CAISI Research Program at CIFAR

2025 Year in Review: Building Safe AI for Canadians

CIFAR

THE NEXT LEAP STARTS **HERE**

Table of Contents

CIFAR's leadership of the Canadian AI Safety Institute Research Program is funded by the Government of Canada, and delivered in collaboration with Amii, Mila and the Vector Institute.

Funded by the Government of Canada | Canada

Introduction

In November 2024, the Government of Canada launched the Canadian AI Safety Institute (CAISI), recognizing that public trust is the primary driver of successful innovation. While CAISI functions as a federal initiative under the Government of Canada's Department of Innovation, Science and Economic Development Canada (ISED), the CAISI Research Program at CIFAR serves as its independent scientific engine. We are charged with mobilizing the nation's AI safety experts across disciplines to address the complex technical and social challenges of advanced AI systems.

The Year in Review: Building Safe AI for Canadians summarizes our progress in 2025. By building Canadian research capacity and fostering a critical mass of skilled talent, CIFAR is positioning Canada as a global leader in developing safe and trustworthy AI systems.

CAISI Research Program at CIFAR by the Numbers

In its first year, the CAISI Research Program at CIFAR delivered significant impacts in AI safety research, training and mobilizing knowledge and insights by:

Building a National AI Safety Research Community

Funding a network of 55+ experts across disciplines, including:

28

Principal Investigators and 6 CAISI Research Council members

11

CIFAR AI Safety Postdoctoral Fellows

10

AI Safety Scientists & Engineers at Amii, Mila and the Vector Institute

Driving High-Impact Research

$2.4M invested to launch 12 new research projects, including:

10

Catalyst projects

2

Solution Networks

Delivering Results

5

Active national and international research partnerships

3

AI safety expert-policymaker roundtables

28+

Knowledge products/ research outputs expected (2026)

Message From the Co-Directors

The year 2025 marked a global turning point for AI safety. With urgent concerns voiced by leading experts, including Canada CIFAR AI Chair Yoshua Bengio and CIFAR Distinguished Fellow Geoffrey Hinton, and the release of the International AI Safety Report, the world recognized the imperative to balance rapid innovation with rigorous risk mitigation.

Catherine Régis
Co-Director, CAISI Research Program at CIFAR // Canada CIFAR AI Chair, Mila, Université de Montréal

Nicolas Papernot
Co-Director, CAISI Research Program at CIFAR // Canada CIFAR AI Chair, Vector Institute, University of Toronto

Canada was uniquely prepared to meet this challenge. Having been appointed to develop and implement the world's first national AI strategy in 2017, CIFAR has been instrumental in building the deep talent pool and research excellence that defines our nation's success today. It is upon this proven foundation that the CAISI Research Program at CIFAR was established. Through this initiative, we are building Canadian research capacity and fostering a critical mass of skilled talent in AI safety, positioning Canada as a global leader in developing safe and trustworthy AI systems. We have leveraged this legacy to rapidly mobilize Canada's scientific community, moving from concept to concrete impact in just one year.

As Co-Directors from the distinct fields of law and computer science, we view AI safety as a sociotechnical challenge. Our program bridges the gap between technical and social considerations, ensuring AI systems are both robust and socially responsible. Enabling multidisciplinary collaborations among the research community, government, and other partners is key to addressing the true complexity of AI safety challenges.

We are already delivering critical results

Defending Democracy

Developing systems to combat foreign influence and disinformation.

Protecting Youth

Creating guardrails to detect and block content that encourages self-harm.

Securing Justice

Launching a Solution Network to safeguard Canadian courts from synthetic AI evidence.

(L-R) Elissa Strome (CIFAR), Joel Martin (National Research Council Canada), Stephen Toope (CIFAR), Yoshua Bengio (Mila, LawZero), François-Philippe Champagne (Government of Canada), Valérie Pisano (Mila), Tony Gaffney (Vector Institute) and Cam Linke (Amii) at the announcement of the Canadian AI Safety Institute in November 2024.

