Kellin Pelrine

Education

Ph.D. Machine Learning, McGill University, 2020-2025 (expected)
Supervisors: Reihaneh Rabbany, Jean-François Godbout
M.A. Economics, Yale University, 2018-2019
M.S. Applied Mathematics, University of Colorado Boulder, 2017-2018
B.A. Economics and Mathematics, University of Colorado Boulder, 2014-2018
Class Rank 1, "Summa cum laude in economics" and "With Distinction"

Current Affiliations

McGill University, PhD Candidate
Mila – Quebec AI Institute, Student Member
Centre for the Study of Democratic Citizenship, Student Member
FAR AI, Research Advisor

Scientific Objectives and Experience

Overall Goal:

Make generative AI a transformatively positive force (instead of a potentially catastrophically negative one) on our ability to find reliable information and build knowledge.

Areas:

AI safety, AI agents, NLP, evaluation, reliability, robustness

Applications:

Manipulation, persuasion, deception, education, political polarization, Go

Professional Experience

Stitch

CTO, December 2023 – November 2024
Led development of a graphical interface for LLMs: https://getstitch.ai/
A visual system enables more organized, faster, and deeper interactions.

Mutual

Chief Scientist, June 2023 – December 2023
Scientific Advisor, May 2023 – June 2023
Led R&D to create consistent, reliable, and powerful generative AI agents.
Improved reliability led to contract with Gaggle Studios for a game character agent, and our AI tutor deployed in multiple classes in Fall 2023.

FAR AI

Research Advisor, June 2023 – Present
Research Scientist Intern, December 2022 – June 2023
Combined ML and Go expertise to better understand vulnerabilities in superhuman Go-playing systems, cited in US Senate hearings, and the "man who beat the machine".
Paper 1 reported in Financial Times, The Times, Ars Technica, Vice
Paper 2 reported in Nature, Ars Technica, Scientific American

Found vulnerabilities in OpenAI models, influencing OpenAI's decision to limit the release of GPT-4 finetuning.
Evaluated scaling laws for data poisoning: more capable models are becoming more vulnerable.

Machine Learning Sports
CTO, October 2019 – December 2020
Created a pitch recommendation system powered by machine learning to help professional baseball pitchers optimize their pitch selection.

Go (strategy board game)
Professional teacher, freelance
Volunteer teacher, multiple Go clubs and online
Past students including but not limited to:
Paul Barchilon, American Go Foundation Teacher of the Year 2006
Eric Wainwright, 2022 US Go Congress Codirector
David Weiss, Boulder Kids and Teens Go Club founder and teacher
Private lessons, group lectures/workshops/classes, game reviews.

Research Leadership

Towards Reliable Misinformation Mitigation
Overall Project Leader, January 2023 – Present
Leading subprojects in detection, retrieval, uncertainty quantification, explainability, datasets, evaluation, and more.
In coordination with co-PIs Reihaneh Rabbany and Jean-François Godbout, building towards tools that will help every individual make better decisions and avoid being misled by malicious manipulation from human to superhuman AI.
$600K funding from Mila, Canadian Heritage Foundation, IVADO, CSDC, SPAR.
2 conference publications, multiple works in progress.
Leading, supervising, and mentoring to facilitate the collaboration of over 30 researchers, undergrad through post-doctorate (ordered roughly by seniority):
Daniel Zhao (Research Scientist, MIT)
Austin Welch (Senior Applied Scientist, AWS Generative AI)
Maximilian Puelma Touzel (Postdoc, Mila)
Andreea Musulan (Postdoc, IVADO)

Gabrielle Peloquin-Skulski (PhD candidate, MIT)
Bijean Ghafouri (PhD candidate, USC)
Aarash Feizi (PhD candidate, McGill/Mila)
Zachary Yang (PhD candidate, McGill/Mila)
Anne Imouza (PhD candidate, McGill/Mila)
Joel Christoph (PhD candidate, European University Institute)
Nikita Agarwal (Machine Learning Researcher, Mayo Clinic)
Jacob-Junqi Tian (Associate Applied ML Specialist, Vector Institute)
Ethan Kosak-Hine (Developer, Atomic Weapons Establishment)
Tom Gibbs (Independent Researcher)
Master's students: Caleb Gupta, Camille Thibault, Mayank Goel, Meilina Reksoprodjo, Michael Walters, Shahrad Mohammadzadeh, Tyler Vergho
Undergraduate students: Annaliese Bissell, Florence Laflamme, James Zhou, Laurence Liang, Lynn Feng, Mauricio Rivera, Hao (Peter) Yu, Svetlana Zhuk, Swagat Bhowmik, Veronica Xia, Yury Orlovskiy

Temporal Graph Learning Workshop @ NeurIPS

Co-organizer, 2022 and 2023
Communications chair. Lead advertising and recruitment of speakers and panelists.
Opening remarks 2022.
Contributed to numerous aspects, such as proposal, recruiting, reviewing, etc.

