VECTOR INSTITUTE

Search Search...

This is a navigation window which overlays the main content of the page. Pressing the "X" at the top right corner of the modal will close the modal and bring you back to where you were on the page.

See posts about

- Announcement
- Blog
- Case Study
- Generative AI
- Health
- Insights
- Intellectual Property
- Masters Programs
- News
- Partnership
- Press

This is a call to action window which overlays the main content of the page. Pressing the "X" at the top right corner of the modal will close the modal and bring you back to where you were on the page.

Let's Collaborate

Partnership Sector

First Name

Last Name

Company

Company Email

City

Country

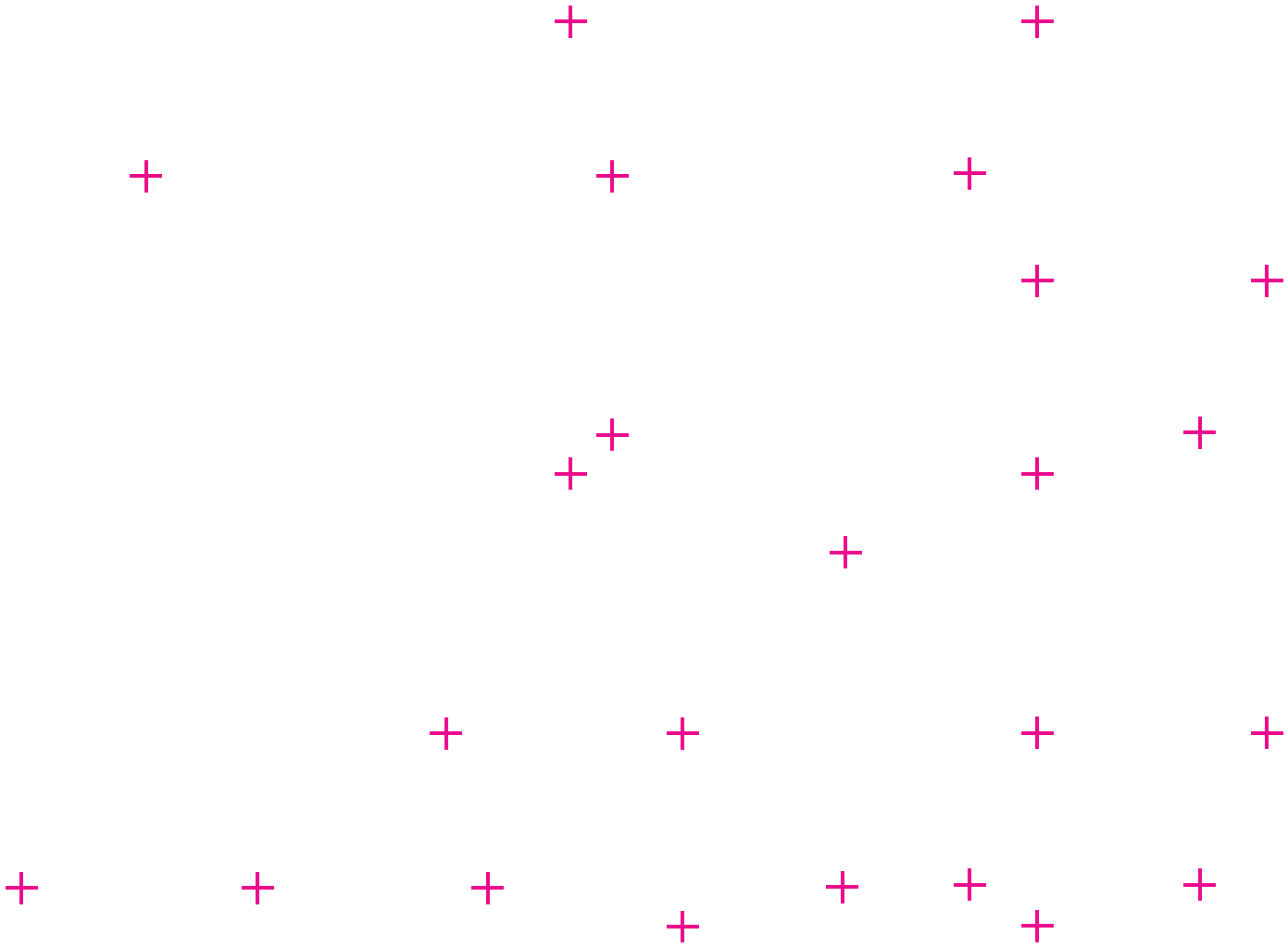
Please select the country you are located in

Tell us about your organization

☐ Yes, contact me

By submitting this form you consent to the given information be used by Vector to contact you about my inquiry. You have read and agreed to the Privacy Policy

Send Message



This is a search window which overlays the main content of the page. Pressing the "X" at the top right corner of the modal will close the modal and bring you back to where you were on the page.