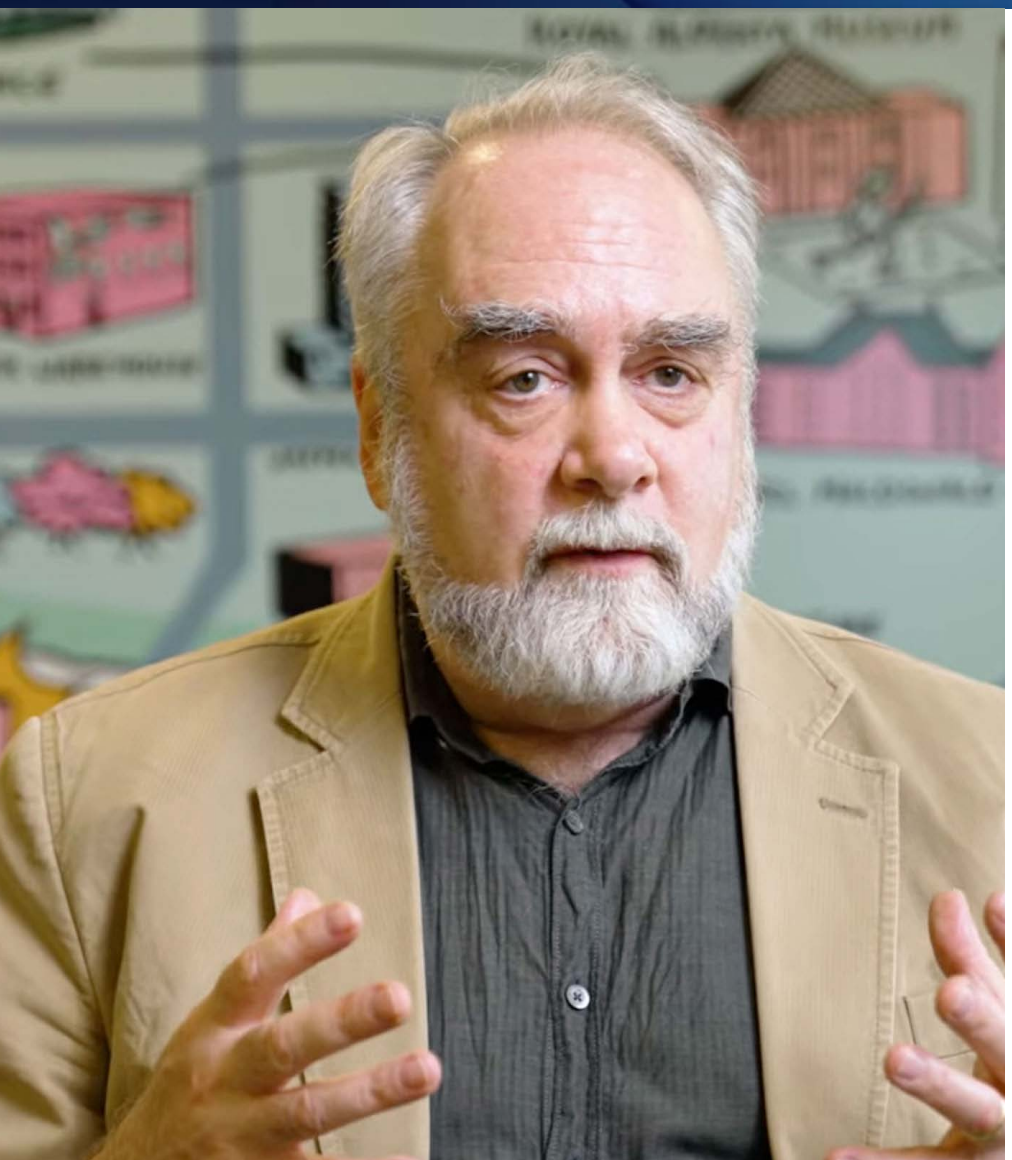
In our first year, we funded 12 projects and supported 28 researchers across disciplines. Our program is advancing innovative research that is developing new techniques and practices to ensure AI systems are safe and reliable, and their deployment takes into account our societal values. In a nutshell, ensuring that AI systems can be trusted.

