

# Henri Lemoine

[henri123lemoine@gmail.com](mailto:henri123lemoine@gmail.com)  
<https://github.com/henri123lemoine>  
Montreal, QC.

McGill University  
Joint Major in **Computer Science** and Statistics, 2021-December 2024

## MARS 2.0 ----- Research Participant

January 2025 - Present

Developing novel AI control frameworks and safety protocols on Tyler Tracy's MARS (Mentorship for Alignment Research Students) team, organized by the Cambridge AI Safety Hub.

Key contributions:

- Development of a feedback monitoring protocol where AI systems iteratively revise outputs based on safety feedback, which pareto-dominates other protocols in our tests
- Contributing to formal modelling of AI control evaluation framework
- Building standardized evaluation infrastructure for testing control protocols across multiple code generation tasks and datasets

## Zeroth Technologies Inc. -- Software Engineer

September 2023 - August 2024

Software Engineer and Platform Architect at [Zeroth Technologies Inc.](#), a software consulting firm targeting startups. Expert in finding robust and cutting-edge data science and machine learning architectures and solutions. Clients:

- **AlphaScript**: Led ML development for a healthcare chatbot AlphaScript, utilizing Meta's LLAMA 2 and SOTA prompt engineering to automate private and local patient transcript processing and information retrieval for SOAP notes. [ Python, Flask, LLAMA 2, Cuda, WhisperX ]
- **Flojoy**: Developed advanced software blocks for [Mecademic](#) robot arm, contributing to a [demo](#) that successfully attracted investor interest. Spearheaded the integration of diverse movement functionalities, exploring the arm's versatility and improving user experience. [ Python, flojoy scripting language, Mecademic ]
- **Marmott Énergies & Olameter**: Oversaw a comprehensive automation project for Marmott Énergies, focusing on the analysis of geological data to assess overburden depth and bedrock thermal conductivity. Collaborated closely with Olameter data science experts to develop innovative solutions and refactored and streamlined the evaluation process, reducing costs and time-to-market for renewable energy initiatives. [ Python, PostGIS, PostgreSQL, SQLAlchemy, Excel, React, FastHTML, AWS S3 ]

## AI Safety.info ----- Software Engineer

March 2023 - August 2023

In March 2023, two friends and I developed a retrieval augmented generation (RAG) platform called [AlignmentSearch](#) ([technical details](#)), a conversational agent that curates the [Alignment Research Dataset](#) (ARD). When asked a query, we source relevant passages from a dataset of academic papers, books, articles, and other resources on AI safety and alignment. We then prompt GPT-3 to synthesize a response while citing back to these passages, and parse out inline citations.

I set up a database for efficient querying of sources using Pinecone, optimized semantic similarity model by fine-tuning terminology-specific embeddings layer, developed automated MySQL database updates for multiple sources, and engineered question-answering assistant prompt for maximum accuracy and helpfulness.

During the following summer, I was employed by Rob Miles' organization ([aisafety.info](#)) to expand this prototype, now available [here](#).

## AI Safety Camp #9 ----- Project Developer

March 2023 - June 2023

Wrote prompts for the prompt library of the AISC#9 Cyborgism project. Essentially, it's a library of prompts that can be used with base (pretraining-only) or fine-tuned LLMs, that lets alignment researchers generate ideas, get automated high-variance and high-quality feedback, find sources for their papers, explore the workings of base models, and much more.

Led the development of [Cyborg-Duck](#), an Obsidian plugin integrated with the [AlignmentSearch](#) chatbot and the prompt library, to assist AI alignment researchers in order to accelerate their research.

