☎ +447824520507
✉ haripriyaaharikumar@gmail.com
LinkedIn Profile
Google Scholar
GitHub Profile

Haripriya Harikumar

About

My research centers on developing mechanisms for AI Safety and Security, spanning both fundamental theory and practical deployment. I focus on ensuring that AI systems remain secure, trustworthy, and robust, particularly in adversarial or malicious settings. This includes designing methods that preserve the privacy of sensitive data and enable reliable predictions and safe decision-making under adversarial threats and distribution shift. I am also deeply interested in the underlying principles of out-of-distribution detection, data manifold learning, machine unlearning and the robustness of Language Vision Models (LVMs) and Large Language Models (LLMs). At the applied level, I have explored human behavioral modeling using social-media data, with the goal of identifying intervention strategies to help prevent addictive behaviors. You can find more details about me on my Homepage.

Education

Ph.D. in Machine Learning *Nov 2015 - Feb 2019*
Deakin University, Australia
Thesis Title: Machine Learning to Fight Addiction Using Social Media

M.Tech in Computer Science and Engineering *2011 - 2013*
Amrita University, India

B.Tech in Information Technology *2007 - 2011*
Calicut University, India

Work Experience

Research Fellow in Machine Learning, University of Manchester *Nov 2024 - Present*

- Under Prof. Samuel Kaski's UKRI Turing AI World-Leading Researcher Fellowship.
- Conduct fundamental research in Machine learning in adversarial robustness in Bayesian, user modeling, and differential privacy (AI Security and Safety).
- Mentor undergraduate, graduate students, PhD and postdocs for their research projects.
- Collaborate with professionals from the University of Manchester and Aalto University, Deakin University Australia, and VinAI Vietnam.
- Contributes to the research laboratory to recruit PhD, Postdoctoral and Research fellowship candidates.

Postdoctoral Research Fellow, Applied Artificial Intelligence Institute (A2I2), Australia *Feb 2019 - Nov 2024*

- **8** plus years of experience in cutting-edge Machine Learning/Deep Learning/Data mining research. Research areas include Backdoor Attacks and Defenses, Out-of-Distribution Detection, Privacy, and Knowledge distillation.
- Led a Safe and Responsible AI research team and contributed to numerous grants.
- Supervise Higher Degree Research students (Finished 1, ongoing - 2), provide mentorship (graduate and research assistants), and deliver guest lectures.
- Successful collaborations with health institutes within (Institute for Health Transformation, Deakin University) and outside (Department of Clinical Data Analytics, Max Super Specialty Hospital, India) Australia with the outcomes are published in top-tier health journals (BioData Mining, and Health Care Management Science).
- Collaborate with Professors and students from top-tier universities (IIT India, VinAI Vietnam, University of Maryland, USA, University of Wollongong, Australia) and industries (Max Super speciality hospital, TrojAI, Simular) globally.

Research Associate, Amrita CREATE, India *June 2013 - May 2016*

- Conducted research using statistical modelling techniques such as classification (linear regression, logistic regression), Clustering (K-Means), dimensionality reduction (PCA, SVM), and Multi-label classification with research outcomes published in machine learning and data mining conference venues.

- Mentored and tutored both undergraduate and graduate students on their research projects and machine learning skill development.

Teaching/Mentorship

- **Lecture**: In person Guest Lecture to BSc Computer Science and Mathematics (Semester 2, Course: Machine Learning (COMP24112), topic - Security Issues in Deep Learning) students at the University of Manchester, UK.

- **PhD research supervision**: Supervise an HDR Candidate (graduated) at Applied Artificial Intelligence Institute, Deakin University, and 2 under enrollment process.

- **Postdoc research supervision**: Mentor Postdocs in Centre for AI Fundamentals in University of Manchester and in University of Sheffield, UK.

- **UG Project Supervision**: Supervising two final year Bachelor students Department of Computer Science (2025) in the area Machine Unlearning.

- **PG Project Supervision**: Successfully supervised the research project of a Master's student (SIT723 and SIT724 Research Project A and B), at the School of Information Technology, Deakin University.