Search What are you looking for?

The Vector Institute's AI Trust and Safety Principles

June 14, 2023

Trustworthy AI

From the inception of the Vector Institute, its community has shared a strong commitment to developing AI that is both safe and trustworthy. It is at the core of our existence. Now, powerful and easily accessible generative AI models have revealed to the general public the short-term and existential risks of AI. As the pace of AI development quickens, policy makers and AI experts around the world are seeking to develop shared guardrails to help guide future development of this technology.

As a central hub of Canada's AI ecosystem, Vector has distilled six AI trust and safety principles for Canadian and international organizations that reflect the ongoing dialogue, shared values, and best practices of the businesses, governments, and globally-renowned research communities with whom we work.

Following these principles from research through to implementation and/or commercialization will help to ensure that AI systems are safe and trustworthy. Vector will be updating its own Code of Conduct to ensure that these principles continue to shape the institute's work.

Vector believes these trust and safety principles can provide guidance for other organizations developing their own code of conduct and AI policies. They are a starting point for organizations at the beginning of their AI journey, while more established organizations can build on them to best suit their needs.

By embracing these principles, organizations developing their own AI policies around the world will signal their commitment to harnessing the power of AI as a means to democratic ends.

A strong commitment to these principles across Canada's AI ecosystem would extend the country's global leadership in the responsible development and deployment of AI.

First principles for AI

The Vector Institute's unique ability to convene stakeholders across AI domains, institutions, and industry sectors gives the institute a novel perspective on the trends and challenges of implementing AI.

These principles are built on international themes gathered from across multiple sectors to reflect the values of AI practitioners in Vector's ecosystem, across Canada, and around the world:

1. AI should benefit humans and the planet.

We are committed to developing AI that drives inclusive growth, sustainable development, and the well-being of society. The responsible development and deployment of AI systems must consider equitable access to them along with their impact on the workforce, education, market competition, environment, and other spheres of society. This commitment entails an explicit refusal to develop

harmful AI such as lethal autonomous weapons systems and manipulative methods to drive engagement, including political coercion.

2. AI systems should be designed to reflect democratic values.

We are committed to building appropriate safeguards into AI systems to ensure they uphold human rights, the rule of law, equity, diversity, and inclusion, and contribute to a fair and just society. AI systems should comply with laws and regulations and align with multi-jurisdictional requirements that support international interoperability for AI systems.

3. AI systems must reflect the privacy and security interests of individuals.

We recognize the fundamental importance of privacy and security, and we are committed to ensuring that AI systems reflect these values appropriately for their intended uses.

4. AI systems should remain robust, secure, and safe throughout their life cycles.

We recognize that maintaining safe and trustworthy AI systems requires the continual assessment and management of their risks. This means implementing responsibility across the value chain throughout an AI system's lifecycle.

5. AI system oversight should include responsible disclosure.

We recognize that citizens and consumers must be able to understand AI-based outcomes and challenge them. This requires the responsible transparency and disclosure of information about AI systems – and support for AI literacy – for all stakeholders.

6. Organizations should be accountable.

We recognize that organizations should be accountable throughout the life cycles of AI systems they deploy or operate in accordance with these principles, and that government legislation and regulatory frameworks are necessary.

Methodology

The Vector Institute's First Principles for AI build upon the approach to ethical AI developed by the OECD. Along with trust and safety principles, clear definitions are also necessary for the responsible deployment of AI systems. As a starting point, the Vector Institute recognizes the Organization for Economic Co-operation and Development (OECD) definition of an AI system. As of May 2023, the OECD defines an AI system as follows:

"An AI system is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy."

Vector also acknowledges that widely-accepted definitions of AI systems may be revised over time. We have seen how the rapid development of AI models can change both expert insight and public opinion on the risks of AI. Through Vector's Managing AI Risk project we collaborated with many organizations and regulators to assess several types of AI risk. These discussions informed the language around risks and impact in the principles.

The dynamic nature of this challenge necessitates that companies and organizations should be prepared to revise their principles as they respond to the changing nature of AI technology.

1. According to a white paper from the Berkman Klein Center for Internet and Society at Harvard, the OECD's statement of AI principles is among the most balanced approaches to articulating ethical and rights-based principles for AI.
2. As AI technology develops, definitions are updated as needed. In May 2023, the OECD was in the process of revising their definition of AI systems.

Related:

Insights

Research

Trustworthy AI

Neutralizing Bias in AI: Vector Institute's UnBIAS Framework Revolutionizes Ethical Text Analysis

News

Trustworthy AI

Key takeaways from the All In 2023 conference

News

Research

Dan Roy named Vector Research Co-Director