Kellin Pelrine

Website | Google Scholar

Leading cross-functional solutions on a foundation of technical research.

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"

Scientific Objectives and Experience

Goal: Make generative AI a transformatively positive force for society, instead of a potentially catastrophically negative one.
Selected Technical Areas: AI security, AI agents, evaluation, reliability, robustness.
Selected Applications: Persuasion, manipulation, education, political polarization, Go.

Professional Experience

FAR AI

Research Scientist, January 2025 – Present
Research Advisor, June 2023 – January 2025
Research Scientist Intern, December 2022 – June 2023

Combined AI and Go expertise to understand vulnerabilities in superhuman Go-playing AI, 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

Leading red-teaming and blue-teaming of fine-tuning vulnerabilities.
Vulnerabilities discovered in OpenAI models influenced OpenAI's decision to limit the release of GPT-4 finetuning.
Discovered jailbreak-tuning, the most severe blackbox fine-tuning attack.
Showed all current models, open and closed, are vulnerable. Led to OpenAI adding fine-tuning to bug bounty program.
Determined scaling trends for data poisoning: more capable models are becoming more vulnerable.
Secured as PI $50K compute funding from CAIS.

Leading projects studying AI persuasion.

Forged collaboration with MIT, Cornell, Université de Montréal, American University, Centro de Investigación y Docencia Económicas. All-senior team includes 5 professors representing 5+ disciplines.
Secured as co-PI $320K from Schmidt Sciences.

Leading AI capability and risk demos.

Stitch

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

Mutual

Cofounder and Chief Scientist, June 2023 – December 2023
Led R&D to create consistent, reliable, and powerful generative AI agents.

Machine Learning Sports

Cofounder and 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
Multiple tournament and award-winning students.

Research Leadership

Information Integrity in the GenAI Era

Project Director, January 2023 – Present
In coordination with PIs Reihaneh Rabbany and Jean-François Godbout, building tools that will help every person find reliable information and avoid being misled by malicious manipulation from both humans and AI.

Built and supervised over a dozen subprojects: evidence retrieval, credibility assessment, simulations, uncertainty quantification, explainability, datasets, evaluation, fieldbuilding, and more.

Secured $1.5M funding from Mila, Canadian Heritage Foundation, UKAISI, IVADO, FLI, CSDC, SPAR.

AI for Human Resilience

Project Director, February 2025 – Present
Demonstrated potential of AI for education as a fallible peer rather than authoritative teacher.
Science of building critical thinking, and AI and AI Safety literacy.
Secured $8500 seed grant from AI Safety Tactical Opportunities Fund and fiscal sponsorship from Ashgro.

Social Simulation with LLMs Workshop @ COLM
Co-organizer, 2025

Future of Information Integrity Research (FIIR) Workshop @ ICDM
Co-organizer, 2025

Temporal Graph Learning (TGL) Workshop @ NeurIPS
Co-organizer, 2022 and 2023
Communications chair. Led 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
Invited to mentor two full-time, funded fellowship recipients.
Project (Simulations to Solve Societal-Scale Manipulation) went on to secure over $500K funding, multiple publications, one mentee joining the lab as a graduate student, and more work in progress.
Mentees: Gayatri Krishna Kumar, Sneheel Sarangi.

Supervised Program for Alignment Research (SPAR)
Mentor, Spring+Fall 2023, Spring+Summer 2024, Spring 2025
Invited to mentor 34 researchers in AI safety, from undergraduate through senior applied scientist.
Mentees co-authored 7 published conference papers, dozens of workshop papers, and multiple works under review and in progress.
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, Toshali Goel, Kushal Dev, Luda Cohen, Sukanya Krishna, Hikaru Tsujimura, Ardy Haroen, Deeraj Nagothu.

Thesis Supervision
Co-supervisor, master's thesis of Ruben Weijers, Utrecht University, defended 2025

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.

Violin

Participant, 2 masterclasses of renowned pedagogue Zakhar Bron, 2012
By invitation and audition.

Conference Papers

A Guide to Misinformation Detection Datasets
Camille Thibault*, Jacob-Junqi Tian*, Gabrielle Péloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, **Kellin Pelrine**
Forthcoming, KDD Datasets and Benchmarks Track 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**
Forthcoming, Findings of ACL 2025

Simulating public discourse in digital societies by giving social media to multimodal AI agents

Maximilian Puelma Touzel*, Sneheel Sarangi*, Gayatri Krishnakumar*, Busra Tugce Gurbuz, Austin Welch, Zachary Yang, Andreea Musulan, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Camille Thibault, Reihaneh Rabbany, Jean-François Godbout†, Dan Zhao†, **Kellin Pelrine†**
Forthcoming, IJCAI Demo Track 2025