## Center for AI Safety ----- Machine Learning Safety Scholars

June 2022 - August 2022

Graduated from an intensive 9-week training program run by the Center for AI Safety to study deep learning for AI safety research, led by Dan Hendrycks. The main topics covered:

- Machine Learning (MIT 6.036 Introduction to Machine Learning course)
- Deep Learning (Convolutional Neural Networks, Recurrent Neural Networks, Attention, Self-Attention, and Transformers)
- AI Safety (Adversarial robustness (robustness to adversarially-generated inputs), and Out-Of-Distribution detection)

Completed a research project on potential solutions to the Eliciting Latent Knowledge (ELK) challenge, the task of reliably translating latent knowledge from an AI system's ontology into our own, when it is otherwise expected and by-default incentivized to conceal it.

## AI Launch Labs ----- AI Research Intern

June 2020 - August 2020

Engaged in a machine learning project to train an RL-truck for self-parking, using reinforcement learning algorithms DDPG and TD3.

## Marmott Énergies ----- Software Engineer

June 2021 - August 2021

Contributed to the development of an autonomous H-VAC inspection robot. focusing on sensor interfacing and 3D modelling.

---

## Projects

---

### Open Low Cost Humanoid

[ Python, PyTorch, IsaacGym, CUDA ]

(WIP) Undergraduate honours research project supervised by Professor Hsiu-Chin Lin for COMP 400 Projects in Computer Science. Developing an accessible, open-source, low-cost humanoid robot for learning-based control. Implemented and optimized PPO for robot locomotion, achieving stable walking gaits in Isaac Gym simulations.

### Lifelogging

[ Rust, FFmpeg, AWS S3 ]

(WIP) Developed a low memory, Rust-based continuous audio recording server for personal lifelogging. Implemented short-term disk persistence and long-term AWS S3 storage solutions. Integrated audio visualization and Opus compression for efficient data management. Optimized audio buffer writes using SIMD instructions, achieving 11.74x performance gain in benchmarks.

### EfficientZero Paper Replication

[ Python, PyTorch ]

Replicated the [Mastering Atari Games with Limited Data](#) paper for COMP 579: Reinforcement Learning with Professor Doina Precup.

### Grokking Paper Replication

[ Python, PyTorch ]

Replicated the [Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets](#) paper by OpenAI for the final project of the Applied Machine Learning class with Professor Isabeau Prémont-Schwarz.

---

## Competitions and Contests

---

### Astral Codex Ten 2023 Prediction Contest

97th / 3296 participants (Top 3%)

### Salem Center forecasting tournament

14th place / 999 participants (Silver medal prize)

### McGill PokerBots Hackathon

1st place

### McHacks 2023

1st place NLP Cohere prize

### MAIS Hacks 2022

3rd place prize

---

## Leadership

---

### Effective Altruism McGill --- Co-President

May 2022 - August 2024

As Co-President, I orchestrated EA McGill's weekly club activities and occasional speaker events and social events, including an Introductory EA Fellowship, an AGI Safety Fundamentals reading group, and EAGxBerkeley and EAGxToronto conferences, enhancing community engagement and knowledge on critical issues.

### AI Alignment McGill ----- Founder and President

March 2023 - May 2024

I founded the club in the context of building AlignmentSearch, in order to find collaborators at McGill. Starting from nothing, we grew the club to over 50 members, with regular weekly meetups to discuss alignment research, social events, and hackathons.

Presented AI alignment to McGill Projects, a large organization of McGill developers, and as a guest speaker at MAIS Hacks 2023, the largest AI hackathon at McGill, to spread interest in AI alignment.

### Astral Codex Ten Montreal -- Organizer

April 2023 - Present

Lead the Montreal chapter of the Astral Codex Ten community meetups, fostering a space for intellectual exchange and networking.

---

## Skills

---

**Proficient With:** Python, Pytorch, Git, SQLAlchemy, Postgres, Machine Learning (esp. Reinforcement Learning)

**Previous Experience With:** Rust, AWS S3, CUDA, R, C++, C, MySQL, Pinecone, Swift, Arduino, Typescript, HTML + CSS, Javascript, React, OCaml, Java, C#, Bash, Django, Redis, Deepgram, Google Cloud, TensorFlow, NumPy, LaTeX