Siddharth Swaroop, Ph.D.

Curriculum Vitae

Data to Actionable Knowledge (DtAK) lab,
School of Engineering and Applied Sciences,
Harvard University, U.S.
✉ *siddharth@seas.harvard.edu*
siddharthswaroop.github.io

Professional Experience

2022– **Postdoctoral Fellow**, *School of Engineering and Applied Sciences, Harvard University*.
Working in the Data to Actionable Knowledge group, with Prof Finale Doshi-Velez.

Summer 2021 **Internship at Microsoft Research**, *Cambridge, UK*.
Investigated ways of leveraging deep language models, such as BERT and GPT-3, to improve an existing large-scale knowledge base construction system. Worked with Pavel Myshkov and Tom Minka.

Summer 2018 **Internship at Microsoft Research**, *Cambridge, UK*.
Working on improvements to a large-scale knowledge base construction system. Worked with Martin Kukla and John Winn.

Education

2017–2022 **PhD in Engineering**, *Churchill College, University of Cambridge*.
Thesis: Probabilistic Continual Learning using Neural Networks.
Supervised by Prof Richard E Turner, advised by Prof Carl E Rasmussen.
Funding: EPSRC Doctoral Training Partnership, Microsoft Research EMEA PhD Award.

2013–2017 **BA and MEng**, *Churchill College, University of Cambridge*.
Graduated with BA and MEng (Honours pass with Distinction)
Awarded Charles Lamb Prize (one candidate in electrical or information engineering)
Achieved a 1st Class with Distinction in Third Year Examinations (Ranked 1st out of ∼250 students)
Achieved a 1st Class Result in Second Year Examinations (1st percentile of ∼250 students)
Achieved a 1st Class Result in First Year Examinations (3rd percentile of ∼300 students)

Publications

Underlined authors are students I mentored.

Selected publications

1. **Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure**.
Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, Finale Doshi-Velez.
International Conference on Intelligent User Interfaces, 2024.
2. **Improving Continual Learning by Accurate Gradient Reconstructions of the Past**.
Erik Daxberger, **Siddharth Swaroop**, Kazuki Osawa, Rio Yokota, Richard E Turner, José Miguel Hernández-Lobato, Mohammad Emtiyaz Khan.
Transactions on Machine Learning Research, 2023 & Conference on Lifelong Learning Agents, 2024.
3. **Probabilistic Continual Learning using Neural Networks**.
Siddharth Swaroop.
PhD thesis.
4. **Knowledge-Adaptation Priors**.
Mohammad Emtiyaz Khan*, **Siddharth Swaroop***.
Neural Information Processing Systems, 2021.
5. **Continual Deep Learning by Functional Regularisation of Memorable Past**.
Pingbo Pan*, **Siddharth Swaroop***, Alexander Immer, Runa Eschenhagen, Richard E Turner, Mohammad Emtiyaz Khan.
Oral presentation at *Neural Information Processing Systems, 2020* (top 1% of submissions).
Oral presentation at *LifeLongML workshop, ICML 2020. Continual Learning workshop, ICML 2020*.

Manuscripts under review

6. **Connecting Federated ADMM to Bayes**.
   Siddharth Swaroop, Mohammad Emtiyaz Khan, Finale Doshi-Velez.
   Under review.
7. **Personalising AI assistance to overreliance rate in AI-assisted decision making**.
   Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, Finale Doshi-Velez.
   Under review.
8. **Aligning Human-AI Knowledge With Contrastive Explanations Can Improve Human Decision-Making Skills in AI-Assisted Decision-Making**.
   Zana Buçinca, **Siddharth Swaroop**, Amanda E. Paluch, Finale Doshi-Velez, Krzysztof Z. Gajos.
   Under review.
9. **Gradient Reconstruction for Continual Learning Memory Selection**.
   Erik Wang*, Theodora Boulouta*, Weiwei Pan, **Siddharth Swaroop**, Finale Doshi-Velez.
   Under review.

