ANDREI MURESANU

andrei.muresanu@uwaterloo.ca | +1 647-213-2129 | Google Scholar | LinkedIn | GitHub | AndreiMuresanu.com

RESEARCH & ENGINEERING HIGHLIGHTS

- Google: Created a new P0 (maximum severity) Google Search attack (5B users); proved no defence without core search redesign
- Vector Institute: First-author ICLR'23 (2k+ cites, WS best paper); built Vector's first multi-node 10B+ param LLM finetuning stack
- Nvidia: Cut data cleaning 50d→17h with a new ID algorithm; found 2.3x speed-up in popular DeepFace library; used in 2.2M cars
- WealthyPlanet: Led 12-engineer team; built personal-finance optimizer saving users $100k by retirement; valuation $3M→$20M

AI SAFETY/SECURITY EXPERIENCE

PhD Student Researcher, Google, New York, NY **April 2026 – Present**
- Created a **new attack** for Google Search with an internal **P0 (maximum severity) vulnerability** classification affecting **5B users**
- Mathematically proved that **defence is impossible** without a core paradigm shift in search mechanics. Targeting ICLR 2026

Graduate Researcher, with Profs. Nicolas Papernot and Zhijing Jin, University of Toronto, Toronto, ON **September 2025 – Present**
- Leading interpretability research to understand how AI generalization mechanisms predict safety properties in long-tail settings

Research Scientist, with Prof. Gillian Hadfield, Vector Institute, Toronto, ON **January 2024 – December 2024**
- 2 papers (NeurIPS, ICLR) in a largely **unexplored** field pioneering AI safety via **normative reasoning** for human-AGI collaboration
- **Designed core** normative module and ran all experiments, achieving a 30% boost in group norm identification and compliance
- Built a fully expressive text environment creation library with multi-agent support, reducing setup time from **5 days to 4 hours**

Research Scientist, with Prof. Nicolas Papernot, Vector Institute, Toronto, ON **May 2023 – February 2024**
- **First author** of ICML 2025 paper, proposing prompting for **exact unlearning** and a new holistic unlearning cost metric
- Designed and ran all experiments, achieving a **99.99% reduction** in exact unlearning cost vs state-of-the-art in language settings

Research Scientist, with Prof. Jimmy Ba, Vector Institute, Toronto, ON **August 2022 – April 2023**
- **First author** of ICLR paper **(2000+ citations, best paper, oral)**, proposed AI safety use and 70% efficiency boost via binary search
- Set up **Vector's first** multi-node framework to fine-tune LLMs with 10+ billion parameters. Edited LLM inference code using Jax

PROFESSIONAL EXPERIENCE

AI Team Lead & CTO, WealthyPlanet, Toronto, ON **April 2023 – Present**
- Led **12 engineers** on Canada's top finance optimizer, saving users ~$100k by retirement; raised valuation from **$3M to $20M**
- Led R&D from **concept to beta** with 100 customers and managed a 40k-line codebase serving up to 80k users/month

Principal Investigator, Silera.ai, Montreal, QC **January 2025 – January 2026**
- Led research and a **4-person** team to create a new synthetic data process, reducing data costs by 95% with no performance loss

Principal Investigator, Triomics, San Francisco, CA **January 2024 – May 2024**
- **Managed a team of 5** to automate cancer-trial eligibility, saving 10h per patient. Featured in **The Globe and Mail** (6M readers)

Research Scientist, with Prof. Animesh Garg, Vector Institute, Toronto, ON **May 2023 – December 2023**
- **Built the fastest** 3D memory benchmark (43% faster) and the **most extensive** memory test, supporting 3 modalities and 9 tasks
- **Conceptualized core idea**/algorithm for **3** projects: formal memory definition, hyperbolic geometry for memory, and text-to-sim

Computer Vision Research Engineer, NVIDIA, Santa Clara, CA **January 2022 – April 2022**
- Cut cleaning time of 0.5 billion images from **50 days to 17 hours** by developing a new stochastic dominant identity algorithm
- Attained a **2.3x speed-up** by redesigning the face-matching module in Meta AI's DeepFace, a library used by 50,000+ developers

Research Scientist, under Prof. Chul Min Yeum, University of Waterloo, Waterloo, ON **September 2021 – December 2021**
- Initiated creation of the world's first autonomous flood-risk system and proposed a novel synthetic data pipeline to enable it

Machine Learning Engineer, Advanced AI & Analytics Research Team, PerkinElmer, Waterloo, ON **May 2021 – August 2021**
- Pioneered a deep learning method for state-of-the-art mass spectrometry; uncovered overfitting causing poor generalization

Machine Learning Engineer, Geminare, Toronto, ON **May 2020 – December 2020**
- Conceptualized an original object-detection process capable of using limited and mislabeled data, saving $15,000 over 3 months

PUBLICATIONS

Large Language Models Are Human-Level Prompt Engineers

Yongchao Zhou*, **Andrei Muresanu***, Ziwen Han*, Keiran Paster, Silviu Pitis, Harris Chan, Jimmy Ba
International Conference on Learning Representations (ICLR). 2023
Best Paper Award at NeurIPS 2022 ML Safety Workshop
Oral presentation at NeurIPS 2022 Foundation Models for Decision Making Workshop
2221 citations as of July 9, 2026

Benchmarks for Physical Reasoning AI

Andrew Melnik, Robin Schiewer, Moritz Lange, **Andrei Muresanu**, Mozhgan Saeidi, Animesh Garg, Helge Ritter
Transactions on Machine Learning Research (TMLR). 2023
Awarded the **Exceptional Survey Certificate**

