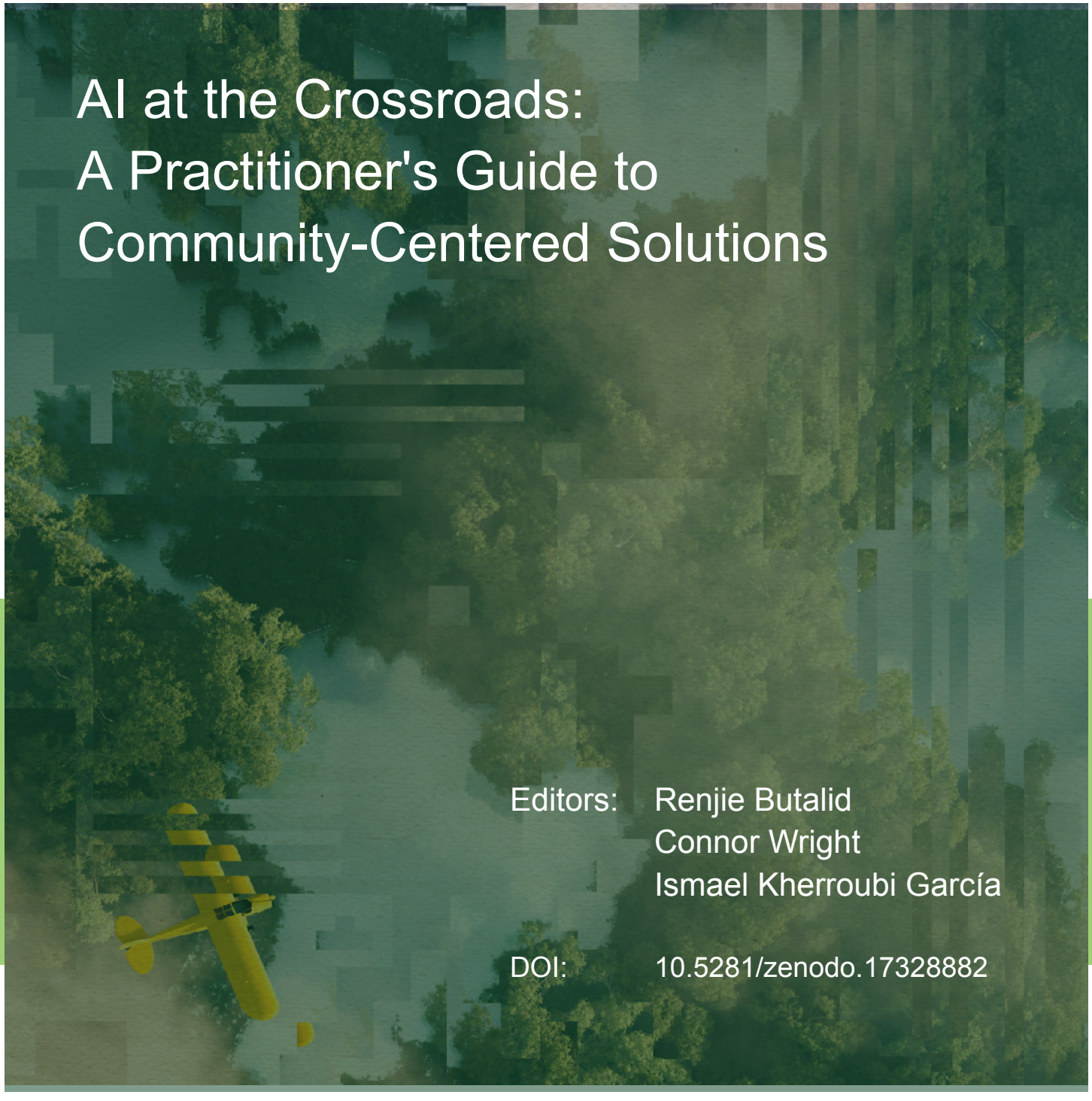
STATE OF AI ETHICS REPORT | VOLUME 7



AI at the Crossroads: A Practitioner's Guide to Community-Centered Solutions

Editors: Renjie Butalid
Connor Wright
Ismael Kherroubi García

DOI: 10.5281/zenodo.17328882

Front & Back Cover Image: Distorted Lake Trees by Lone Thomasky & Bits&Bäume / Better Images of AI / CC BY 4.0. ALT text: A bird's eye view photo of a small yellow aeroplane flying over a river or lake, interspersed with trees and clouds. However, the image is slightly distorted with digital artifacts.