All of this wouldn't be possible without the dedication and commitment of Canada's AI safety research community, the Government of Canada and ISED for their ongoing investment in AI safety research and talent, and our many supporters, partners and collaborators, including the National Research Council of Canada, the national AI institutes (Amii, Mila and Vector Institute), the International Development Research Centre (IDRC), the UK AI Security Institute and many others.

Together, we are building safer AI for Canadians.

Safeguarding Society

Protecting our collective future from the large-scale risks of advanced AI. This means confronting and mitigating systemic harms, like mass disinformation and economic disruption, and building the tools and policies needed to ensure AI remains a force for public good.



Intelligent Ideas with Geoffrey Rockwell

Canada CIFAR AI Chair at Amii, Geoffrey Rockwell, uses his ethics expertise to bring a philosophical perspective to AI safety research, discussing the role of government in mitigating harm and applying existing safety knowledge and infrastructure to AI deployment.

See video

Spotlight

Safeguarding Mental Health from AI Companions

As more Canadians turn to AI chatbots for companionship and self-validation, there is growing proof that misuse and overuse of AI companion chatbots cause mental health harm, ranging from dependency to full AI psychosis and suicide assistance. As many as 70% of young people now regularly turn to AI companions, necessitating the need for independent safeguards, including technological guardrails, policies, and education.

To mitigate the risks of harmful chatbot interactions, the CAISI Research Program at CIFAR provided funding to support Mila's AI Safety Studio to undertake this work. This initiative focuses on creating independent, trustworthy AI guardrails and developing exhaustive benchmarks that reflect Canadian cultural and societal diversity to objectively measure the harm.

To date, the Studio has developed its first iteration of a mental health guardrail and benchmark for AI chatbots. They are now working to extend their reach across multiple large language model (LLM) vendors, languages and cultural specificities, using anonymized real-world data and input from mental health experts.

Cross-Disciplinary Collaboration and Future Focus

"The most exciting aspect of this work is the unanimous, cross-disciplinary support of a web of partners. The socio-technical collaboration across disciplines — bridging AI expertise, mental health, policy, education specialists and impacted communities grassroots up, ensures that we'll create a robust, multidisciplinary protection against companion AI mental health harm," said Simona Grandrabur, Mila's AI Safety Studio Lead.

"Mila's mental health guardrail and benchmarks will establish an independent and trustworthy means to measure the extent of harmful interactions with AI companions to safeguard our most vulnerable populations, including our children, against suicide assistance."

Simona Grandrabur
AI Safety Lead, Mila

Over the next year, the Studio plans to develop intelligent filters to block AI-generated content that assists or encourages self-harm or suicide, as well as reliability testing protocols to evaluate the safety and robustness of conversational and generative AI models. Additionally, the Studio will develop psychological and ethical risk assessment tools.

The first official version of the AI Safety Studio benchmark dashboard and guardrails will be released to the public in 2026.

Read the full story

(L-R) Kianna Adams (AltaML), Golnoosh Farnadi (Mila), Mohamed Abdalla (Amii) and Elissa Strome (CIFAR) discuss how to build AI systems that are aligned with the values and safety of society and how Canada can lead in this endeavor.

Spotlight

Securing Canada Against Disinformation

Malicious foreign influence and AI-driven disinformation pose a direct threat to Canadian democracy, aiming to erode trust in our institutions, media and civil society. In response, a 2025 CIFAR AI Safety Catalyst project is developing an advanced AI tool to protect Canadians against disinformation campaigns.

Defending against the malicious use of AI is the focus of this research, which is led by Canada CIFAR AI Chair Matthew E. Taylor (Amii, University of Alberta), Brian McQuinn (University of Regina), and CIFAR AI Safety Postdoctoral Fellow James Benoit.

An AI Defense System

The team is developing CIPHER, an advanced human-in-the-loop AI system. The core purpose of this tool is to empower civil society organizations by equipping them to identify and combat sophisticated and coordinated disinformation campaigns. The initial focus of this work is to detect Russian operations across both textual and visual media, providing a vital shield of protection for Canadian society.

“The CIPHER project treats safe and reliable information as a matter of national security. Identifying state-backed disinformation news campaigns can help us all remain rooted in Canadian facts and values. Our goal is to ensure outside influencers don’t poison our debates and security decisions,” the team told CIFAR.

