Alan Chan

alan.chan@mila.quebec achan.ca Substack Google Scholar

Research Statement

I work on how to manage risks from AI agents as they become increasingly capable of automating a wide variety of commercial, scientific, governmental, and personal activities. Much of my current focus is on figuring out what technical infrastructure we need to understand and respond to problems that agents may cause.

Academic Background

Ph.D. Computer Science 2024 (expected)
Université de Montréal and Mila, Montréal, Canada
- Advised by Nicolas Le Roux and David Krueger.

M.Sc. Computer Science 2020
University of Alberta, Edmonton, Canada
- Specialization in statistical machine learning, advised by Martha White.
- Thesis: *Greedification Operators for Policy Optimization: Investigating Forward and Reverse KL Divergences.*

B.Sc. Mathematics 2018
University of Alberta, Edmonton, Canada
- Thesis title: *A Comparison of Efficient Particle Filters.*

Professional Experience

Research Scholar Oct. 2023 - Oct. 2024
Centre for the Governance of AI, Oxford, UK
- Performing research on technical methods for enabling better governance of AI agents. See Visibility into AI Agents and IDs for AI Systems.
- Policy advising for various stakeholders in government and civil society.

Research Visitor Oct. 2022 - Apr. 2023
Cambridge University, Cambridge, UK
- Hosted by Dr. David Krueger.
- Projects included: Reclaiming the Digital Commons: Public Data Trusts for Training Data, Harms from Increasingly Agentic Algorithmic Systems, Characterizing Manipulation from AI Systems.

Contract Researcher Feb. 2023 - Mar. 2023
CEIMIA, Montréal, Canada
- Performed research on gaps in the proposed bill C-27 on AI regulation before the Canadian House of Commons.
- Co-authored a policy report.

Summer Research Fellow June 2022 - Dec. 2022
Center on Long-Term Risk, London, UK
- Supervised by Jesse Clifton and Julian Stastny.
- Led a project to evaluate cooperativeness in language models.

Intern Research Scientist May 2019 - Jan. 2020
Huawei Canada - Noah's Ark, Edmonton, Canada
- Supervised by Daniel Graves.

- Contributed to a project on improving the stability of value-based algorithms in reinforcement learning.

Student Researcher Jan. 2018 - Apr. 2019
University of Alberta, Edmonton, Canada
- Supervised by Dr. Michael Kouritzin.
- Developed faster particle filters for dynamic time series estimation.

Student Researcher May 2018 - June 2018
University of Alberta, Edmonton, Canada Edmonton, Canada
- Supervised by Dr. Xinwei Yu.
- Developed understanding the differential geometry of map projections of the earth.

Student Researcher May 2017 - Aug. 2017
University of Alberta, Edmonton, Canada
- Supervised by Dr. Xinwei Yu.
- Studied the energy method for solution of non-linear PDEs.

Research Student May 2017 - Aug. 2017
University of Alberta, Edmonton, Canada
- Supervised by Dr. Jack Tuszynski.
- Performed data-mining to help determine the mechanism of action for a specific drug.

Research Student May 2016 - Aug. 2016
University of Alberta, Edmonton, Canada
- Supervised by Dr. Jack Tuszynksi.
- Led a project on the use of fractal dimension as a feature in the diagnosis of breast cancer.

Grants and Awards

- Long-Term Future Fund Grant (for Cambridge research visit): 10 500 GBP 2023
- Open Philanthropy Early-Career Funding: 16 000 USD 2022-24
- Top 10% reviewer for NeurIPS 2020
- Queen Elizabeth II Graduate Scholarship: 6 000 CAD 2018
- NSERC Summer Research Grant: 8 000 CAD 2018
- NSERC Summer Research Grant: 8 000 CAD 2017
- Yahya Mathematics Scholarship 2017
- Erist Wilson Sheldon Memorial Prize in Mathematics 2017
- A. Murray Thomas Gibson Memorial Scholarship in Mathematics 2016
- University of Alberta Summer Research Grant: 8 000 CAD 2016
- Cyril G. Wates Memorial Scholarship 2016
- International Baccalaureate Diploma Scholarship 2014
- Robert Tegler Entrance Scholarship 2014
- Alexander Rutherford Scholarship 2014

Refereed Papers

9. **Alan Chan**, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, Markus Anderljung. Visibility into AI Agents, *FAccT 2024*, June 2024.
8. Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, **Alan Chan**, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, Dylan Hadfield-Menell . Black-Box Access is Insufficient for Rigorous AI Audits , *FAccT 2024*, June 2024.
7. Micah Carroll*, **Alan Chan*** (joint first author), Henry Ashton. Characterizing Manipulation from AI Systems, *EAAMO 2023*, Oct. 2023.
6. **Alan Chan**, Herbie Bradley, Nitarshan Rajkumar. Reclaiming the Digital Commons: Public Data Trusts for Training Data, *AIES 2023*, Aug. 2023.
5. **Alan Chan**, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, Tegan Maharaj. Harms from Increasingly Agentic Algorithmic Systems, *FAccT 2023*, June 2023.
4. **Alan Chan**, Hugo Silva, Sungsu Lim, Tadashi Kozuno, A. Rupam Mahmood, Martha White. Greedification Operators for Policy Optimization: Investigating Forward and Reverse KL Divergences, *JMLR*, Aug. 2022.
3. Somjit Nath, Vincent Liu, **Alan Chan**, Adam White, Martha White. Training Recurrent Neural Networks Online by Learning Explicit State Variables, *ICLR 2020*, Apr. 2020.
2. Kris De Asis, **Alan Chan**, Silviu Pitis, Richard S. Sutton, and Daniel Graves. Fixed-horizon Temporal Difference Methods for Stable Reinforcement Learning, *AAAI 2020*, Feb. 2020.
1. **Alan Chan** and Jack A. Tuszynski. Automatic Prediction of Tumour Malignancy in Breast Cancer with Fractal Dimension *Royal Society Open Science*, Dec. 2016.

