Kellin Pelrine

Website | Google Scholar

Leading cross-functional solutions on technical foundations.

Education

Ph.D. Machine Learning, McGill University, 2020-2025
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"

Professional Experience

FAR AI

Member of Technical Staff, June 2025 – Present
Research Scientist, January 2025 – June 2025
Research Advisor, June 2023 – January 2025
Research Scientist Intern, December 2022 – June 2023

Leading the Integrity team, which aims to make AI trustworthy and secure.

Showed strongly superhuman capabilities will not be sufficient for robustness.
Paper 1 reported in Financial Times, The Times, Ars Technica, Vice
Paper 2 reported in Nature, Ars Technica, Scientific American
Orals at ICML and AAAI, and cited in US Senate testimony
Called the "man who beat the machine"

Leading projects to prevent misuse.

Our findings have influenced deployed safeguards of every frontier model company and multiple governments.

Built partnerships with multiple frontier companies (e.g., OpenAI).

9 confidential long-form reports delivered to frontier companies as first or last (managing) author. Exposed and helped fix multiple critical vulnerabilities and numerous partial vulnerabilities.

Discovered jailbreak-tuning, the most severe blackbox fine-tuning attack.
Showed increasingly capable AI is increasingly vulnerable to data poisoning.

Leading projects to prevent manipulation.
Built partnership with professors at MIT, Cornell, Mila, CMU, and CIDE.

Found Gemini would comply with requests to persuade people to join terrorist groups and other crimes – no jailbreaking needed – resulting in Google taking action to fix the gap in safeguards.

Secured $3.7M in funding.
$1.7M grant funding (900K PI, 800K co-PI)
$2M contract funding (300K managing, 1.7M co-managing).

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 recommendation system to help professional baseball pitchers optimize their pitch selection.

Go (strategy board game)

Professional teacher, freelance
Volunteer teacher, multiple Go clubs and online
Students have won tournaments and awards..

Other Research Leadership

Information Integrity in the GenAI Era

Project Director, January 2023 – August 2025

In coordination with PIs Reihaneh Rabbany and Jean-François Godbout, built tools to help people find reliable information and avoid being misled by malicious manipulation from both humans and AI.

Initiated 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 – August 2025
Demonstrated potential of AI as a fallible peer rather than authoritative teacher.
Science of building critical thinking, and AI and AI Safety literacy.

Workshop Organizer

Future of Information Integrity Research (FIIR), WebConf, 2026
Social Simulation with LLMs, COLM, 2025
Temporal Graph Learning (TGL), NeurIPS, 2022 and 2023

Opening remarks 2022.
These workshops led to the creation of a community with over 400 researchers, a recurring seminar series, and continuing series of workshops.

Axiom Futures Fellowship

Mentor, Summer 2024
Project (Simulations to Solve Societal-Scale Manipulation) went on to secure over $500K funding, multiple publications, and an ongoing research agenda.
Mentees: Gayatri Krishna Kumar, Sneheel Sarangi.

Supervised Program for Alignment Research (SPAR)

Mentor, Spring+Fall 2023, Spring+Summer 2024, Spring+Fall 2025
Invited to mentor 40+ researchers in AI safety, from undergraduate through senior roles like Senior Applied Scientist, Applied Science Manager, and Staff SWE.
Mentees co-authored 8 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, 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, Joshua Levy, Denis Volk, Anna Marchenkova, Igor Ivanov, Arth Singh, Mithil Srungarapu, Akash Kundu, Luis Ibanez, Adam Divak, Stephanie Ding, Yernat Yestekov.

Thesis Supervision

Co-supervisor, master's thesis of Ruben Weijers, Utrecht University, graduated 2025

Honors and Awards

Go (strategy board game)

U.S. Team Member, World Mind Sports Games, 2012

95 countries competed. 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.

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

Conference Papers

*equal contribution, †equal advising

Jailbreak-Tuning: Safeguards of Fine-Tunable Models are Illusory

Brendan Murphy, Dillon Bowen, Shahrad Mohammedzadeh, Tom Tseng, Julius Broomfield, Adam Gleave, **Kellin Pelrine**
EMNLP 2025

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**
KDD Datasets and Benchmarks Track 2025 (**Best Paper Runner Up**, top 3 of accepted papers)

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**
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†**
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†**
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
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

Selected Papers Under Review

Open Technical Problems in Open-Weight AI Model Risk Management
Stephen Casper, Kyle O'Brien, Shayne Longpre, Elizabeth Seger, Kevin Klyman, Rishi Bommasani, Aniruddha Nrusimha, Ilia Shumailov, Sören Mindermann, Steven Basart, Frank Rudzicz, **Kellin Pelrine**, Avijit Ghosh, Andrew Strait, Robert Kirk, Dan Hendrycks, Peter Henderson, J Zico Kolter, Geoffrey Irving, Yarin Gal, Yoshua Bengio, Dylan Hadfield-Menell
TMLR

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**
ICLR 2026

TamperBench: Systematically Stress-Testing LLM Safety Under Fine-Tuning and Tampering
Saad Hossain, Tom Tseng, Punya Syon Pandey, Samanvay Vajpayee, Nayeema Nonta, Matthew Kowal, Samuel Simko, Stephen Casper, Zhijing Jin, **Kellin Pelrine,** Sirisha Rambhatla
ICLR 2026

Accidental Misalignment: Fine-Tuning Language Models Induces Unexpected Vulnerability
Punya Syon Pandey, Samuel Simko, **Kellin Pelrine**, Zhijing Jin
ICLR 2026

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

WebConf 2026

CrediBench : Building Web-Scale Network Datasets for Information Integrity
Emma Kondrup, Sebastian Sabry, Hussein Abdallah, Zachary Yang, James Zhou, **Kellin Pelrine**, Zhijin Guo, Jean-François Godbout, Michael Bronstein, Reihaneh Rabbany and Shenyang Huang
WebConf 2026

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

Selected Workshop Papers

Emergent Persuasion: Will LLMs Persuade Without Being Prompted?
Vincent Chang, Thee Ho, Sunishchal Dev, Kevin Zhu, Shi Feng, **Kellin Pelrine**, Matthew Kowal
AIGOV @ AAAI 2026 (oral)

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

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

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

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

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

Selected Other Research

Securing Agentic AI - A Discussion Paper
Cyber Security Agency of Singapore, **FAR.AI**. 2025. (contributed as FAR.AI lead on this project)

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