Alex J. Chan

alexjchan.com | +44 (0) 777 112 8977 | me@alexjchan.com
LinkedIn: alexjchan | GitHub: XanderJC

I am interested in developing **safe** and **robust** machine learning systems that behave in the way we expect, and thus able to work with humans effectively. My work has often explored **inverse reinforcement learning** and **imitation learning** in order to learn from humans, and I'm interested in using this knowledge and **generative AI** such as **large language models** to develop **personalised, human-centric,** decision making systems.

Experience

Spotify London, UK

Research Scientist Intern - *Satisfaction, Interaction and Algorithms (SIA) team* *Jun. - Oct. 2023*

- Project on improving the podcast content moderation pipeline by incorporating large language models (LLMs).
- Created various cultural media datasets through web-scraping in order to finetune a family of models to be more aware of transient news and location/cultural specific issues to improve detection of potential violative content.
- Incorporated these cultural models as tools for larger models to generate personalised guidance to content moderators, especially in the case when there is a cultural mismatch between the moderator and original content.
- Project resulted in writing an academic paper, filing for a patent, and I received a full-time return offer.

Stanford Existential Risk Initiative Berkeley, CA

Researcher *Nov. 2022 – Jun. 2023*

- Working with Dr Owain Evans examining LLMs, in particular their ability and willingness to be deceptive and how that might be detected with only black-box query knowledge.
- Designed and implemented experiments in Python using LLMs with various APIs and HuggingFace models.
- Fine-tuned and ran large (up to 65 billion parameters) models on remote servers with extensive shell scripting using various distributed computing packages including PyTorch Distributed and DeepSpeed.

Microsoft Research Cambridge, UK

PhD Scholar *Oct. 2020 – Present*

- Student researcher co-supervised by Dr Aditya Nori (previously Dr Danielle Belgrave before she joined DeepMind) "A Smart Care System for Healthcare using Contextual Reinforcement Learning".

University of Cambridge Cambridge, UK

Research Assistant - *van der Schaar lab* *Dec 2019. – Aug. 2020*

- Research project on uncertainty calibration in Bayesian neural networks under distributional shift.
- Developed a novel method for improving calibration in transfer learning - resulted in publication at ICML.

University College London London, UK

Research Assistant - *Dr Sam Livingstone's Research Group* *Jun. – Aug. 2018*

- Implemented a variety of MCMC samplers and practical convergence diagnostic methods in Python, running experiments to compare their output with recently proposed theoretical bounds.

Education

University of Cambridge Cambridge, UK

PhD Machine Learning - *Supervisor: Professor Mihaela van der Schaar* *Oct. 2020 – Present*

- Understanding decision making through interpretable imitation learning and inverse reinforcement learning.
- First-author publications in all three of the major machine learning conferences: ICML, NeurIPS, and ICLR.
- In total my published work has been cited more than 150 times, and I have an h-index of 6.

MPhil Machine Learning and Machine Intelligence *Oct. 2019 – Sep. 2020*

- Awarded with Commendation and an average of 79%. 92% for my thesis "Interpretable Policy Learning" - developing interpretable imitation learning algorithms for decision making in high stakes environments.
- Taught component focusing on probabilistic machine learning, with modules on natural language processing, reinforcement learning, and computational neuroscience.

University College London London, UK

BSc Statistics *Oct. 2016 – June 2019*

- 1^{st} Class Honours - 81% average. Achieved 84% on my final year project on probabilistic deep learning, focusing on flexible Bayesian approximations in deep neural networks.
- Modules include significant mathematical and statistical content covering probability, linear models, mathematical analysis, and advanced linear algebra.

Peer-Reviewed Publications

How to Catch an AI Liar: Lie Detection in Black-box LLMs by Asking Unrelated Questions
L. Pacchiardi*, **A. J. Chan***, S. Mindermann, I. Moscovitz, A. Pan, Y. Gal, O. Evans, and J. M. Brauner. *International Conference on Learning Representations (ICLR)* 2024.

Optimising Human-AI Collaboration by Finding Convincing Explanations
A. J. Chan, A. Hüyük, and M. van der Schaar. *NeurIPS XAI in Action* 2023.

AllSim: Systematic Simulation and Benchmarking of Repeated Resource Allocation Policies in Multi-User Systems with Varying Resources
J. Berrevoets, D. Jarrett, **A. J. Chan**, and M. van der Schaar. *Proceedings of the Neural Information Processing Systems (NeurIPS) track on Datasets and Benchmarks* 2023.

GAUCHE: A Library for Gaussian Processes in Chemistry
R. Griffiths, L. Klarner, H. Moss, A. Ravuri, S. T. Truong, Y. Du, S. Don Stanton, G. Tom, B.Ranković, A. R. Jamasb, A. Deshwal, J. Schwartz, A. Tripp, G. Kell, S. Frieder, A. Bourached, **A. J. Chan**, J. Moss, C. Guo, J. P. Dürholt, S. Chaurasia, J. W. Park, F. Strieth-Kalthoff, A. Lee, B. Cheng, A. Aspuru-Guzik, P. Schwaller, J. Tang. *Advances in Neural Information Processing Systems (NeurIPS)* 2023.

Synthetic Model Combination: A New Machine Learning Method for Pharmacometric Model Ensembling
A. J. Chan, R. Peck, M. Gibbs, and M. van der Schaar. *CPT: Pharmacometrics & Systems Pharmacology* 2023.