Axiom Futures Fellowship

Mentor, Summer 2024 (and invited for upcoming Winter 2025)
Invited to mentor two full-time, funded fellowship recipients.
Project: Simulations to Solve Societal-Scale Manipulation.
Mentees: Gayatri Krishna Kumar, Sneheel Sarangi.

Supervised Program for Alignment Research (SPAR)

Mentor, Spring+Fall 2023, Spring+Summer 2024
Invited to mentor over 25 junior researchers in AI safety, from undergraduate through senior applied scientist. Multiple projects in misinformation and manipulation, LLM vulnerabilities, and education.
Mentees co-authored 2 published conference papers and 3 under review, 5 workshop papers and 7 under review.
Mentees (ordered roughly chronologically): Caleb Gupta, Joel Christoph, Meilina Reksoprodjo, James Zhou, Lynn Feng, Mayank Goel, Raghav Ravi, Roman Hauksson, Tyler Vergho, Yury Orlovskiy, Arjun Verma, Arturs Semenuks, George Ingebretsen, Ruben Weijers, Gabrielle Castilho, Dylan Tabarini, Michael Walters, Will Cai, Ethan Kosak-Hine, George Ingebretsen, Jason Zhang, Julius Broomfield, Reihaneh Iranmanesh, Sara Pieri, Tom Gibbs, Nikita Agarwal, Austin Welch.

Thesis Supervision

Co-supervisor, master's thesis of Ruben Weijers, Utrecht University, 2024-Present

Honors and Awards

Research and Academic (graduate)

Doctoral Training Scholarship, CAD 58k, Fonds de Recherche du Québec, 2023
Graduate Excellence Award, CAD 48k total, McGill University, 5 times, 2020-2023
GREAT Award, CAD 2k total, McGill University, 2 times, 2022 and 2023
IVADO PhD Excellence Scholarship, CAD 75k, IVADO, 2021
Max Stern Recruitment Fellowship, CAD 14k, McGill University, 2020
Cowles Foundation Fellowship, USD 32k, Yale University, 2018

Academic (undergraduate)

Chancellor's Recognition Award, CU Boulder, for perfect GPA, 2018
Sieglinde Talbott Haller Economics Scholarship, CU Boulder, 2016 and 2017
Jim and Laura Marshall Scholarship, CU Boulder, mathematics, 2016
Flock Leader Scholarship, CU Boulder, 2015
CU Esteemed Scholars Program: President Joseph A. Sewall Award, 2014

Go (strategy board game)

U.S. Team Member, 2012 World Mind Sports Games, 2012
Lille, France. 95 countries were represented. Invitation based on U.S. ranking and tournament results.

U.S. Team Member, China-US Internet Go Tournament, 2020
One of 6 players selected by ranking to represent the U.S. in a friendship match with top Chinese amateurs.

Playoff for North American representative to Li Min Cup, 2014 and 2016
One of 8 participants by invitation.

Conference Papers

Web Retrieval Agents for Evidence-Based Misinformation Detection
Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, Kellin Pelrine
COLM 2024

Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4
Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, Reihaneh Rabbany
EMNLP 2023

SWEET: Weakly Supervised Person Name Extraction for Fighting Human Trafficking
Javin Liu*, Vidya Sujaya*, Peter Yu*, Pratheeksha Nair, Kellin Pelrine, Reihaneh Rabbany
Findings of EMNLP 2023

Party Prediction for Twitter

Kellin Pelrine, Anne Imouza, Gabrielle Desrosiers-Brisebois, Sacha Lévy, Jacob-Junqi Tian, Zachary Yang, Aarash Feizi, Cécile Amadoro, André Blais, Jean-François Godbout, Reihaneh Rabbany
ICWSM 2024

Adversarial Policies Beat Superhuman Go AIs
Tony Tong Wang[*], Adam Gleave[*], Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, Stuart Russell
ICML 2023 (Oral)

Towards Better Evaluation for Dynamic Link Prediction
Farimah Poursafaei[*], Andy Huang[*], Kellin Pelrine, Reihaneh Rabbany
NeurIPS Datasets and Benchmarks Track 2022

Extracting Person Names from User Generated Text: Named-Entity Recognition for Combating Human Trafficking
Yifei Li, Pratheeksha Nair, Kellin Pelrine, Reihaneh Rabbany
Findings of the Association for Computational Linguistics 2022