- **Mentor**: An invited mentor at the Women in Machine Learning (WiML) workshop, co-located with NeurIPS 2023 Conference.

- **Panellist**: An invited panellist at Black in AI Emerging Leaders Grad prep program entitled, "MSc/Ph.D: What to expect (pre-admission)", January 2024.

- **Guest lecture**: Invited guest lecture at College of Engineering, Kerala, India. (Topic: Introduction to Deep learning and research scopes.), 2024.

- **Workshop**: Organised a workshop **Backdoors in Deep Learning: The Good, the Bad, and the Ugly** at the **Rank A*** Machine learning conference **Neural Information Processing Systems** (NeurIPS, 2023). Collaborated with academic and industry fellows from VinUniversity - Vietnam, VinAI Research - Vietnam, Cornell University - USA, University of Maryland - USA, Center for Data Science at NYU - USA, Clemson University - USA, **Simular** - USA (Information Technology Company).

Invited Talks/Presentations

- **Turing AI Fellowship Annual Event, November 2025 at The Royal Society, London** - Engage with potenatial collaborators from Alan Turing Institute to Government officials about numerous grants and potential opportunities (presented a poster).

- **Seminar Series 2024-till date**: Active member of four seminar series - Joint Aalto University and University of Manchester weekly seminar series, Bayesian Design reading Group, a fortnightly reading group, and Department of Computer Science, University of Manchester weekly Seminar and AI-Fun & ELLIS Invited Speaker Series.

- **Poster Presentation**: Trustworthy AI, Secure and Safe Foundation Models, at CISPA Helmholtz Center for Information Security, 2025.

- **Invited Keynote Speaker**: An Invited keynote speaker at the International Conference on FOSS Approaches towards Computational Intelligence and Language Technology, 2024 (FOSS-CIL T24), India.

- **Invited Speaker**: An invited speaker at **IEEE Student Branch** at **Deakin University**, (Topic: AI Security: Exploring Backdoor Attacks, Defense Strategies, and a Path to Good), 2024.

- **Invited Speaker**: An invited speaker at MAIL Research Lab at **VinUni**, (Topic: Exploring Backdoor Intrusions, Defense Strategies, and the Path to Social Good.), 2023.

- **Seminar Series**: Active member of a seminar series named "Sample Efficient AI" conducted by A2I2 from 2017 to 2024.
- **Research paper at conferences**: Presented posters and research papers at top conferences (UAI, ECML, ICDM, PAKDD, ADMA etc).