This CIFAR AI Safety Catalyst project will deliver tangible impacts by producing:

- A rigorously evaluated proof-of-concept of the CIPHER tool, tested in the real world by Canadian and global civil society partners.
- Actionable policy briefs to guide government and industry response.
- A new public dataset to accelerate further research and development in this critical area, ensuring the project’s impact extends far beyond its initial scope.

"AI is becoming increasingly common. Rather than outsourcing important decisions to AI, our design makes sure humans are always in the loop. The CIPHER project aims to earn the trust of decision-makers and users to collaboratively defend democratic spaces from disinformation and misinformation."

Matthew E. Taylor
Canada CIFAR AI Chair, Amii

Funded Projects

Solution Network: Safeguarding Courts from Synthetic AI Content

- Ebrahim Bagheri (University of Toronto)
- Maura Grossman (University of Waterloo)

AI Safety Catalyst Project: On the Safe Use of Diffusion-based Foundation Models

- Mi Jung Park (Canada CIFAR AI Chair, Amii, University of British Columbia)

AI Safety Catalyst Project: CIPHER: Countering Influence Through Pattern Highlighting and Evolving Responses

- Matthew E. Taylor (Canada CIFAR AI Chair, Amii, University of Alberta)
- Brian McQuinn (University of Regina)
- James Benoit (CIFAR AI Safety Postdoctoral Fellow, Amii)