The Surprising Performance of Simple Baselines for Misinformation Detection
Kellin Pelrine[*], Jacob Danovitch[*], Reihaneh Rabbany
The Web Conference 2021

Conference Papers Under Review

Epistemic Integrity in Large Language Models
Bijean Ghafouri[*], Shahrad Mohammadzadeh[*], James Zhou, Pratheeksha Nair, Jacob-Junqi Tian, Mayank Goel, Reihaneh Rabbany, Jean-François Godbout, Kellin Pelrine
ICLR 2025

A Guide to Misinformation Detection Datasets
Camille Thibault, Gabrielle Péloquin-Skulski, Jacob-Junqi Tian, Florence Laflamme, Reihaneh Rabbany, Jean-François Godbout, Kellin Pelrine
ICLR 2025

Can Go AIs be adversarially robust?
Tom Tseng, Euan McLean, Kellin Pelrine[†], Tony Tong Wang[†], Adam Gleave[†]
AAAI 2025

Scaling Laws for Data Poisoning in LLMs
Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave[†], Kellin Pelrine[†]
AAAI 2025

The Structural Safety Generalization Problem
Tom Gibbs[*], Julius Broomfield[*], George Ingebretsen[*], Ethan Kosak-Hine[*], Tia Nasir, Jason Zhang, Reihaneh Iranmanesh, Sara Pieri, Reihaneh Rabbany, Kellin Pelrine
ACL ARR

[*] Equal contribution

[†] Equal advising

Workshop Papers

A Simulation System Towards Solving Societal-Scale Manipulation
Maximilian Puelma Touzel*, Sneheel Sarangi*, Austin Welch*, Gayatri Krishnakumar, Dan Zhao, Zachary Yang, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Andreea Musulan, Camille Thibault, Busra Tugce Gurbuz, Reihaneh Rabbany, Jean-François Godbout, Kellin Pelrine
NeurIPS 2024 Workshops: SoLaR, SATA, SafeGenAI

Epistemic Integrity in Large Language Models
Bijean Ghafouri, Shahrad Mohammadzadeh, James Zhou, Pratheeksha Nair, Jacob-Junqi Tian, Mayank Goel, Reihaneh Rabbany, Jean-François Godbout, Kellin Pelrine
SafeGenAI @ NeurIPS 2024

The Structural Safety Generalization Problem
Tom Gibbs, Julius Broomfield, George Ingebretsen, Ethan Kosak-Hine, Tia Nasir, Jason Zhang, Reihaneh Iranmanesh, Sara Pieri, Reihaneh Rabbany, Kellin Pelrine
SafeGenAI @ NeurIPS 2024

Decompose, Recompose, and Conquer: Multi-modal LLMs are Vulnerable to Compositional Adversarial Attacks in Multi-Image Queries
NeurIPS 2024 Workshops: RBFM, Red Teaming GenAI

Can Go AIs be adversarially robust?
Tom Tseng, Euan McLean, Kellin Pelrine†, Tony Tong Wang†, Adam Gleave†
Next Generation of AI Safety 2024

Web Retrieval Agents for Evidence-Based Misinformation Detection
Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, Kellin Pelrine
Workshop on Online Harms and Abuse 2024

An Evaluation of Language Models for Hyperpartisan Ideology Detection in Persian Twitter
Sahar Omidi Shayegan, Isar Nejadgholi, Kellin Pelrine, Hao Yu, Sacha Levy, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany
Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) 2024

Uncertainty Resolution in Misinformation Detection
Yury Orlovskiy, Camille Thibault, Anne Imouza, Jean-François Godbout, Reihaneh Rabbany, Kellin Pelrine
UncertaiNLP Workshop 2024

Combining Confidence Elicitation and Sample-based Methods for Uncertainty Quantification in Misinformation Mitigation
Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, Kellin Pelrine
UncertaiNLP Workshop 2024

Quantifying learning-style adaptation in effectiveness of LLM teaching
Ruben Weijers, Gabrielle Fidelis de Castilho, Jean-François Godbout, Reihaneh Rabbany, Kellin Pelrine

Personalization of Generative AI Workshop 2024

Comparing GPT-4 and Open-Source Language Models in Misinformation Mitigation
Tyler Vergho, Jean-Francois Godbout, Reihaneh Rabbany, Kellin Pelrine
Responsible Language Models (ReLM) 2024

Better Bridges Between Model and Real World
Kellin Pelrine
Canadian AI Conference Graduate Student Symposium 2023.

