

# UDARI MADHUSHANI SEHWAG

Visiting Postdoc, Stanford University, California, CA 94305  
AI Research Scientist, JPMorgan AI Research, Palo Alto, CA 94304

☎ (609)-933-7329 ✉ [udarim@stanford.edu](mailto:udarim@stanford.edu)  
🌐 [udarimadhu.github.io](https://github.com/udarimadhu)

## RESEARCH INTEREST

My research vision is to embed AI agents with the ability to enhance their capabilities through collective intelligence, ultimately enabling them to seamlessly coexist with humans, augmenting their cognitive and physical capabilities. In my research work I primarily focus on developing socially intelligent and aligned AI agents.

**Key Topics:** Generative AI (LLMs - collective alignment, social intelligence, human behaviour proxy, policy and governance, responsible AI), Multi-agent reinforcement learning (generalization, evolution of cooperation, communication), Game theory, Robotics, Bandit theory.

**PhD thesis:** Learning through social interactions and learning to socially interact in multi-agent learning

## EDUCATION

Degree	Institution	Years
Ph.D. (Artificial Intelligence)	Princeton University, USA	2017 - 2023
M.A. (Artificial Intelligence)	Princeton University, USA	2017 - 2019
B.Sc. (Robotics & Control Theory)	University of Peradeniya, Sri Lanka	2011 - 2015

## HONORS

### Graduate Honors (Princeton University)

- Harold W. Dodds Fellowship (prestigious honorific fellowship from Graduate School) 2021-22
- Britt and Eli Harari Fellowship 2020-21
- Larisse Rosentweig Klein Memorial Award Fall 2019
- Martin Summerfield Graduate Fellowship 2018-19
- Athena-Feron Prize for Mathematical Excellence Fall 2018
- Elliotte Robinson Little '25 Student Aid Fund Fellowship 2017-18

### Other Awards

- Presidential Award for Scientific Research (Sri Lanka) 2019
- Best Undergraduate Project Award (IEEE Sri Lanka Section) December 2015

## EXPERIENCE

### Visiting Postdoc

- Stanford University, California, USA August 2023 - Present  
*Focus:* Collective alignment, Socially intelligent generative agents, Policy and governance, Safety, Robotics

### Research Scientist

- JPMorgan AI Research, New York, USA July 2023 - Present  
*Focus:* Collective alignment, Socially intelligent generative agents, Policy and governance, LLMs as human proxy, Safety

### Research Internships

- Google Deepmind, London, UK May 2022 - Sep 2022  
*Mentors:* Daniel Hennes and Edgar Duenez-Guzman  
*Project:* Zero-shot Generalization in Mixed Motive Games

- Facebook AI Research, Menlo Park, California, USA May 2021 - August 2021  
*Mentors:* Kalesha Bullard and Roberto Calandra  
*Project:* Incentivizing Coordination in Multi-Agent Reinforcement Learning
- Siemens, Princeton, New Jersey, USA May 2020 - August 2020  
*Mentors:* Biswadip Dey and Amit Chakraborty  
*Project:* Using Hamiltonian Monte Carlo Sampling for Reinforcement Learning Problems in High-dimension

#### Assistant in Research

- Princeton University September 2017 - May 2023
- Sri Lanka Technological Campus (SLTC), Sri Lanka January 2016 - August 2017

#### Guest Lectures

- MAE 545 Collective Intelligence: Dynamics and Control of Multi-Agent Systems Fall 2022  
(Princeton University)

#### Assistant in Teaching

- MAE 542 Advanced Dynamics (Princeton University) Fall 2020
- MAE 502/APC 506 Mathematical Methods of Engineering Analysis (Princeton University) Spring 2020
- MAE 345/MAE 549 Introduction to Robotics (Princeton University) Fall 2019
- EE 554 Microwave Techniques (University of Peradeniya), Sri Lanka October 2015 - December 2015

## OUTREACH

---

#### Mentoring

Stanford University

- Arjun Karanam (M.Sc. Computer Science) 2024

University of California, Berkeley

- Marwa Abdulhai (Ph.D. Electrical Engineering and Computer Sciences) 2024
- Nivasini Ananthakrishnan (Ph.D. Electrical Engineering and Computer Sciences) 2024

University of Maryland, College Park

- Yuan Cheng Xu (Ph.D. Applied Mathematics) 2024
- Souradip Chakraborty (Ph.D. Computer Science) 2024

