Honors

2024 **Kavli Fellow**, US National Academy of Sciences

2017, 2020 Top Reviewer recognition for International Conference on Machine Learning

2018 Top Reviewer recognition for Neural Information Processing Systems conference

2016 **Best Paper Award**, Recsys Workshop on Deep Learning for Recommendation Systems

Professional Activities

Program Committee Member and External Reviewer

International Association for Safe and Ethical Artificial Intelligence (IASEI)

Neural Information Processing System (NeurIPS)

International Conference on Machine Learning (ICML)

International Conference on Learning Representations (ICLR)

Autonomous Agents and Multiagent Systems (AAMAS)

Association for the Advancement of Artificial Intelligence (AAAI)

The Web Conference (WWW)

Transactions on Machine Learning Research (TMLR)

Journal of Machine Learning Research (JMLR)

Journal of Artificial Intelligence Research (JAIR)

IEEE Transactions on Knowledge and Data Engineering (TKDE)

VLDB Journal

Workshops and Seminars Co-organized

2025 Workshop on Cooperation, Oversight and Trust in Multi-Agent LLMs, *Under Review*

2025 AI and Society Seminar Series, *MIT*

2024 Workshop on The Concordia Contest: Advancing the Cooperative Intelligence of Language Agents, *NeurIPS*

2023 Workshop on The Melting Pot Contest: Charting the Future of Generalized Cooperative Intelligence, *NeurIPS*

Grant Proposals

* All current contributions towards grant proposals are at non-PI level: includes design and development of problem formulation and solution approaches, writing relevant technical and budget sections, partial design of proposal structure and proofreading.

2021–2022 **DARPA**: Learning Compositional Policies for Economic Environments through Economics. Award Amount: $127,000. PI: David C. Parkes

2017-2021 **NSF Small**: Topics in Temporal Marked Point Processes: Granger Causality, Imperfect Observations and Intervention. Award Amount: $450,000.00. PI: Hongyuan Zha

2017-2020 **NSF EAGER:** SSDIM - Leveraging Point Processes and Mean Field Games Theory for Simulating Data on Interdependent Critical Infrastructures. Award Amount: $200,000. PI: Hiongyuan Zha

2018 unsuccessful **JP Morgan Research Grant Proposal:** Learning to explain and optimize risk in Financial Systems. Final two proposals. Proposed Amount: $150,000. PI: Hongyuan Zha

Selected Talks

2024 Tutorial on Cross-disciplinary insights into alignment in humans and machines, Neural Information Processing Systems (NeurIPS)

2023 Foundations for Learning in Multi-agent Ecosystems: Modeling, Imitation and Equilibria, CS Colloquium, University of Southern California

2022 Learning from Interactions in Networked Systems. Beneficial AI Seminar at Center for Human Compatible AI, UC Berkeley

2022 Learning from Interactions as an Inverse Problem. Invited talk at University of Illinois, Chicago

2022 Learning from Interactions as an Inverse Problem. Invited talk at University of Rochester, New York

2020 Temporal Graph Representation Learning. Guest Lectures for Web Search course taught by Prof. Srijan Kumar at Georgia Tech

2019 Learning Dynamic Processes over Graphs. Invited talk at Translational Data Analytics Lab, Georgia Tech

2018 Learning Dynamic Graph Representations. Workshop Spotlight at NeurIPS

2017 Fake News Mitigation via Point Process Based Intervention (oral). ICML

2017 Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs (oral) ICML

2013 Discourse Connectors for Latent Subjectivity in Sentiment Analysis (oral). NAACL-HLT

2012 CODD: COnstructing Dataless Databases (spotlight). Workshop at ACM SIGMOD/PODS

Pre-prints and Working Papers

[1] Gillian Hadfield, **Rakshit Trivedi**, Dylan Hadfield-Menell. Building AI for the Democratic Matrix: A Technical Research Agenda for Normative Competence and Normative Institutions. *To Appear, 2025 (Preliminary version disseminated at Knight Symposium on AI and Democratic Freedom).*

[2] Kartik Sharma, **Rakshit Trivedi**. COLD-Steer: Steering Large Language Models via In-Context One-step Learning Dynamics. *Under Review.*

[3] **Rakshit Trivedi**, Gillian Hadfield, Dylan Hadfield-Menell. Altared Environments: Institutions Enable normative competence in AI for Cooperation in Multi-Agent Systems. *Working Paper, 2025.*