Active Keyword Selection to Track Evolving Topics on Twitter
Sacha Lévy, Farimah Poursafaei, Kellin Pelrine, Reihaneh Rabbany
Workshop on Utility-Driven Mining and Learning 2022

OPPVIS: Visualizing Online Partisan Polarization of COVID-19
Zachary Yang, Anne Imouza, Kellin Pelrine, Sacha Lévy, Jiewen Liu, Gabrielle Desrosiers-Brisebois, Jean-François Godbout, André Blais, Reihaneh Rabbany
Poster, IEEE Visualization & Visual Analytics 2021

Online Partisan Polarization of COVID-19
Zachary Yang, Anne Imouza, Kellin Pelrine, Sacha Lévy, Jiewen Liu, Gabrielle Desrosiers-Brisebois, Jean-François Godbout, André Blais, Reihaneh Rabbany
International Conference on Data Mining Workshops 2021

ComplexDataLab at WNUT-2020 Task 2: Detecting Informative COVID-19 Tweets by Attending over Linked Documents
Kellin Pelrine, Jacob Danovitch, Albert Orozco Camacho, Reihaneh Rabbany
Workshop on Noisy User-generated Text (WNUT) 2020

Other Research

Data Poisoning in LLMs: Jailbreak-Tuning and Scaling Laws
Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave†, Kellin Pelrine†
Earlier version at Workshop on Data Contamination 2024

Exploiting novel GPT-4 APIs
Kellin Pelrine*, Mohammad Taufeeque*, Michał Zając, Euan McLean, Adam Gleave. 2024.

Open, Closed, or Small Language Models for Text Classification?
Hao Yu, Zachary Yang, Kellin Pelrine, Jean Francois Godbout, Reihaneh Rabbany. 2023.

A Note on the Unconditional Bias of the Nadaraya-Watson Regression Estimator
Kellin Pelrine. Supervisor: Carlos Martins-Filho. Undergraduate Honors Thesis, 2018.

Referee

IEEE Transactions on Information Forensics and Security
ACM Computing Surveys

International Journal of Human-Computer Interaction
ICLR 2025
ACL ARR 2024
The Web Conference, 2023
NeurIPS Temporal Graph Learning Workshop, 2022 and 2023
NeurIPS, Datasets and Benchmarks Track, 2022
Workshop on Noisy User-generated Text, 2020

Invited Talks

Misinformation Detection with Generative AI
Research by: Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, Kellin Pelrine
Presented at American Political Science Association Annual Meeeting, 2024

Can Go AIs be adversarially robust?
Research by: Tom Tseng, Euan McLean, Kellin Pelrine[†], Tony Tong Wang[†], Adam Gleave[†]
Presented at Mila AI Safety Reading Group

Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4
Research by: Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, Reihaneh Rabbany
Presented at American Political Science Association Annual Meeeting, 2023
Presented at FAR Labs, 2023

Adversarial Policies Beat Superhuman Go Ais
Research by: Tony Tong Wang[*], Adam Gleave[*], Tom Tseng, Nora Belrose, Kellin Pelrine, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, Stuart Russell
Co-presenter: Tony Tong Wang
Presented at Cross Labs, 2023

Party Prediction for Twitter
Research by: Kellin Pelrine, Anne Imouza, Gabrielle Desrosiers-Brisebois, Sacha Lévy, Jacob-Junqi Tian, Zachary Yang, Aarash Feizi, Cécile Amadoro, André Blais, Jean-François Godbout, Reihaneh Rabbany
Presented at Université de Montréal Political Science, 2023
Presented at IVADO Digital October, 2022
Presented at American Political Science Association Annual Meeting, 2022

Social Graphs
Guest Lecture
Presented at COMP 599 - Network Science, McGill University, 2022
Presented at COMP 599 - Network Science, McGill University, 2021

Political Polarization on Social Media
Research by: Kellin Pelrine, Anne Imouza, Gabrielle Desrosiers-Brisebois, Zachary Yang, Sacha Lévy, Aarash Feizi, Jiewen Liu, André Blais, Jean-François Godbout, Reihaneh Rabbany

Presented at American Political Science Association Annual Meeting, 2021

Using Social Media Data to Measure Polarization
Guest Lecture
Presented at PLU6904A - Les États-Unis de Trump à Biden, Université de Montréal/CÉRIUM, 2021

Marginal GAN
Research by: Kellin Pelrine
Presented at CU Boulder Econometrics Workshop, November 2019

ShapeAttack: Genetic Algorithm for Shape-Constrained Adversarial Robustness Testing
Research by: Kellin Pelrine
Presented at CU Boulder Econometrics Workshop, October 2019