University College London, UK

- Oliver Shumbers (Ph.D. Computer Science) 2024

Princeton University

- Hui Yuan (Ph.D. Electrical and Computer Engineering) 2024
- Justin Lidard (Ph.D. Mechanical Engineering) 2021-23
- Kathryn Wantlin (M.Sc. Computer Science) 2021-22
- Sarah Dillender (B.S.E. Mechanical Engineering) 2020-21  
*Thesis:* Message passing structures for improved policy finding in decentralized multi-agent Q-learning.
- Gargi Sadalgekar, Samarie Wilson and Jacob Walrath (B.S.E. Mechanical Engineering) 2020-21  
*Thesis:* Decision making and task allocation in a multi-robot system.

Peradeniya University, Sri Lanka

- Lasitha Weerakon (Post graduate, Mechanical Engineering) 2016-17
- Isuru Basnayake (Post graduate, Mechanical Engineering) 2016-17
- Kusal Tennakoon (Post graduate, Mechanical Engineering) 2016-17

#### Commitment to Diversity and Inclusion

- Co-organized "Re-education Monthly Book Club" aimed towards improving awareness of implicit biases (Princeton University) 2020-23
- Co-organized the workshop "Diversity and Mentorship" at American Control Conference May 2021
- Co-organized the workshop "Inclusive Teaching" (Princeton University) Jan 2021

#### Leadership Roles

- Committee member of Women in STEM Leadership Council (Princeton University) 2020-23
- Committee member of Graduate Women in Science & Engineering (Princeton University) 2020-23

- Future Digileaders Training (Stockholm, Sweden) November 2019
- Founder of EE Instructors Scholarship (Peradeniya University, Sri Lanka) 2016

#### Volunteer for

- World Maker Faire September 2018
- American Control Conference June 2018

## ACADEMIC SERVICES

---

#### Program committee Member of

- Games Agents and Incentives Workshop (**GAIW**) at International Conference on Autonomous Agents and Multi Agent Systems (**AAMAS**) 2023, 2024
- Adaptive and Learning Agents Workshop (**ALA**) at International Conference on Autonomous Agents and Multi Agent Systems (**AAMAS**) 2022, 2023
- Cooperative AI Workshop at Conference on Neural Information Processing Systems (**NeurIPS**) 2021

#### Co-organizer of

- Special issue on "Collective Artificial Intelligence" at Proceedings of National Academy of Science (**PNAS**) 2023
- Proceedings of National Academy of Science (**PNAS**) special issue symposium on "Collective Artificial Intelligence" 2022

#### Editor for

**Journal:** Proceedings of National Academy of Science (**PNAS**) 2023, 2024

#### Reviewer for

**Journal:** Automatica, IEEE Transactions on Automatic Control, IEEE Control Systems Letters, Journal of Field Robotics

**Conference:** Conference on Neural Information Processing Systems, International Conference on Machine Learning, International Conference on Learning Representations, IEEE Conference on Decision and Control, American Control Conference, International Conference on Intelligent Robots and Systems, European Control Conference, International Conference of Industrial and Information Systems

#### Member of

IEEE (Institute of Electrical and Electronics Engineers), IEEE Young Professionals, IEEE Women in Engineering, APS (American Physical Society)

## PUBLICATIONS

---

#### Technical Reports

[T1] **Melting Pot 2.0**

John Agapiou, Alexander Vezhnevets, Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, **Udari Madhushani**, Kavya Kopparapu, Ramona Comanescu, DJ Strouse, Michael B. Johanson, Sukhdeep Singh, Julia Haas, Igor Mordatch, Dean Mobbs, Joel Z Leibo  
ArXiv 2023

#### Peer-reviewed Journal Articles

[J4] **Autocratic Learning and Unilateral Incentive Alignment in Two-player Stochastic Games**  
Alex McAvoy, **Udari Madhushani Sehwal**, Christian Hilbe, Wolfram Barfuss, Krishnendu Chatterjee, Qi Su, Naomi Ehrich Leonard, Joshua B. Plotkin  
Proceedings of National Academy of Science - Special issue on Collective Artificial Intelligence (**PNAS**) 2024 (accepted for publication)