Unlearnable Algorithms for In-context Learning

Andrei Muresanu, Anvith Thudi, Michael Zhang, Nicolas Papernot
International Conference on Machine Learning (ICML). 2025

Normative Modules: A Generative Agent Architecture for Learning Norms that Supports Multi-Agent Cooperation

Atrisha Sarkar, **Andrei Muresanu**, Carter Blair, Aaryam Sharma, Rakshit S Trivedi, Gillian K Hadfield
Submitted to the Conference on Neural Information Processing Systems (NeurIPS). 2024
Accepted to the workshop on Foundation Models and Game Theory (FMGT). 2024

Altared Environments: The Role of Normative Infrastructure in AI Alignment

Rakshit Trivedi, Nikhil Chandak, **Andrei Muresanu**, Shuhui Zhu, Atrisha Sarkar, Joel Leibo, Dylan Hadfield-Menell, Gillian Hadfield
Submitted to the International Conference on Learning Representations (ICLR). 2025

CauSciBench: Can LLMs Automate Causal Inference in Real-World Scientific Research?

Sawal Acharya, Terry Jingchen Zhang, Andrew Kim, Anahita Haghighat, Xianlin Sun, Pepijn Cobben, Rahul Babu Shrestha, Maximilian Mordig, Jacob T. Emmerson, Furkan Danisman, Yuen Chen, Clijo Jose, **Andrei Muresanu**, Justin Cui, Jiarui Liu, Yahang Qi, Punya Syon Pandey, Yinya Huang, Bernhard Schölkopf, Zhijing Jin
International Conference on Machine Learning (ICML). 2026

EDUCATION

MSc + Ph.D. in Computer Science — **September 2025 – April 2030**
University of Toronto — Toronto, Ontario, Canada
Advisors: Prof. Nicolas Papernot and Prof. Zhijing Jin

Bachelor of Statistics — **April 2024 – August 2025**
University of Waterloo — Waterloo, Ontario, Canada

Bachelor of Computer Science — **September 2019 – April 2024**
University of Waterloo — Waterloo, Ontario, Canada
Recipient of Research Certificate

SELECTED PROJECTS

Superhuman Poker AI — **March 2021 – March 2022**

- Recreated Facebook AI's 2019 "Pluribus" project **from scratch** and corrected 5+ errors in one of the supporting papers

Unity Neural Network Library — **February 2019 – April 2019**

- Constructed the **first-ever** Unity neural network library from scratch, used to create backpropagation neural networks

Indie Game Developer — **September 2013 – June 2019**

- Built **25+ games over 6 years**. Released on desktop, mobile, and in the browser. Primarily developed with Unity in C#

AWARDS, FELLOWSHIPS, & GRANTS

- (2026) ($39,282) Google DeepMind AI Masters Scholarship
- (2025) ($80,000) **As PI** for an **Open Philanthropy** grant on AI Safety
- (2025) ($17,500) Vector Master's Scholarship (**1 of 2** recipients for the Computer Science category at the University of Toronto)

- (2024) ($8,000) Vector Institute Research Grant
- (2024) ($33,000) Vector Institute Research Grant
- (2023) ($8,000) Georgia Tech Research Grant
- (2023) ($8,000) University of Toronto Research Grant
- (2022) ($7,500) Vector Institute Research Grant
- (2021) ($1,000) University of Waterloo Undergraduate Research (URA) Grant
- (2019) ($2,000) University of Waterloo President's Entrance Scholarship

Competition Awards:

- (2019) Top 20 finalist (Top 0.00045%) in C1 Terminal International AI Programming Competition
- (2019) 2nd Place in Toronto Police Hackathon. Presented our idea to the mayor in a televised board meeting
- (2019) Won 1st place and $5,000 in the DMZ Basecamp pitch competition as co-founder of a non-invasive insulin patch startup

INVITED TALKS

- **What Is Memory and How Ought We Store It?**: Amii Upper Bound 2026
  1 of 6 *graduate students selected for the* ***Pan-Canadian*** *research pitch competition.*
- **What Is Memory and How Ought We Store It?**: Vector Institute, Remarkable Conference 2026
- **What does it mean to trust an AI system?**: Steven's Institute of Technology 2026
- **What does it mean to trust an AI system?**: University of Toronto, Conference X 2026
- **Fast exact unlearning for in-context learning data for LLMs**: University of Waterloo 2025
- **Large Language Models Are Human-Level Prompt Engineers**: Vector Institute 2024

SERVICES

- Reviewer for the Neural Information Processing Systems (NeurIPS) international conference 2026
- Reviewer for the Association for Computational Linguistics (ACL) international conference 2025
- NLP and LLM Workshop Lead at MacHacks | McMaster University 2023
- Computer Vision Workshop Lead at MacHacks | McMaster University 2022

ADDITIONAL INFORMATION

- **Coding Languages:** Proficient: **C**, **C#**, **C++**, **Python**, **MIPS**, **ARM,** and **Scheme/Lisp**; working knowledge: **SQL**, **R**, **MATLAB**, and **Java**
- **Languages**: Fluent in English and Romanian; professional working proficiency in French
- **Skills:** Git, Docker, NumPy, SciPy, Pandas, OpenCV, CUDA, Scikit-Learn, CNN, data mining, data visualization, computer vision, web scraping, big data, data analytics, deep learning, GPU, parallel programming, simulation, reinforcement learning, PyTorch, TensorFlow, algorithms, GCP, Azure, AWS, alignment, mechanistic interpretability, scalable oversight, robustness, jailbreaks