Refereed Conference papers

- **Haripriya Harikumar** and Santu Rana. TRUST: Test-time resource utilization for superior trustworthiness. **arXiv preprint**, 2025
- Alex Hämäläinen Lukas Prediger Amir Sonee*, **Haripriya Harikumar*** and Samuel Kaski. Privacy-preserving neural processes for probabilistic user modeling. **UAI**, 2025, ***equal contribution**
- Banibrata Ghosh, **Haripriya Harikumar**, Santu Rana, and Svetha Venkatesh. Robust nearest neighbour retrieval using targeted manifold manipulation. arXiv preprint (under review), 2025
- **Haripriya Harikumar**, Santu Rana, Kien Do, Sunil Gupta, Wei Zong, Willy Susilo, and Svetha Venkatesh. Defense against multi-target multi-trigger backdoor attacks. **Data Science Foundations and Applications, Special edition, PAKDD**, 2025
- Banibrata Ghosh, **Haripriya Harikumar**, Santu Rana, and Svetha Venkatesh. Targeted manifold manipulation against adversarial attack. **SaTML**, 2025
- Banibrata Ghosh*, **Haripriya Harikumar***, Khoa Doan, Svetha Venkatesh, and Santu Rana. Composite concept extraction through backdooring. In International Conference on Pattern Recognition **ICPR 2024**, ***equal contribution**
- Banibrata Ghosh*, **Haripriya Harikumar***, Khoa Doan, Svetha Venkatesh, and Santu Rana. Composite concept extraction through backdooring. In *In FGVC, **CVPR Workshop, non-archival, *equal contribution***, 2024
- Kien Do, Dung Nguyen, Hung Le, Thao Le, Dang Nguyen, **Haripriya Harikumar**, Truyen Tran, Santu Rana, and Svetha Venkatesh. Revisiting the dataset bias problem from a statistical perspective. In European Conference on Artificial Intelligence **(ECAI)**, 2024
- Kien Do, **Haripriya Harikumar**, Hung Le, Dung Nguyen, Truyen Tran, Santu Rana, Dang Nguyen, Willy Susilo, and Svetha Venkatesh. Towards effective and robust neural trojan defenses via input filtering. In *In European Conference on Computer Vision (**ECCV**)*, 2022
- Kien Do, Thai Hung Le, Dung Nguyen, Dang Nguyen, **Haripriya Harikumar**, Truyen Tran, Santu Rana, and Svetha Venkatesh. Momentum adversarial distillation: Handling large distribution shifts in data-free knowledge distillation. *In Advances in Neural Information Processing Systems (**NeurIPS**)*, 2022
- **Haripriya Harikumar**, Vuong Le, Santu Rana, S Bhattacharya, Sunil Gupta, and Svetha Venkatesh. Scalable backdoor detection in neural networks. In *In European Conference on Machine Learning and Knowledge Discovery in Databases (**ECML**)*, 2021
- **Haripriya Harikumar**, Kien Do, Santu Rana, Sunil Gupta, and Svetha Venkatesh. Semantic host-free trojan attack. *arXiv*, 2021
- **Haripriya Harikumar**, Santu Rana, Sunil Gupta, Thin Nguyen, Ramachandra Kaimal, and Svetha Venkatesh. Differentially private prescriptive analytics. In *In International conference on data mining (**ICDM**)*, 2018
- **Haripriya Harikumar**, Santu Rana, Sunil Gupta, Thin Nguyen, Ramachandra Kaimal, and Svetha Venkatesh. Prescriptive analytics through constrained bayesian optimization. In *In Pacific-Asia Conference Knowledge Discovery and Data Mining (**PAKDD**)*, 2018
- **Haripriya Harikumar**, Thin Nguyen, Santu Rana, Sunil Gupta, Ramachandra Kaimal, and Svetha Venkatesh. Extracting key challenges in achieving sobriety through shared subspace learning. In *In Advanced Data Mining and Applications (**ADMA**)*, 2016
- **Haripriya Harikumar**, Thin Nguyen, Sunil Gupta, Santu Rana, Ramachandra Kaimal, and Svetha Venkatesh. Understanding behavioral differences between short and long-term drinking abstainers from social media. In *In Advanced Data Mining and Applications (**ADMA**)*, 2016

- Prema Nedungadi and **Haripriya Harikumar**. Feature and search space reduction for label-dependent multi-label classification. In *In International Conference on Computer and Communication Technologies*, 2016

- **Haripriya Harikumar**, CP Prathibhamol, Yashwant R Pai, M Sai Sandeep, Arya M Sankar, Srinivas Nag Veerla, and Prema Nedungadi. Multi label prediction using association rule generation and simple k-means. In *In International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)*, 2016

- **Haripriya Harikumar**, Shaji Amrutha, R Veena, and Prema Nedungadi. Integrating apriori with paired k-means for cluster fixed mixed data. In *In International Symposium on Women in Computing and Informatics*, 2015

- **Haripriya Harikumar**, R Devisree, Dinesh Pooja, and Prema Nedungadi. A comparative performance analysis of self organizing maps on weight initializations using different strategies. In *In International Conference on Advances in Computing and Communications (ICACC)*, 2015

- Prema Nedungadi and **Haripriya Harikumar**. Exploiting label dependency and feature similarity for multi-label classification. In *In International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2014

- Prema Nedungadi, **Haripriya Harikumar**, and Maneesha Ramesh. A high performance hybrid algorithm for text classification. In *In International Conference on the Applications of Digital Information and Web Technologies (ICADIWT)*, 2014

- Harikumar Sandhya, **Haripriya Harikumar**, and MR Kaimal. Implementation of projected clustering based on sql queries and udfs in relational databases. In *In Recent Advances in Intelligent Computational Systems (RAICS)*, 2013