Other Papers

10. **Alan Chan**, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, Markus Anderljung. IDs for AI Systems, *under review*.
9. Ross Gruetzemacher, **Alan Chan**, Štěpán Los, Kevin Frazier, Siméon Campos, Matija Franklin, James Fox, Jose Hernandez-Orallo, Christin Manning, Philip Tomei, Kyle Kilian. An International Consortium for AI Risk Evaluations, *Socially Responsible Language Modelling Research Workshop at NeurIPS 2023*, Dec. 2023.
8. Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, **Alan Chan***, Jesse Clifton*. Welfare Diplomacy: Benchmarking Language Model Cooperation, *Preprint*, Sep. 2023.
7. **Alan Chan**, Maxime Riché, Jesse Clifton. Towards the Scalable Evaluation of Cooperativeness in Language Models, *Preprint*, Mar. 2023.
6. Montréal Society and Artificial Intelligence Collective. Comments on the Toronto Police Services Board Proposed Policy on AI Technologies, Dec. 2021.

5. Andrew Jacobsen, **Alan Chan**. Parameter-free Gradient Temporal Difference Learning, May 2021.
4. **Alan Chan**. Scoring Rules for Performative Binary Prediction, *Strategic ML workshop at NeurIPS 2021*, Dec. 2021.
3. **Alan Chan**. Loss of Control: Normal Accidents and AI Systems, *Responsible AI Workshop at ICLR 2021*, May 2021.
2. **Alan Chan**, Chinasa Okolo, Zach C. Terner, Angelina Wang. The Limits of Global Inclusion in AI Development, *Rethinking Diversity in AI workshop at AAAI 2021*, Feb. 2021.
1. **Alan Chan**, Kris De Asis, Richard S. Sutton. Inverse Policy Evaluation for Value-based Sequential Decision-making, Aug. 2020.

Articles

1. Jan Brauner*, **Alan Chan***. AI Poses Doomsday Risks—But That Doesn't Mean We Shouldn't Talk About Present Harms Too, *Time*, Aug. 2023.

Supervision

10. Neel Alex. GovAI Summer Fellowship. June 2024 - Aug 2024
9. Arjun Karanam. GovAI Summer Fellowship. June 2024 - Aug 2024
8. Carson Ezell. Existential Risk Alliance Fellowship. May 2024 - Now
7. Ben Bucknall. Krueger Lab Internship. June 2023 - Jan. 2024
6. Gabriel Mukobi. Existential Risk Alliance Fellowship. June 2023 - Sep. 2023
5. Gabriel Weil. PIBBSS Fellowship. June 2023 - Aug. 2023
4. Sammy Martin. PIBBSS Fellowship. June 2023 - Aug. 2023
3. Stuart Burrell, Niki Kyriacou, Hadyn Cheong, Aafiya Hussain, Talha Chafekar. Cambridge AI Safety Labs. Dec. 2022 - Sep. 2023
2. Anish Upadhayay. Undergraduate research. June 2022 - Sep. 2022
1. Kay Kozaronek. CHERI Summer Fellowship. June 2022 - Aug. 2022

Talks

9. Governing AI Agents. ICML 2024 workshop on Trustworthy Multi-modal Foundation Models and AI Agents, Vienna, Austria. Jul. 2024
8. Infrastructure for AI Agents. RAND, Washington, DC, USA. May 2024
7. Technical AI Governance. Vector AI Safety Group, Toronto, Canada. Sep. 2023
6. Technical AI Governance. Concordia AI, Beijing, China. Jul. 2023
5. Briding AI Safety with Other Fields. PIBBSS Speaker Series, Prague, Czechia (Virtual). Jul. 2023
4. The Case for AI Governance. Cambridge AI Safety Hub, Cambridge, UK. Apr. 2023
3. Alignment. Digital Law and Innovation Society, Edmonton, Canada (Virtual). Mar. 2021
2. Colonialism and AI Development. Mila, Montréal, QC, Canada (Virtual). Feb. 2021
1. Problems with Fair Machine Learning. University of Alberta, Edmonton, Canada. July 2020

Other Activities

- Reviewer for NeurIPS, ICML, ICLR, TMLR, FAccT. 2019 - Present
- Organizer of the Socially Responsible Language Modelling (SoLaR) workshop at NeurIPS 2023. Dec. 2023
- Member of the Mila EDI committee. Aug. 2021 - May 2022
- Organizer of the Mila AI Governance Reading Group. Nov. 2020 - Mar. 2022
- Lecturer and mentor for the MISE Foundation. June 2021 - July 2021
- Teacher Assistant for the AI4Good Lab June 2020 - July 2020
- Volunteer for the UAlberta Sexual Assault Centre. Sep. 2017 - May 2019
- VP Training of the UAlberta Debate Society. May 2017 - May 2018
- Volunteer for The Landing Sep. 2015 - Sep. 2017
- Competitive debate regionally, nationally, and internationally. 2011 - 2018