Additional published research

10. **AI Agents & Liability – Mapping Insights from ML and HCI Research to Policy**.
    Connor Dunlop*, Weiwei Pan*, Julia Smakman*, Lisa Soder*, **Siddharth Swaroop***.
    Workshop on Socially Responsible Language Modelling Research, NeurIPS 2024.
11. **Levels of Autonomy: Liability in the age of AI Agents**.
    Julia Smakman*, Lisa Soder*, Connor Dunlop*, Weiwei Pan, **Siddharth Swaroop**.
    Workshop on Socially Responsible Language Modelling Research, NeurIPS 2024.
12. **Understanding Model Bias Requires Systematic Probing Across Tasks**.
    Helen Zhao*, Susannah Su*, Soline Boussard*, **Siddharth Swaroop**, Finale Doshi-Velez, Weiwei Pan.
    Workshop on Socially Responsible Language Modelling Research, NeurIPS 2024.
13. **Where do doctors disagree? Characterizing Decision Points for Safe Reinforcement Learning in Choosing Vasopressor Treatment**.
    Esther Brown, Shivam Raval, Alex Rojas, Jiayu Yao, Sonali Parbhoo, Leo A Celi, **Siddharth Swaroop**, Weiwei Pan, Finale Doshi-Velez.
    AMIA Journal, 2024.
14. **AMBER: An Entropy Maximizing Environment Design Algorithm for Inverse Reinforcement Learning**.
    Paul Nitschke, Lars Lien Ankile, Eura Nofshin, **Siddharth Swaroop**, Finale Doshi-Velez, Weiwei Pan.
    Models of Human Feedback for AI Alignment Workshop, ICML 2024.
15. **Rethinking Discount Regularization: New Interpretations, Unintended Consequences, and Solutions for Regularization in Reinforcement Learning**.
    Sarah Rathnam, Sonali Parbhoo, **Siddharth Swaroop**, Weiwei Pan, Susan Murphy, Finale Doshi-Velez.
    Journal of Machine Learning Research, 2024.
16. **Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning**.
    Zana Buçinca, **Siddharth Swaroop**, Amanda E. Paluch, Susan A. Murphy, Krzysztof Z. Gajos.
    arXiv preprint: 2403.05911.
17. **Reinforcement Learning Interventions on Boundedly Rational Human Agents in Frictionful Tasks**.
    Eura Shin, **Siddharth Swaroop**, Weiwei Pan, Susan A. Murphy, Finale Doshi-Velez.
    International Conference on Autonomous Agents and Multiagent Systems, 2024.
18. **Lifelong Learning for Deep Neural Networks with Bayesian Principles**.
    Cuong V. Nguyen*, **Siddharth Swaroop***, Thang D. Bui, Yingzhen Li, Richard E. Turner.
    Book chapter in "Towards Human Brain Inspired Lifelong Learning", 2024.

19. **Discovering User Types: Mapping User Traits by Task-Specific Behaviors in Reinforcement Learning**.
Lars L Ankile*, Brian S Ham*, Kevin Mao, Eura Shin, **Siddharth Swaroop**, Finale Doshi-Velez, Weiwei Pan.
Honourable mention for best paper award at *Artificial Intelligence & Human-Computer Interaction Workshop, ICML 2023*.

20. **Memory Maps to Understand Models**.
Dharmesh Tailor, Paul Edmund Chang, **Siddharth Swaroop**, Eric Nalisnick, Arno Solin, Mohammad Emtiyaz Khan.
Duality Principles for Modern ML Workshop, ICML 2023.

21. **Adaptive interventions for both accuracy and time in AI-assisted human decision making**.
Siddharth Swaroop, Zana Buçinca, Finale Doshi-Velez.
Artificial Intelligence & Human-Computer Interaction Workshop, ICML 2023.

22. **Soft prompting might be a bug, not a feature**.
Luke Bailey*, Gustaf Ahdritz*, Anat Kleiman*, **Siddharth Swaroop**, Finale Doshi-Velez, Weiwei Pan.
Challenges of Deploying Generative AI Workshop, ICML 2023.