Veracity: An Open-Source AI Fact-Checking System
Taylor Lynn Curtis, Maximilian Puelma Touzel, William Garneau, Manon Gruaz , Mike Pinder, Li Wei Wang, Sukanya Krishna, Luda Cohen, Jean-François Godbout†, Reihaneh Rabbany†, **Kellin Pelrine†**
Forthcoming, IJCAI Demo Track 2025

Can Go AIs be adversarially robust?
Tom Tseng, Euan McLean, **Kellin Pelrine†**, Tony Tong Wang†, Adam Gleave†
AAAI 2025 (Oral)

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

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

* Equal contribution

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 ACL 2022

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

Conference Papers Under Review

Jailbreak-Tuning: Safeguards of Fine-Tunable Models are Illusory
Brendan Murphy, Dillon Bowen, Shahrad Mohammedzadeh, Julius Broomfield, Adam Gleave, **Kellin Pelrine**
ACL ARR 2025

Accidental Misalignment: Fine-Tuning Language Models Induces Unexpected Vulnerability
Punya Syon Pandey, Samuel Simko, **Kellin Pelrine**, Zhijing Jin
ACL ARR 2025

It's the Thought that Counts: Evaluating the Attempts of Frontier LLMs to Persuade on Harmful Topics
Matthew Kowal, Jasper Timm, Jean-François Godbout, Thomas Costello, Antonio A. Arechar, Gordon Pennycook, David Rand, Adam Gleave, **Kellin Pelrine**
NeurIPS Datasets and Benchmarks 2025

Blueprint: A Social Media User Dataset for LLM Persona Evaluation and Training
Aurélien Bück-Kaeffer, Je Qin Chooi, Dan Zhao, Maximilian Puelma Touzel, **Kellin Pelrine**, Jean-François Godbout, Reihaneh Rabbany, Zachary Yang
NeurIPS Datasets and Benchmarks 2025

Online Influence Campaigns: Strategies and Vulnerabilities
Andreea Musulan, Veronica Xia, Ethan Kosak-Hine, Tom Gibbs, Vidya Sujaya, Reihaneh Rabbany, Jean-François Godbout[†], **Kellin Pelrine[†]**
Big Data and Society

Workshop Papers

Simulating public discourse in digital societies by giving social media to multimodal AI agents
Maximilian Puelma Touzel[*], Sneheel Sarangi[*], Gayatri Krishnakumar[*], Busra Tugce Gurbuz, Austin Welch, Zachary Yang, Andreea Musulan, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Camille Thibault, Reihaneh Rabbany, Jean-François Godbout[†], Dan Zhao[†], **Kellin Pelrine[†]**
NLP4PI @ ACL 2025

From Intuition to Understanding: Using AI Peers to Overcome Physics Misconceptions

Ruben Weijers, Denton Wu, Hannah Betts, Tamara Jacod, Yuxiang Guan, Vidya Sujaya, Kushal Dev, Toshali Goel, William Delooze, Reihaneh Rabbany, Ying Wu, Jean-François Godbout, **Kellin Pelrine**
ICLR 2025 Workshops: AI4CHL (oral), FM-WILD, Bi-Align

Rethinking Anti-Misinformation AI
Vidya Sujaya, **Kellin Pelrine**, Andreea Musulan, Reihaneh Rabbany
ICLR 2025 Workshops: Bi-Align, HAIC

A Guide to Misinformation Detection Datasets
Camille Thibault*, Jacob-Junqi Tian*, Gabrielle Péloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, **Kellin Pelrine**
ICLR 2025 Workshops: MLDPR (spotlight), SCSL

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: SATA (oral), SoLaR, 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
Julius Broomfield, George Ingebretsen, Reihaneh Iranmanesh, Sara Pieri, Ethan Kosak-Hine, Tom Gibbs, Reihaneh Rabbany, **Kellin Pelrine**
NeurIPS 2024 Workshops: RBFM, Red Teaming GenAI

Scaling Laws for Data Poisoning in LLMs
Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave†, **Kellin Pelrine†**
Workshop on Data Contamination @ ACL 2024

Can Go AIs be adversarially robust?
Tom Tseng, Euan McLean, **Kellin Pelrine†**, Tony Tong Wang†, Adam Gleave†
Next Generation of AI Safety @ ICML 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 @ NAACL 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) @ LREC-COLING 2024

Uncertainty Resolution in Misinformation Detection
Yury Orlovskiy, Camille Thibault, Anne Imouza, Jean-François Godbout, Reihaneh Rabbany, **Kellin Pelrine**
UncertaiNLP Workshop @ EACL 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 @ EACL 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 @ EACL 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) @ AAAI 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 @ ICDM 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
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 (ICDMW) 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) @ EMNLP 2020

Other Research

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

International Journal of Human-Computer Interaction
IEEE Transactions on Information Forensics and Security
ACM Computing Surveys
International Journal of Human-Computer Interaction
ICLR 2025
ACL ARR 2024, 2025
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†
Co-presenter: Tom Tseng
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