See details of all projects

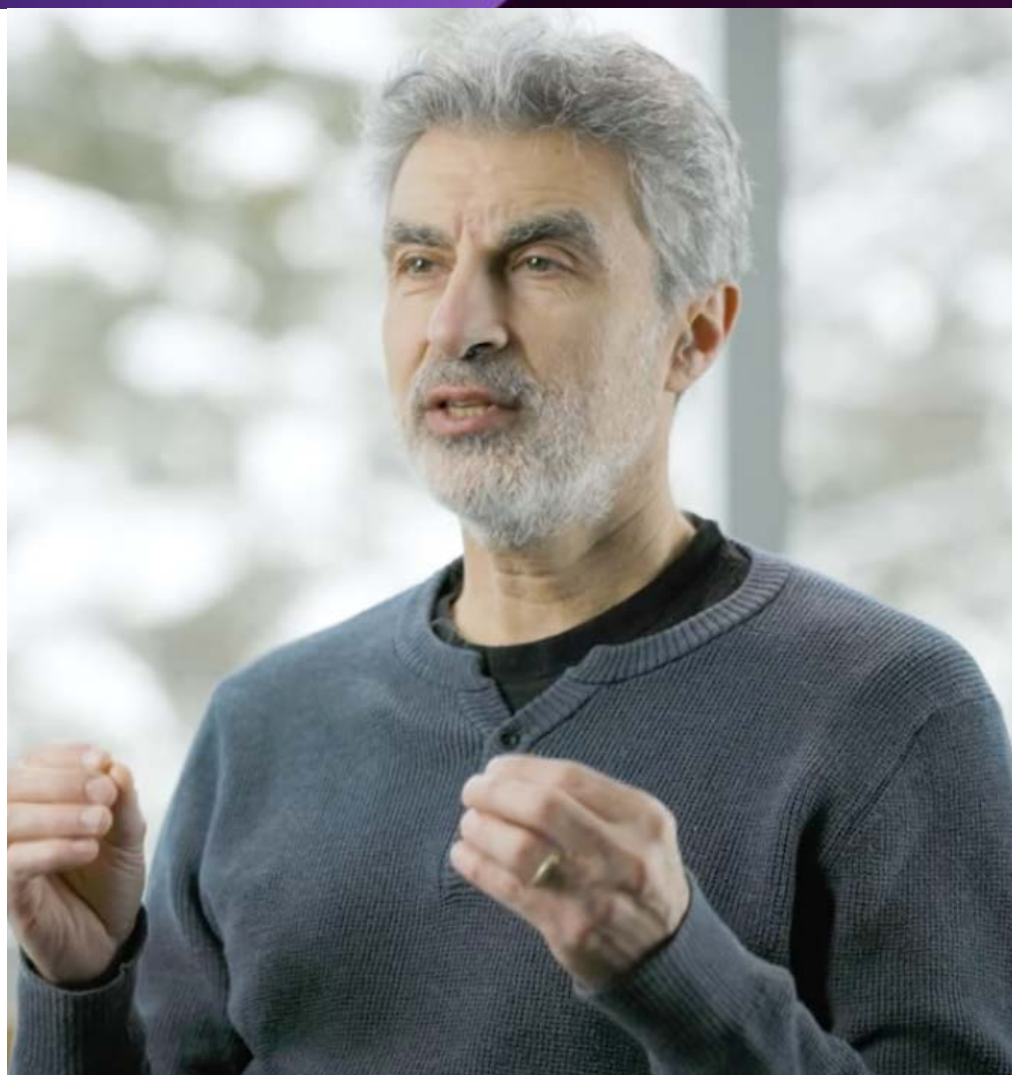
Building Trust & Fairness

Actively embedding human values and equity into AI systems ensures that AI is aligned with the public good and does not perpetuate existing societal harms such as bias and discrimination.

Intelligent Ideas with Yoshua Bengio

How do we design AIs that will not harm people? Canada CIFAR AI Chair Yoshua Bengio (Mila, University of Montreal, LawZero) explores the various ways that AI could harm society and why he remains optimistic about the future.

See video



CAISI Research Program at CIFAR

Spotlight

Building Safe AI Through More Reliable Reasoning

How can we trust an AI that knows more than we do? AI systems draw on immense data, leaving users unable to verify their outputs. While AI debate was proposed to solve this by forcing models to cite evidence, current systems fail to argue reliably — misinterpreting facts and taking sources out of context. To earn trust, an AI must learn to construct and support justifications that can be taken at face value.

"The highly interdisciplinary nature of this collaboration helps reframe the researchers' thinking about AI as not just a technical matter, but also a social, political and economic one. Learning from experts in other fields as well as your own is crucial for making meaningful progress in today's AI research."

Maria Ryskina
CIFAR AI Safety Postdoctoral Fellow

One research team aims to address this issue by developing an AI Debate Framework, a goal pursued by Gillian Hadfield, (Johns Hopkins University, University of Toronto (status only), Vector Institute) and her team, including Maria Ryskina, in their 2025 AI Safety Catalyst project funded by CIFAR.

“Our goal is to teach language models to construct reasonable justifications and to support them with valid evidence from available sources of information,” says Maria Ryskina, a CIFAR AI Safety Postdoctoral Fellow at Vector Institute. “Such models will be more worthy of users’ trust as they can be reliably overseen by non-experts.”

Trust Between Humans and AI Systems

Drawing on insights from law, economics, cultural evolution, and political science, the AI Debate Framework project aims to equip AI with normative reasoning — making choices informed by the rules that coordinate behaviour, mirroring how people’s actions are informed by their societies’ norms and laws. The project aims to make AI systems more trustworthy, resilient and better aligned with human social structures.

The team plans to deliver several key technical assets: novel AI agent architectures, a new dataset of disciplined normative reasoning, and a debate-based framework to test and enforce stable, shared norms in AI agent interactions.

The end goal is to make AI systems safer by raising their awareness of the unwritten rules and complex trade-offs of our diverse human society.

“The unconstrained growth of artificial intelligence has already started to have major ripple effects on people’s lives. I am proud to be part of an effort to steer the trajectory of AI development and usage towards a more positive future,” added Ryskina.

Funded Projects

Solution Network: Mitigating Dialect Bias

- Laleh Seyyed-Kalantari (York University)
- Blessing Ogbuokiri (Brock University)

AI Safety Catalyst Project: Sampling Latent Explanations From LLMs for Safe and Interpretable Reasoning

- Yoshua Bengio (Canada CIFAR AI Chair, Mila, Université de Montréal, LawZero)

AI Safety Catalyst Project: Adversarial Robustness of LLM Safety

- Gauthier Gidel (Canada CIFAR AI Chair, Mila, Université de Montréal)

AI Safety Catalyst Project: Advancing AI Alignment Through Debate and Shared Normative Reasoning

- Gillian Hadfield (Johns Hopkins University, University of Toronto)
- Maria Ryskina (CIFAR AI Safety Postdoctoral Fellow, University of Toronto)

AI Safety Catalyst Project: Formalizing Constraints For Assessing and Mitigating Agentic Risk

- Sheila McIlraith (Canada CIFAR AI Chair, Vector Institute, University of Toronto)

See details of all projects

Renowned computer scientist Deborah Raji (Mozilla) speaks about the sociotechnical dimensions of AI safety at the inaugural CAISI Research Program Annual Meeting in October 2025.