Synthetic Model Combination: An Instance-wise Approach to Unsupervised Ensemble Learning
A. J. Chan and M. van der Schaar. *Advances in Neural Information Processing Systems (NeurIPS)* 2022.

Practical Approaches for Fair Learning with Multitype and Multivariate Sensitive Attributes
T. Liu, **A. J. Chan**, B. van Breugel, and M. van der Schaar. *NeurIPS Algorithmic Fairness through the Lens of Causality and Privacy (AFCP)* 2022.

Inverse Online Learning: Understanding Non-Stationary and Reactionary Policies
A. J. Chan, A. Curth, and M. van der Schaar. *International Conference on Learning Representations (ICLR)* 2022.

POETREE: Interpretable Policy Learning with Adaptive Decision Trees
A. Pace, **A. J. Chan**, and M. van der Schaar. *International Conference on Learning Representations (ICLR)* 2022.

The Medkit-learn(ing) Environment: Medical Decision Modelling through Simulation
A. J. Chan, I. Bica, A. Hüyük, D. Jarrett, and M. van der Schaar. *Proceedings of the Neural Information Processing Systems (NeurIPS) track on Datasets and Benchmarks* 2021.

Scalable Bayesian Inverse Reinforcement Learning
A. J. Chan and M. van der Schaar. *International Conference on Learning Representations (ICLR)* 2021.

Generative Time Series Modelling with Fourier Flows
A. M. Alaa, **A. J. Chan**, and M. van der Schaar. *International Conference on Learning Representations (ICLR)* 2021.

Unlabelled Data Improves Bayesian Uncertainty Calibration under Covariate Shift
A. J. Chan, A. M. Alaa, Z. Qian, and M. van der Schaar. *International Conference on Machine Learning (ICML)* 2020.

Preprints / Under-Review

Harmonizing Global Voices: Culturally-Aware Models for Enhanced Content Moderation
A. J. Chan, J. L. R. García, F. Silvestri, C. O'Donnel, and K. Palla. https://arxiv.org/abs/2312.02401. Under review at *Nature Machine Intelligence*.

Dense Reward for Free in Reinforcement Learning from Human Feedback
A. J. Chan, H. Sun, S. Holt, and M. van der Schaar. https://arxiv.org/abs/2402.00782. Under review at ICML.

*Equal contribution.

Awards/Prizes

Machine Learning Alignment Theory Scholarship (20k USD)
- Scholarship funding for research project on how large language models can lie when articulating decision rules.

Microsoft Research PhD Scholarship (≈160k GBP)
- Received the award for full funding of my PhD co-supervised with Microsoft Research (Dr Aditya Nori, and Dr Danielle Belgrave) "A Smart Care System for Healthcare using Contextual Reinforcement Learning".

G-Research PhD Prize in Maths and Data Science (7k GBP)
- Runner up in the G-Research competition for best draft PhD dissertation.

EPSRC Vacation Grant (≈2k GBP)
- Awarded funding grant by the Engineering and Physical Sciences Research Council to conduct a research project during the summer on Markov chain Monte Carlo methods.

Skills

Machine Languages: Python, R, MATLAB, PostgreSQL, HTML.
Human Languages: English, Conversational French.
Libraries/Tools: PyTorch, JAX, TensorFlow, Transformers, DeepSpeed, pandas, NumPy, Matplotlib, Git, Azure, GCP.

Supervision

University of Cambridge Cambridge, UK
MPhil Machine Learning and Machine Intelligence Theses *Mar. – Aug. 2021*
- **Tennison Liu:** *Fair Policy Learning.* (Work published at AFCP 2022).
- **Alizée Pace:** *Adaptive Decision Tree Policies* (Resulted in a Spotlight at ICLR 2022).

University of Oxford Oxford, UK
MSc Statistical Science Thesis *Mar. – Aug. 2021*
- **Yuling Chen:** *Clustered Bayesian Inverse Reinforcement Learning Via Variational Inference.*

Service

AAAI Workshop on Representation Learning for Responsible Human-Centric AI
Invited Area Chair *2023*

NeurIPS SyntheticData4ML Workshop
Program Committee / Area Chair *2022*

NeurIPS Workshop on Causality for Real-world Impact
Invited Reviewer *2022*

ICML/ICLR/NeurIPS
Invited Reviewer ICML21-23, NeurIPS21-23, ICLR21-24 *2021 – Present*

Code First Girls
Volunteer course instructor for "Introduction to Python Programming" *Jan. – March 2021*

University of Cambridge Cambridge, UK
Club Captain, Wolfson College Boat Club *Aug. 2021 – Aug. 2022*
- As Captain of the boat club, I was in charge of the overall running of the club, organising the training of the members as well as broader events and the alumni network.

University College London London, UK
Vice President/Treasurer - Pure Krav Maga Society *Oct. 2018 – June 2019*
- I oversaw the organisation and finances behind sessions while helping to run classes as a trainee instructor.

Welfare Officer, Effective Altruism Society *Oct. 2018 – June 2019*
- I was responsible for engaging with the wider community to develop more of an understanding of the aims of Effective Altruism as well as looking out for the welfare of our members and helping develop the society further.

Electric Eels Swimming Club Windsor, UK
Volunteer Swimming Coach *2011 – 2015*
- I spent four years volunteering with the club, which aims to provide special coaching for children with Down syndrome, coaching both groups and 1-on-1 at a range of swimming ability
- I became ASA certified in Teaching Aquatics, allowing me to develop my technical and communication skills to be a more effective coach.