Alex J. Chan

alexjchan.com | me@alexjchan.com | LinkedIn: alexjchan | Google Scholar: Alex J. Chan | GitHub: XanderJC

I am interested in developing **safe** and **effective** machine learning systems that can successfully complete complex tasks in the real world. My work has often explored **(inverse) reinforcement learning**, **imitation learning** and **simulation** in order to learn from intrinsic rewards and humans, applied most recently to **large language/multimodal models** for autonomous computer-using agents.

Experience

Salesforce London, UK

Director / Principle MTS – Autonomous Agents ML Lead *Jul. 2025 – Present*

- I lead our machine learning research agenda, focussed mainly on post-training vision-language(-action) models for agentic tasks. This includes fine-tuning for GUI grounding and reasoning, as well as RL techniques such as PPO and inference-time model-based tree-search.

Convergence Labs (*acquired by Salesforce*) London, UK

Founding Member of Technical Staff (Research Scientist) *Sep. 2024 – Jul 2025*

- First Research Scientist employee, took our capability to train and serve our own in-house models from 0-1, moving away from 3rd party (e.g. OpenAI) to serve $> 50\%$ of the traffic from our $125,000+$ users.
- Developed and trained models for our agent, Proxy, that achieved state-of-the-art WebVoyager performance, using supervised fine-tuning and reinforcement learning on large vision-language models (VLMs).
- Set-up data flywheels for automatically collecting training examples and preference data as well as building infrastructure for training and serving models on our large scale H100 cluster using FastAPI, Ray, and vLLM.

Spotify London, UK

Research Scientist *Mar. - Sep. 2024*

- Worked on reinforcement learning training of LLMs on domains including tool use and controllable generation.
- Outperformed GPT-4 models using a 2B local model optimised with PPO for playlist generation.

Research Scientist Intern *Jun. - Oct. 2023*

- Developed a framework for culturally-dependent content moderation, including fine-tuning a family of models for cultural awareness on specially curated and web media datasets to provide personalised guidance to moderators.

Stanford Existential Risk Initiative Berkeley, CA

Researcher *- Dr Owain Evans' Research Group* *Nov. 2022 – Jun. 2023*

- Investigated deceptive behaviour in LLMs and possible detection with only black-box query knowledge.
- Fine-tuned, ran, and evaluated large (up to 65 billion parameters) models on remote servers with extensive shell scripting using various distributed computing packages including PyTorch Distributed, FSDP, and DeepSpeed.

Microsoft Research Cambridge, UK

PhD Scholar *Oct. 2020 – Mar. 2024*

- Student researcher co-supervised by Dr Aditya Nori and Dr. Danielle Belgrave for our project: "A Smart Care System for Healthcare using Contextual Reinforcement Learning".

Education

University of Cambridge Cambridge, UK

PhD Machine Learning *- Supervisor: Professor Mihaela van der Schaar* *Oct. 2020 – Mar. 2024*

- Thesis Title: Aligning Models for Human-Centric Decision Systems.
- 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 500 times, and I have an h-index of 10.

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.

University College London London, UK

BSc Statistics *Oct. 2016 – June 2019*

- 1^{st} Class Honours - 81% average. 84% for my thesis on flexible Bayesian approximations in deep neural networks.

Conference Publications

Discovering Preference Optimization Algorithms with and for Large Language Models
C. Lu, S. Holt, C. Fanconi, **A. J. Chan**, J. Foerster, M. van der Schaar, and R. T. Lange. *Advances in Neural Information Processing Systems (NeurIPS)* 2024.

Dense Reward for Free in Reinforcement Learning from Human Feedback
A. J. Chan, H. Sun, S. Holt, and M. van der Schaar. *International Conference on Machine Learning (ICML)* 2024.

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.

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

Journal Publications

When is Off-Policy Evaluation Useful in Contextual Bandits? A Data-Centric Perspective
Hao Sun*, **A. J. Chan***, Nabeel Seedat, A. Hüyük, and M. van der Schaar. *Journal of Data-centric Machine Learning Research (DMLR)* 2024.

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.

Workshop Publications

WebGames: Challenging General-Purpose Web-Browsing AI Agents
G. Thomas, **A. J. Chan**, J. Kang, W. Wu, F. Christianos, F. Greenlee, A. Toulis, and M. Purtorab. *ICML Workshop on Computer Use Agents* 2025.

LM2: Large Memory Models
J. Kang, W. Wu, F. Christianos, **A. J. Chan**, F. Greenlee, G. Thomas, M. Purtorab, and A. Toulis. *ICLR Workshop on Reasoning and Planning for Large Language Models* 2025.

*Equal contribution.

Actions Speak Louder than Words: Superficial Fairness Alignment in LLMs
Q. Wei, **A. J. Chan**, L. Goetz, D. Watson, and M. van der Schaar. *ICLR Workshop on Reliable and Responsible Foundation Models* 2024.

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

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.

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.

Awards/Prizes

OpenAI Superalignment Fellowship (150k USD)
- Grant for research project on activation steering generalisation in large language models.

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, TF, Transformers, DeepSpeed, TRL, Azure, GCP, Ray, Kubernetes, Slurm, vLLM.

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.

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.