Securing Critical Systems

Developing rigorous tools to evaluate the safety, accuracy, and reliability of frontier AI is essential for securing critical systems and enabling responsible industry innovation.

Intelligent Ideas with Nicolas Papernot

How do we build safe, trustworthy AI systems? Nicolas Papernot's research explores how machines can learn important information without compromising personal user data. Securing critical systems is essential for building trust among Canadians and keeping their information safe and secure while accelerating the adoption of AI.

See video

CAISI Research Program at CIFAR

Spotlight

Landmark Evaluation Study Measures Safety and Reliability of Frontier AI

In a revolutionary benchmark study, the Vector Institute conducted an independent evaluation to measure the safety and reliability of the world's LLMs.

Led by Vector's AI Engineering team, the project assessed 11 prominent frontier AI models from around the world. The evaluation examined both open- and closed-source systems, including the January 2025 release of DeepSeek-R1, testing each one against a comprehensive suite of 16 different benchmarks.

This project marks a critical milestone in AI safety research. As the global AI race intensifies, with the development of increasingly powerful LLMs, developing trusted and widely accepted benchmarks is essential. This research provides a crucial tool for helping researchers, developers and policymakers understand how these models perform in terms of accuracy, reliability and fairness.

Enabling Safe and Reliable AI Adoption

"This study gives people and organizations a clear, independent picture of how these models actually behave so that AI can be adopted safely and with confidence," says Deval Pandya, VP of AI Engineering, Vector Institute.

To promote transparency and accountability, the Vector Institute has open-sourced the entire study, including its results and underlying code. "By releasing our work openly, we are giving everyone the ability to verify, learn and build on it, which supports smarter and safer use of AI across the country," he added.

"I am excited that we are helping create a future where AI earns trust because people can see how it works and test it for themselves. By open-sourcing our evaluations and tools, we are enabling a broader community to replicate the results, spot gaps and improve on them. This is how we support safe adoption, not just in labs but in hospitals, classrooms, businesses and public services."

Deval Pandya
VP AI Engineering, Vector Institute

The evaluation features powerful benchmarks like MMLU-Pro, MMMU, and OS-World, which are now widely used in the field. These specific benchmarks, developed by University of Waterloo professors and Canada CIFAR AI Chairs at the Vector Institute, Wenhu Chen and Victor Zhong, whose research is improving how we approach benchmark techniques, are now widely adopted by major companies, including OpenAI, Google, and Anthropic.

Read the full story

Funded Projects

AI Safety Catalyst Project: Safe Autonomous Chemistry Labs

- Alán Aspuru-Guzik (Canada CIFAR AI Chair, Vector Institute, University of Toronto)

AI Safety Catalyst Project: Adversarial Robustness in Knowledge Graphs

- Ebrahim Bagheri (University of Toronto)
- Jian Tang (Canada CIFAR AI Chair, Mila, HEC Montréal & Université de Montréal)
- Benjamin Fung (Mila, McGill University)

AI Safety Catalyst Project: Maintaining Meaningful Control: Navigating Agency and Oversight in AI-Assisted Coding