23. **Differentially private partitioned variational inference**.
Mikko A. Heikkilä, Matthew Ashman, **Siddharth Swaroop**, Richard E Turner, Antti Honkela.
Transactions on Machine Learning Research, 2023.

24. **Modeling Mobile Health Users as Reinforcement Learning Agents**.
Eura Shin, **Siddharth Swaroop**, Weiwei Pan, Susan Murphy, Finale Doshi-Velez.
Contributed talk at *Workshop on AI for Behavior Change, AAAI 2023*.

25. **Partitioned Variational Inference: A Framework for Probabilistic Federated Learning**.
Matthew Ashman, Thang D Bui, Cuong V Nguyen, Stratis Markou, Adrian Weller, **Siddharth Swaroop**, Richard E Turner.
arXiv preprint: 2202.12275.

26. **Collapsed Variational Bounds for Bayesian Neural Networks**.
Marcin B Tomczak, **Siddharth Swaroop**, Andrew YK Foong, Richard E Turner.
Neural Information Processing Systems, 2021.

27. **Generalized Variational Continual Learning**.
Noel Loo, **Siddharth Swaroop**, Richard E Turner.
International Conference on Learning Representations, 2021.

28. **Efficient Low Rank Gaussian Variational Inference for Neural Networks**.
Marcin B Tomczak, **Siddharth Swaroop**, Richard E Turner.
Neural Information Processing Systems, 2020.

29. **Combining Variational Continual Learning with FiLM Layers**.
Noel Loo, **Siddharth Swaroop**, Richard E Turner.
Oral presentation at *LifeLongML workshop, ICML 2020. Continual Learning workshop, ICML 2020*.

30. **Practical Deep Learning with Bayesian Principles**.
Kazuki Osawa, **Siddharth Swaroop***, Anirudh Jain*, Runa Eschenhagen, Richard E Turner, Rio Yokota, Mohammad Emtiyaz Khan.
Neural Information Processing Systems, 2019.

31. **Differentially Private Federated Variational Inference**.
Mrinank Sharma, Michael Hutchinson, **Siddharth Swaroop**, Antti Honkela, Richard E Turner.
Privacy in Machine Learning Workshop, NeurIPS 2019.

32. **Improving and Understanding Variational Continual Learning**.
Siddharth Swaroop, Thang D Bui, Cuong V Nguyen, Richard E Turner.
Oral presentation at *Continual Learning Workshop, NeurIPS 2018*.

33. **Partitioned Variational Inference: A unified framework encompassing federated and continual learning**.
Thang D Bui, Cuong V Nguyen, **Siddharth Swaroop**, Richard E Turner.
arXiv preprint: 1811.11206, Spotlight at *Bayesian Deep Learning Workshop, NeurIPS 2018*.

34. **Neural network ensembles and variational inference revisited**.
Marcin B Tomczak, **Siddharth Swaroop**, Richard E Turner.
Advances in Approximate Bayesian Inference Symposium 2018.

35. **Understanding Expectation Propagation**.
Siddharth Swaroop and Richard E Turner.
Advances in Approximate Bayesian Inference workshop, NIPS 2017.

Talks

Nov 2024 **Quick and Accurate Knowledge Adaptation in Machine Learning**.
CS Colloquium Series, Harvard, US, 2024 (Invited speaker)

June 2024 **Federated learning with a Laplace approximation**.
2nd Bayes-Duality Workshop, Tokyo, Japan, 2024 (Invited speaker)

April 2024 **Quick and accurate knowledge adaptation in machine learning**.
SPIRAL Seminar Series, Northeastern University, USA
Tufts CS Colloquium, TUFTS, USA

March 2024 **Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure**.
International Conference on Intelligent User Interfaces, Greenville, SC, USA

June 2023 **Bayesian continual learning and adaptation**.
Bayes-Duality Workshop, Tokyo, Japan, 2023 (Invited speaker)

June 2022 **Knowledge-adaptation priors for continual learning**.
Workshop on Continual Learning in Computer Vision, CVPR 2023 (Invited talk)