[J3] **Collective Cooperative Intelligence**  
Wolfram Barfuss, Jessica Flack, Chaitanya S. Gokhale, Lewis Hammond, Christian Hilbe, Joel Leibo, Tom Lenaerts, Naomi Leonard, Simon Levin, **Udari Madhushani Sehwal**, Alex McAvoy, Janusz M. Meylahn, Fernando P. Santos  
Proceedings of National Academy of Science - Special issue on Collective Artificial Intelligence (**PNAS**) 2024 (accepted for publication)

- [J2] **Heterogeneous Explore-Exploit Strategies on Multi-Star Networks**  
**Udari Madhushani**, Naomi Leonard  
IEEE Control Systems Letters, 2021.
- [J1] **Semi-globally Exponential Trajectory Tracking for a Class of Spherical Robots**  
**Udari Madhushani**, Sanjeeva Maithripala, Janaka Wijayakulasooriya, Jordan Berg  
Automatica, 2017.

#### Peer-reviewed Conference Articles

- [C15] **SORRY-Bench: A Systematic Evaluation on Large Language Model Safety Refusal Behaviors**  
Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, **Udari Madhushani Sehwag**, Boyi Wei, Luxi He, Kaixuan Huang, Dacheng Li, Ying Sheng, Bo Li, Danqi Chen, Kai Li, Peter Henderson, Prateek Mittal  
(Under review at NeurIPS 2024)
- [C14] **O3D: Offline Data-Driven Discovery and Distillation for Sequential Decision Making with Large Language Models**  
Yuchen Xiao, Yanchao Sun, Mengda Xu, **Udari Madhushani Sehwag**, Jared Vann, Deepeka Garg, Sumitra Ganesh  
Conference on Language Modeling **COLM** 2024.
- [C13] **On Using Hamiltonian Monte Carlo Sampling for Reinforcement Learning**  
**Udari Madhushani**, Biswadip Dey, Naomi Leonard, Amit Chakraborty  
Conference on Decision and Control, (**CDC**) 2022. (**Invited Paper**)
- [C12] **A Regret Minimization Approach to Multi-Agent Control**  
Udaya Ghai, **Udari Madhushani**, Naomi Leonard, Elad Hazan  
International Conference on Machine Learning, (**ICML**) 2022.
- [C11] **Provably Efficient Multi-Agent Reinforcement Learning with Fully Decentralized Communication**  
Justin Lidard, **Udari Madhushani**, Naomi Leonard  
American Control Conference, (**ACC**) 2022.
- [C10] **One More Step Towards Reality: Cooperative Bandits with Imperfect Communication**  
**Udari Madhushani**, Abhimanyu Dubey, Naomi Leonard, Alex Pentland  
Conference on Neural Information Processing Systems, (**NeurIPS**) 2021.
- [C9] **Multi-robot Learning and Coverage of Unknown Fields**  
Maria Santos, **Udari Madhushani**, Alessia Benevento, Naomi Leonard  
IEEE International Symposium on Multi-Robot and Multi-Agent Systems, (**MRS**) 2021.
- [C8] **Distributed Bandits: Probabilistic Communication on  $d$ -regular Graphs**  
**Udari Madhushani**, Naomi Leonard  
European Control Conference, (**ECC**) 2021.
- [C7] **A Dynamic Observation Strategy for Multi-agent Multi-armed Bandit Problem**  
**Udari Madhushani**, Naomi Leonard  
European Control Conference, (**ECC**) 2020.
- [C6] **Heterogeneous Stochastic Interactions for Multiple Agents in a Multi-armed Bandit Problem**  
**Udari Madhushani**, Naomi Leonard  
European Control Conference, (**ECC**) 2019. (**Invited Paper**)
- [C5] **Feedback Regularization and Geometric PID Control for Robust Stabilization of a Planar Three-link Hybrid Bipedal Walking Model**  
Lasitha Weerakoon, **Udari Madhushani**, Sanjeeva Mathripala, Jordan Berg  
American Control Conference, (**ACC**) 2018.
- [C4] **Multi-armed Bandit Based Approach for Performance Enhancement of Window Intensity Test (WIT) Detector**  
Kusal Tennakoon, **Udari Madhushani**, Sanjeeva Mathripala  
International Conference of Industrial and Information Systems, (**ICIIS**) 2017.