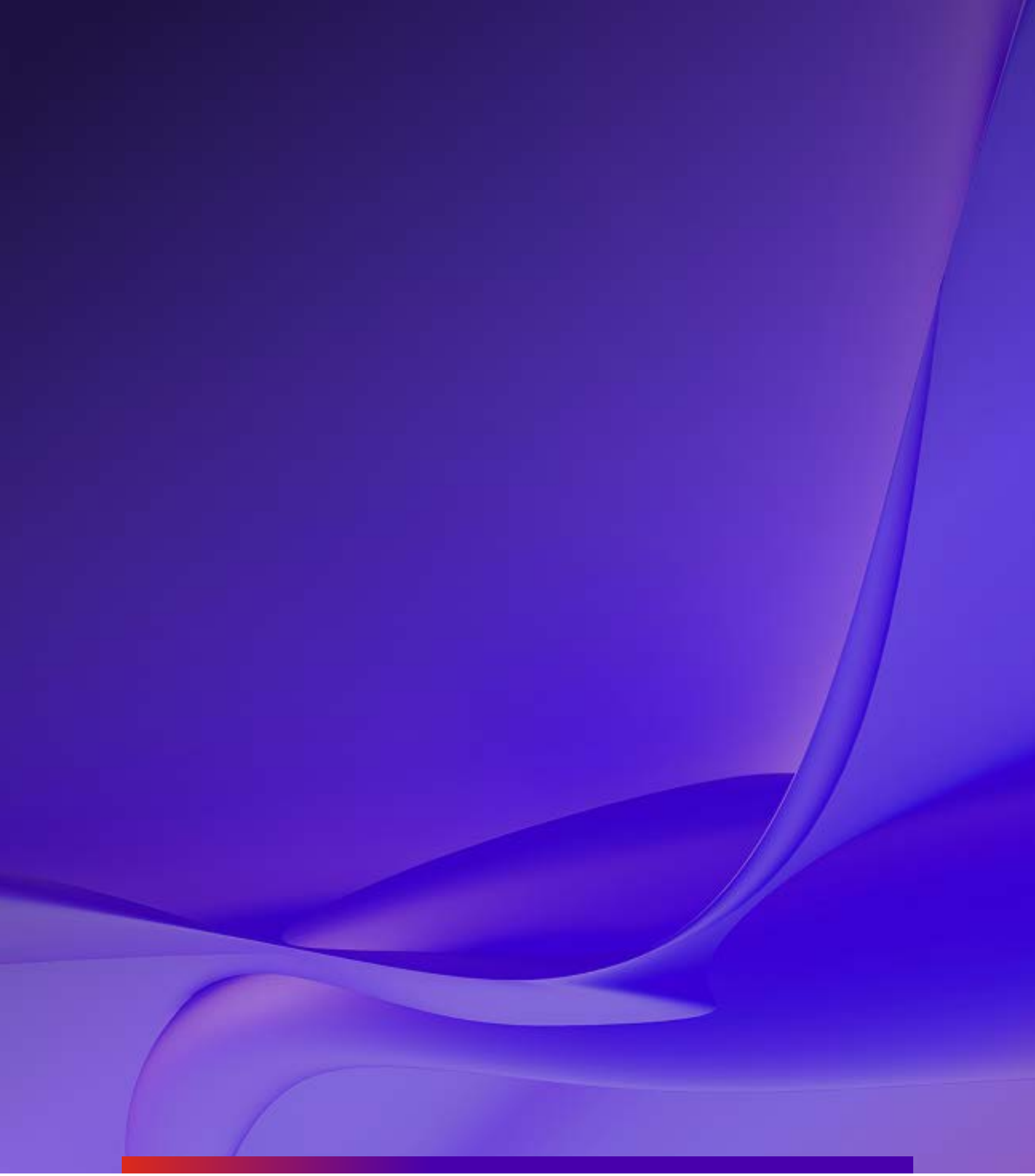
- Jackie Chi Kit Cheung (Canada CIFAR AI Chair, Mila, McGill University)
- Jin Guo (McGill University)

AI Safety Catalyst Project: Safety Assurance and Engineering for Multimodal Foundation Model-enabled AI Systems

- Foutse Khomh (Canada CIFAR AI Chair, Mila, Polytechnique Montréal)
- Lei Ma (Canada CIFAR AI Chair, Amii, University of Alberta)
- Randy Goebel (Amii, University of Alberta)

See details of all projects

Blessing Ogbuokiri (Brock University), Co-director of the Mitigating Dialect Bias Solution Network attends the inaugural CAISI Research Program Annual Meeting in October 2025.



CIFAR

THE NEXT LEAP STARTS **HERE**

CIFAR.CA