Refereed journal papers

- **Haripriya Harikumar**, Santu Rana, Sunil Gupta, Thin Nguyen, Ramachandra Kaimal, and Svetha Venkatesh. Prescriptive analytics with differential privacy. *In International Journal of Data Science and Analytics (**JDSA**)*, 2022

- **Haripriya Harikumar**, Thomas P Quinn, Santu Rana, Sunil Gupta, and Svetha Venkatesh. Personalized single-cell networks: a framework to predict the response of any gene to any drug for any patient. *In BioData Mining*, 2021

- Md Shahid Ansari, Dinesh Jain, **Haripriya Harikumar**, Santu Rana, Sunil Gupta, Sandeep Budhiraja, and Svetha Venkatesh. Identification of predictors and model for predicting prolonged length of stay in dengue patients. *In Health Care Management Science*, 2021

Professional Services

Program Committee/Conference Reviewer

- Computer Vision and Pattern Recognition (CVPR): 2021 – current
- Neural Information Processing Systems (NeurIPS): 2023 – current
- International Conference on Machine Learning (ICML): 2024 – current
- Association for the Advancement of Artificial Intelligence (AAAI): 2022 – current
- International Conference on Computer Vision (ICCV): 2023, and 2025
- International Conference on Artificial Intelligence and Statistics (AISTAT): 2024 – current
- European Conference on Computer Vision (ECCV): 2024, 2022
- Women in Computer Vision (WiCV) Workshop: 2024
- Women in Machine Learning (WiML) Workshop: 2025
- Winter Conference on Applications of Computer Vision (WACV): 2025
- Asian Conference on Computer Vision (ACCV): 2022
- British Machine Vision Conference (BMVC): 2021
- European Conference on Signal Processing (EUSIPCO): 2019 – 2021
- Australasian Data Mining Conference (AusDM): 2019

Journal Reviewer

- Journal Mathematics - 2025
- Springer Nature Computer Science - 2023
- Transactions on Knowledge and Data Engineering - 2018 – 2019

Other

- Involved in Ph.D. student interviews, Postdocs and other staff selection (Machine Learning Engineer and AI System Technician) interviews.
- Volunteer as a Fire Warden at Applied Artificial Intelligence Institute, Deakin University, 2022 - 2024.

Membership

- **Women in AI**: An active member in Women in AI community (Profile).

News

- **Celebrating our Women in Science** (link).

Grants

- **ELSA Grant 2025**: Awarded a grant from European Lighthouse on Secure and Safe AI funds (ELSA) to attend a CISPA-ELLIS-Summer School 2025 in Trustworthy AI (August 4-8, 2025).

- **SaTML Travel Grant 2025**: Awarded travel grant for attending IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 2025.

Scholarships

- **Higher Degree by Research (HDR) Scholarship, 2015**: Awarded Higher Degree by Research Scholarship from Deakin University, Australia to pursue a PhD.

- **Engineering Scholarship, 2007**: Awarded a merit-based, government-funded scholarship from Kerala, India, to pursue a four-year undergraduate degree (B.Tech in Information Technology).

Training

Recruiting Staff at Manchester to Academic/Research Posts

University of Manchester, UK 2025

Foundations of Teaching and Learning

University of Manchester, UK 2025

Certifications

Project Management for Professionals

RMIT University, Australia

Technical Skills

Programming Languages: Python, C, C++, R, Matlab, Java, and PostgreSQL.

Libraries: Numpy, NLTK, Pytorch, Tensorflow, Pandas, Scipy, Scikit-learn, BeautifulSoup, LIWC, PRAW (Reddit API wrapper), LDA(topic extractor), sklearn, OpenCV, PIL, LLM and CLIP (large language vision model)

Visualisation tools: Matplotlib, Seaborn, plotly, PowerBI, and Tableau.

Platforms: Windows, Ubuntu, and Ubuntu servers.

Frameworks: VSCode, Latex, git, Monday.com, and LyX.

Online Courses: Fundamentals of Reinforcement Learning (University of Alberta - 2023, audited), Machine learning (Stanford University - 2015, audited).

Soft Skills: Teamwork, Organisation, Problem-solving, Communication, and Project Management.

References

provide upon request