Dec 2021 **Adaptive and Robust Learning with Bayes**.
Bayesian Deep Learning workshop, NeurIPS 2021 (Invited talk, with Emtiyaz Khan & Dharmesh Tailor)

July 2021 **Continual Deep Learning with Bayesian Principles**.
Theory and Foundations of Continual Learning workshop, ICML 2021 (Invited oral)
Machine learning reading group, Microsoft Research Cambridge, UK
Healthcare intelligence reading group, Microsoft Research Cambridge, UK

Apr-Jun 2021 **Continual Deep Learning by Functional Regularisation of Memorable Past**.
OATML, University of Oxford, UK
DtAK lab, Harvard University, USA
University of Toronto, Canada

2020 **Continual Deep Learning by Functional Regularisation of Memorable Past**.
NeurIPS 2020 (Oral presentation)
LifeLongML workshop, ICML 2020 (Oral presentation)

May 2020 **Natural-gradient variational inference for Bayesian Neural Networks**.
Gatsby Machine Learning MLJC, University College London, UK

Nov 2019 **Variational inference: scaling Bayesian neural networks, distributed inference, and continual learning**.
ARM, Cambridge, UK

April 2019 **Improving Variational Continual Learning**.
RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

Dec 2018 **Improving and Understanding Variational Continual Learning**.
Continual Learning Workshop, NeurIPS 2018, Montréal, Canada (Oral presentation)

Teaching and Mentoring

Guest lectures and teaching.
Guest lectures for *Advanced Topics in Data Science*, and *Inverse Reinforcement Learning* (2024).
Taught for 2 years in the *Computational Science and Engineering Capstone Project* course (2023-2024).
Co-instructor on and helped design new iteration of Harvard's course on *Diversity, Inclusion, Ethics and Leadership in Tech* (2024).

Mentorship.
Mentored 60+ students in research (from undergraduate to PhD) over seven years, leading to 15+ theses and 15+ publications.

Supervising small groups of undergraduate students, *University of Cambridge*.
Engineering Tripos IIA 3F8 (Inference), 2018, 2019, 2021
Engineering Tripos IIA 3F3 (Signal and pattern processing), 2018
Engineering Tripos IB Paper 7 (Mathematical methods), 2017

Awards and Funding

2020 Microsoft Research EMEA PhD Award ($15k)

2017-2021 Honorary Vice-Chancellor's Award, Cambridge Trust (length of PhD)

2020 Instrumental in obtaining £100k unrestricted gift from ARM to group for work on BNNs

2014–2017 Charles Lamb Prize (2017), The Institution of Civil Engineers Baker prize (2016), Bill Browne Engineering Prize and Scholar of Churchill College (awarded every year 2014-2017)

Academic Service

Committee experience: Workshop chair for the Conference on Lifelong Learning Agents, 2025; Sponsorship chair for the Advances in Approximate Bayesian Inference, 2024; Pre-registration paper chair for the ContinualAI Un-conference, 2023; Organiser of Continual Lifelong Learning workshop at the Asian Conference on Machine Learning, 2022.

Area Chair: AABI 2024, ICLR 2025.

Reviewing for Journals: JMLR; TMLR; IEEE TAI.

Reviewing for Conferences: NeurIPS (2020-2022); AISTATS (2021-2023, top reviewer in 2022); ICML (top reviewer in 2020, expert reviewer in 2021, 2024); ICLR (2020-2023, top/highlighted reviewer in 2021 & 2022); UAI (2023); CoLLAs (Senior PC 2022-2024); CHI (2024); ACML (2019).

Reviewing for Workshops: ICBINB, NeurIPS 2020; Workshop on Continual Learning, ICML 2020; Uncertainty & Robustness in Deep Learning, ICML 2020; AABI Symposium, 2018-2021 & 2023; Bayesian Deep Learning Workshop, NeurIPS 2019 & NeurIPS 2021; AAAI23 Bridge Continual Causality.

Additional mentorship activities.
Hosted online mentorship sessions on mementor.net (2021-present).
Mentor at mentorship roundtables at the Asian Conference on Machine Learning 2022 and the Conference on Lifelong Learning Agents 2024.