[4] **Rakshit Trivedi**. Steerable Alignment Objective. *In preparation, 2025.*

[5] Stephen Casper, Luke Bailey, Rosco Hunter, Carson Ezell, Emma Cabalé, Michael Gerovitch, Stewart Slocum, Kevin Wei, Nikola Jurkovic, Ariba Khan,Phillip Christoffersen, A. Pinar Ozisik, **Rakshit Trivedi**, Dylan Hadfield-Menell, Noam Kolt. The AI Agent Index. *Pre-print, 2025.*

[6] Marko Tesic, Yue Zhao, Joel Z Leibo, **Rakshit Trivedi**, Jose Hernandez-Orallo. Beyond the High Score: Prosocial Ability Profiles of Multi-Agent Populations. *Pre-print, 2025.*

[7] Kartik Sharma, Yiqiao Jin, **Rakshit Trivedi**, and Srijan Kumar. Efficient Knowledge Probing of Large Language Models by Adapting Pre-trained Embeddings. *Pre-print, 2025.*

[8] Atrisha Sarkar, Andrei Ioan Muresanu, Carter Blair, Aaryam Sharma, **Rakshit Trivedi**, Gillian K Hadfield. Normative Modules: A Generative Agent Architecture for Learning Norms that Supports Multi-Agent Cooperation. *Pre-print, 2024.*

Refereed Conference Publications

[17] **Rakshit Trivedi**, Kartik Sharma, David Parkes. Inner Speech as Behavior Guides: Steerable Imitation of Diverse Behaviors for Human-AI coordination. *Neural Information Processing System (NeurIPS), 2025*, **Spotlight paper, among top 3.2% of total submissions.**

[16] Chandler Smith, Marwa Abdulhai, Manfred Diaz, Marko Tesic, **Rakshit Trivedi**, Alexander Sasha Vezhnevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar A. Duéñez-Guzmán, John P. Agapiou, Jayd Matyas, Danny Karmon, Dylan Hadfield-Menell, Natasha Jaques, Tim Baarslag, Jose Hernandez Orallo, Joel Leibo. Evaluating Generalization Capabilities of LLM-Based Agents in Mixed-Motive Scenarios Using Concordia. *Dataset and Benchmark Track at Neural Information Processing System (NeurIPS), 2025.*

[15] **Rakshit Trivedi**, Akbir Khan, Jesse Clifton, Lewis Hammond, Edgar Duéñez-Guzmán, John Agapiou, Jayd Matyas, Sasha Vezhnevets, Natasha Jaques, Jakob Foerster, Vincent Conitzer, José Hernández-Orallo, Dylan Hadfield-Menell, Joel Leibo. Melting Pot Contest: Charting the Future of Generalized Cooperative Intelligence. *Dataset and Benchmark Track at Neural Information Processing System (NeurIPS), 2024.*

[14] Kartik Sharma, Srijan Kumar, **Rakshit Trivedi**. Diffuse, Sample, Project: Plug-And-Play Controllable Graph Generation. *International Conference on Machine Learning (ICML), 2024.*

[13] Kartik Sharma, **Rakshit Trivedi**, Rohit Sridhar, Srijan Kumar. Temporal Dynamics Aware Adversarial Attacks On Discrete-Time Graph Models. *Knowledge Discovry and Data Mining (KDD), 2023.*

[12] Matthias Gertsgrasser, **Rakshit Trivedi**, David C. Parkes. CrowdPlay: Crowdsourcing human demonstrations for offline learning. *International conference on Learning Representations (ICLR), 2023.*

[11] Jiachen Yang, Ethan Wang, **Rakshit Trivedi**, Tuo Zhao, Hongyuan Zha. Adaptive Incentive Design with Multi-Agent Meta-Gradient Reinforcement Learning. *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2022.*

[10] **Rakshit Trivedi** and Hongyuan Zha. Learning Strategic Network Emergence Games. *Neural Information Processing Systems (NeurIPS), 2020.*

[9] **Rakshit Trivedi**, Jiachen Yang and Hongyuan Zha. GraphOpt: Learning Optimization Models of Graph Formation. *International Conference on Machine Learning (ICML), 2020.*

[8] **Rakshit Trivedi**, Mehrdad Farajtabar, Prasenjeet Biswal and Hongyuan Zha. DyRep: Representation Learning over Dynamic Graphs. *International Conference on Learning Representations (ICLR), 2019.*

[7] **Rakshit Trivedi**, Bunyamin Sisman, Jun Ma, Christos Faloutsos, Hongyuan Zha and Xin Luna Dong. LinkNBed: Multi-Graph Representation Learning with Entity Linkage. *Annual Meeting of the Association for Computational Linguistics (ACL), 2018.*

[6] Jiachen Yang, Xiaojing Ye, **Rakshit Trivedi**, Huan Xu and Hongyuan Zha. Deep Mean Field Games for Learning Optimal Behavior Policy of Large Populations. *International Conference on Learning Representations (ICLR), 2018*, **Oral presentation, top 2% of submissions**.