- [C3] **Intrinsic PID Controller for a Segway Type Mobile Robot**  
Isuru Basnayake, **Udari Madhushani**, Sanjeeva Mathripala  
International Conference of Industrial and Information Systems, (**ICIIS**) 2017.
- [C2] **Feedback Regularization and Geometric PID Control for Trajectory Tracking of Mechanical Systems: Hoop Robots on an Inclined Plane**  
**Udari Madhushani**, Sanjeeva Mathripala, Jordan Berg  
American Control Conference, (**ACC**) 2017.
- [C1] **WIT: Window Intensity Test Detector and Descriptor**  
**Udari Madhushani**, Sanjeeva Mathripala, Janaka Wijayakulasooriya  
International Conference of Industrial and Information Systems, (**ICIIS**) 2016.

#### Peer-reviewed Workshop Articles

- [W10] **A Heterogeneous Agent Model of Mortgage Servicing: An Income-based Relief Analysis**  
Deepeka Garg, Benjamin Patrick Evans, Leo Ardon, Annapoorani Lakshmi Narayanan, Jared Vann, **Udari Madhushani**, Makada Henry-Nickie, Sumitra Ganesh  
AI in Finance for Social Impact (**AIFinSI**), AAAI Conference on Artificial Intelligence, (**AAAI**) 2024.
- [W9] **O3D: Offline Data-Driven Discovery and Distillation for Sequential Decision Making with Large Language Models**  
Yuchen Xiao, Yanchao Sun, Mengda Xu, **Udari Madhushani**, Jared Vann, Deepeka Garg, Sumitra Ganesh  
Foundation Models for Decision Making Workshop (**FMDM**), Conference on Neural Information Processing Systems, (**NeurIPS**) 2023.
- [W8] **Heterogeneous Social Value Orientation Leads to Meaningful Diversity in Sequential Social Dilemmas**  
**Udari Madhushani**, Kevin McKee, John Agapiou, Joel Z Leibo, Thomas Anthony, Richard Everett, Edward Hughes, Karl Tuyls, Edgar Duéñez-Guzmán  
Adaptive and Learning Agents Workshop (**ALA**), International Conference on Autonomous Agents and Multi Agent Systems (**AAMAS**) 2023)
- [W7] **Multi-robot Learning and Coverage of Unkown Fields**  
Maria Santos, **Udari Madhushani**, Alessia Benevento, Naomi Leonard  
Autonomous Robots and Multirobot Systems (**ARMS**), International Conference on Autonomous Agents and Multiagent Systems (**AAMAS**) 2022.
- [W6] **A Regret Minimization Approach to Multi-Agent Control**  
Udaya Ghai, **Udari Madhushani**, Naomi Leonard, Elad Hazan  
Workshop on Gamification and Multiagent Solutions (**GMS**), International Conference on Learning Representations (**ICLR**) 2022. (**Oral - Won the best poster award**)
- [W5] **Provably Efficient Decentralized Communication for Multi-Agent RL**  
Justin Lidard, **Udari Madhushani**, Naomi Leonard  
Reinforcement Learning in Games (**RLG**), AAAI Conference on Artificial Intelligence (**AAAI**) 2022. (**Oral**)
- [W4] **On Using Hamiltonian Monte Carlo Sampling for Reinforcement Learning Problems in High-dimension**  
**Udari Madhushani**, Biswadip Dey, Naomi Leonard, Amit Chakraborty  
Deep Reinforcement Learning Workshop (**Deep RL**), Conference on Neural Information Processing Systems (**NeurIPS**) 2021.
- [W3] **When to Call Your Neighbor? Strategic Communication in Cooperative Stochastic Bandits**  
**Udari Madhushani**, Naomi Leonard  
Learning and Decision-Making with Strategic Feedback (**StratML**), Conference on Neural Information Processing Systems (**NeurIPS**) 2021.

- [W2] **It Doesn't Get Better and Here's Why: A Fundamental Drawback in Natural Extensions of UCB to Multi-agent Bandits**  
**Udari Madhushani**, Naomi Leonard  
I can't Believe It's Not Better! Workshop (**ICBINB**), Conference on Neural Information Processing Systems (**NeurIPS**) 2020. (**Spotlight**)
- [W1] **Distributed Learning: Sequential Decision Making in Resource-Constrained Environments**  
**Udari Madhushani**, Naomi Leonard  
Practical ML for Developing Countries Workshop (**PML4DC**), International Conference on Learning Representations, (**ICLR**) 2020. (**Oral**)

## SKILLS

---

**Computer Skills:** Python, PyTorch, TensorFlow, JAX, Matlab, L<sup>A</sup>T<sub>E</sub>X

**Language Skills:** Sinhala (Native), English (Fluent)