Jonathan Colaço Carr

jonathan.colaco-carr@mail.mcgill.ca | j-c-carr.github.io | linkedin.com/in/jonathan-colaco-carr

Education

M.Sc in Computer Science, McGill University 2023 – 2025 (Expected)

- Supervisors: Prakash Panangaden and Doina Precup. GPA: 4.0/4.0.

Visiting Graduate Student (Computer Science), Stanford University 2024 – 2025

- Supervisor: Benjamin Van Roy

B.Sc in Honours Mathematics, McGill University 2018 – 2023

- Minors in Computer Science and Physics. GPA: 3.8/4.0.

Research Experience

Visiting Student Researcher, Stanford University October 2024 – March 2025

- Supervisor: Benjamin Van Roy
- Developed a sequential decision-making algorithm that accounts for societal disagreement when choosing actions.

Visiting Student Researcher, UC Berkeley (CHAI Lab) June 2024 – September 2024

- Supervisors: Cameron Allen and Stuart Russell
- Designed a skill discovery algorithm that allows reinforcement learning agents to avoid negative side effects in underspecified tasks.

Student Researcher, Mila – Quebec AI Institute May 2022 – May 2024

- Supervisors: Prakash Panangaden and Doina Precup
- Audited the Anthropic Helpful and Harmless preference dataset for label quality and topic coverage. Fine-tuned nine LLMs on this dataset, showing how demographic imbalances lead to disparities in model safety.
- Proved that it was possible for sequential decision-making agents to satisfy human preferences, even when those preferences cannot be expressed by a reward function.

Student Research Engineer, McGill Space Institute May 2021 – December 2021

- Supervisor: Adrian Liu
- Implemented a deep computer vision algorithm (a modified U-Net) to recover signals from noisy telescope data.

Papers in Preparation

[1] **J. Colaço Carr**, B. Van Roy, D. Precup, and P. Panangaden. *Handling Societal Disagreements in Long-Term Decision Problems*. In preparation for the *International Conference on Learning Representations (ICLR)*. 2025.

Conference Publications

* denotes co-first authorship.

[2] K. Chehbouni*, **J. Colaço Carr***, Y. More, J. C. Cheung, and G. Farnadi. "Beyond the Safety Bundle: Auditing the Helpful and Harmless Dataset". In: *Proceedings of the Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL)*. 2025. **Selected for an oral presentation**.

[3] **J. Colaço Carr**, P. Panangaden, and D. Precup. "Conditions on Preference Relations that Guarantee the Existence of Optimal Policies". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2024.

Journal Publications

[4] **J. Colaço Carr**, Q. Sun, and C. Allen. "Focused Skill Discovery: Learning to Control Specific State Variables while Minimizing Side Effects". In: *Reinforcement Learning Journal (RLJ)*. 2025.

[5] J. Kennedy, **J. Colaço Carr**, S. Gagnon-Hartman, A. Liu, J. Mirocha, and Y. Cui. "Machine-Learning Recovery of Foreground Wedge-Removed 21-cm Light Cones for High-z Galaxy Mapping". In: *Monthly Notices of the Royal Astronomical Society (MNRAS)*. 2024.

Work Experience

Lead Software Developer, Hortus AI December 2024 – Present

- Lead a team of four software developers with startup founder, Thomas Krendl Gilbert, to create a platform that helps government entities manage AI alignment and safety concerns in procurement decisions.
- Improved the user interface through iterative design, integrating user feedback from focus groups and usability studies with the GovAI coalition.

Teaching Assistant, McGill University January 2024 – April 2024

- Held weekly office hours and created assignments for a graduate-level course in machine learning (COMP 551).

Associate Machine Learning Developer, AltaML January 2022 – April 2022

- Implemented decision trees and clustering algorithms to identify cross-selling opportunities and enhance customer segmentation. Presented technical insights to the client's sales and marketing teams.

Extracurricular Experience

Software Contributor, Center for Human-Compatible AI July 2024

- Built a docker container for the *DignityIndex*, a tool that detects disrespectful language in political discourse.

Co-founder, McGill AI Safety x Law Group April 2022 – December 2023

- Co-founded a 16-week reading program for McGill Law students interested in AI regulation. Organized a panel discussion with speakers from AI Governance & Safety Canada.

VP Technology, McGill Research and Sustainability Network April 2021 – April 2022

- Organized monthly panel discussions with students, professors and professionals interested in sustainability research.

Editorial Board Member, The McGill Tribune April 2019 – April 2020

- Collaborated with journalists and graphic designers to publish online articles for McGill's largest student newspaper.

Awards

McGill Graduate Mobility Award October 2024

Research award for McGill graduate students studying abroad, determined by academic standing.

MITACS Globalink Research Award October 2024

Competitive research award for Canadian graduate students studying abroad.

NSERC Canada Graduate Scholarship – Master's May 2023

Graduate scholarship determined by academic excellence, research potential and interpersonal skills.

McGill Computer Science Undergraduate Research Award May 2022

Summer research award determined by academic record and research aptitude.

McGill Space Institute Summer Undergraduate Research Award May 2021

Summer research award determined by academic record and extra-curricular leadership.

RBC x Microsoft AI for Social Impact Challenge, 2nd Place Prize April 2020

Led a team of four students in an eight-month competition to design a robotic arm for automating textile recycling. Placed 2nd out of 167 teams across Canada.