[5] **Rakshit Trivedi**, Hanjun Dai, Yichen Wang and Le Song. Know-Evolve: Deep Temporal Reasoning for Dynamic Knowledge Graphs. *International Conference on Machine Learning (ICML), 2017.*

[4] Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, **Rakshit Trivedi**, Elias Khalil, Shuang Li, Huan Xu, Le Song and Hongyuan Zha. Fake News Mitigation via Point Process Based Intervention. *International Conference on Machine Learning (ICML), 2017.*

[3] Yichen Wang, Nan Du, **Rakshit Trivedi** and Le Song. Coevolutionary Latent Feature Processes for Continuous-Time User-Item Interactions. *Neural Information Processing Systems (NeurIPS), 2016.*

[2] Nan Du, Hanjun Dai, **Rakshit Trivedi**, Utkarsh Upadhyay, Manuel Gomez-Rodriguez and Le Song. Recurrent Marked Temporal Point Process. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2016.*

[1] **Rakshit Trivedi** and Jacob Eisenstein. Discourse Connectors for Latent Subjectivity in Sentiment Analysis. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2013.*

Workshop Publications

[10] Chandler Smith, **Rakshit Trivedi**, Jesse Clifton, Lewis Hammond, Akbir Khan, Marwa Abdulhai, Alexander Sasha Vezhnevets, John P. Agapiou, Edgar A. Duéñez-Guzmán, Jayd Matyas, Danny Karmon, Dylan Hadfield-Menell, Natasha Jaques, Tim Baarslag, Joel Z. Leibo. The Concordia Contest: Advancing the Cooperative Intelligence of Language Model Agents. *Workshop at NeurIPS, 2024.*

[9] **Rakshit Trivedi**, Nikhil Chandak, Carter Blair, Atrisha Sarkar, Tehilla Weltman, Dylan Hadfield-Menell, Gillian K Hadfield. Altared Environments: The Role of Normative Infrastructure in AI Alignment. *Agentic Markets Workshop at ICML, 2024.*

[8] Atrisha Sarkar, Andrei Ioan Muresanu, Carter Blair, **Rakshit Trivedi**, Gillian K Hadfield. Normative Modules: A Generative Agent Architecture for Learning Norms that Supports Multi-Agent Cooperation. *EC workshop on Foundation Models and Game Theory, 2024.*

[7] **Rakshit Trivedi**, Akbir Khan, Jesse Clifton, Lewis Hammond, John Agapiou, Edgar Dueñez-Guzman, Jayd Matyas, Dylan Hadfield-Menell, Joel Z Leibo. Melting Pot Contest. *Workshop at NeurIPS, 2023.*

[6] Kartik Sharma, Srijan Kumar, **Rakshit Trivedi**. Plug-and-Play Controllable Graph Generation with Diffusion Models. *ICML Workshop on Structured Probabilistic Inference and Generative Modeling, 2023.*

[5] Kartik Sharma, **Rakshit Trivedi**, Rohit Sridhar, Srijan Kumar. Imperceptible Adversarial Attacks on Discrete-Time Dynamic Graph Models. *Neurips Workshop on Temporal Graph Learning, 2022.*

[4] **Rakshit Trivedi**, Jiachen Yang and Hongyuan Zha. Learning Optimization Models of Graphs. *NeurIPS Workshop on Perception as Generative Reasoning (PGRSCP), 2019.*

[3] **Rakshit Trivedi**, Mehrdad Farajtabar, Prasenjeet Biswal and Hongyuan Zha. Learning Dynamic Graph Representations. *NeurIPS Workshop on Modeling and Decision-making in SpatioTemporal Domain (SpatioTemporal), 2018.*

[2] Hanjun Dai, Yichen Wang, **Rakshit Trivedi** and Le Song. Recurrent Coevolutionary Feature Embedding Processes for Recommendation. *Recsys Workshop on Deep Learning for Recommendation Systems (DLRS), 2016.*

[1] **Rakshit Trivedi**, I. Nilavalagan and Jayant Haritsa. CODD: COnstructing Dataless Databases. *International Workshop on Testing Database Systems DBTest, 2012.* (Accepted for presentation in **2012 ACM SIGMOD conference**).