DOI: 10.5281/zenodo.17328882

Dedicated to

Abhishek Gupta

Founder & Principal Researcher
Montreal AI Ethics Institute

In Memoriam (Dec 20, 1992 – Sep 30, 2024)

DOI: 10.5281/zenodo.17328882

Published by the Montreal AI Ethics Institute (MAIEI)

Montreal, Quebec, Canada
November 4, 2025
The State of AI Ethics Report (Volume 7)

DOI: 10.5281/zenodo.17328882

Suggested Citation

For citing the entire report:

Butalid, R., Wright, C. & Kherroubi García, I. (Eds.). (2025). The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide to Community-Centered Solutions. *Montreal AI Ethics Institute*. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

For citing individual pieces within the report:

[Author Surname], [Initials]. (2025). [Title of piece]. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. [xx-xx]. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Disclaimer:

The views and opinions expressed in this report are those of the individual contributors and do not necessarily reflect the official policy or position of the Montreal AI Ethics Institute (MAIEI) or its affiliated organizations. The information contained in this report is for general informational purposes and should not be construed as professional advice. While every effort has been made to ensure accuracy, MAIEI makes no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, or suitability of the information contained herein.

Contact Information

Website: https://montrealethics.ai
Email: support@montrealethics.ai

Contributors

This year's *State of AI Ethics Report (SAIER) Volume 7* represents the collective expertise and insights of **researchers, practitioners, policymakers, industry advocates, MAIEI collaborators** and **advisors** from around the world. We are deeply grateful to each contributor who shared their knowledge, analysis, and perspectives to make this report a comprehensive resource for the AI ethics community.

Your contributions help ensure that conversations about AI ethics remain grounded, inclusive, and forward-looking. Thank you for lending your voice to this important work.

Anna Sikorski
ACTRA Montreal

Kent Sikstrom
ACTRA National

Shi Kang'ethe
AIVERSE

Trisha Ray
Atlantic Council

Linda Solomon Wood
Canada's National Observer

Rosa E. Martín Peña
Centre for Ethics and Law in the Life Sciences (CELLS), Leibniz University Hannover

Kirthi Jayakumar
civitatem resolutions

Aimee Li, Anna Zhou, Chelsea Sun, Kanika Singh Pundir, Roberto Concepcion, Rose Simon, & Tao Liu
Encode Canada

Michelle Baldwin
Equity Cubed

Katrina Ingram
Ethically Aligned AI

Jae-Seong Lee
Electronics and Telecommunications Research Institute (ETRI), South Korea

Denise Williams
First Nations Technology Council *(former CEO)*

David Atkinson
Georgetown University

Rachel Adams,
Global Center on AI Governance;
Leverhulme Centre for the Future of Intelligence, University of Cambridge

Daniel S. Schiff
Governance and Responsible AI Lab (GRAIL), Purdue University

Jennifer Laplante
Government of Nova Scotia, Canada

Joahna Kuiper
HiirAI

Kate Arthur
Independent, Author

Jake Wildman-Sisk
Independent, Lawyer

Jonathan van Geuns
Independent Researcher

Priscila Chaves Martínez
Independent Researcher

Amanda Silvera
Independent, Voice Actor

Wan Sie Lee
Infocomm Media Development Authority of Singapore

Ismael Kherroubi Garcia *(Editor)*
Kairoi; The Responsible Artificial Intelligence Network (RAIN); MAIEI

Tariq Khan
London Borough of Camden County Council, United Kingdom

Connor Wright *(Editor)*
MAIEI

Renjie Butalid *(Editor)*
MAIEI

Marianna Ganapini
MAIEI; University of North Carolina at Charlotte

DOI: 10.5281/zenodo.17328882

Ana Brandusescu
McGill University

Renée Sieber
McGill University

Jimmy Y. Huang
McGill University; MAIEI

Elizabeth M. Adams
Minnesota Responsible AI Institute

Seher Shafiq
Mozilla Foundation

Kathy Baxter
Salesforce

Ivy Seow
Singapore Management University

Tamas Makanay
Singapore Management University

Burkhard Mausberg
Small Change Fund

Shay Kennedy
Small Change Fund

Alex Tveit
Sustainable Impact Foundation

Bryan Lozano
Tech:NYC Foundation

Jenni Warren
Tech:NYC Foundation

Adnan Akbar
tekniti.ai

Ayaz Syed
The Dais, Toronto Metropolitan University

Blair Attard-Frost
University of Alberta; Alberta Machine Intelligence Institute

Eliot Tretter
University of Calgary

Fabio Tollon
University of Edinburgh

Roxana Akhmetova
University of Oxford

Zoya Yasmine
University of Oxford

Jess Reia
University of Virginia

Maria Lungu
University of Virginia

Ryan Burns
University of Washington Bothell

Tania Duarte
We and AI

DOI: 10.5281/zenodo.17328882

Acknowledgments

Thank you to our **Paid Subscribers** who make it possible to keep The AI Ethics Brief free and accessible to everyone. Your support sustains our work in democratizing AI ethics literacy and honouring Abhishek Gupta's legacy. You are part of the **SAIER Champions Circle 2025.**

SAIER Vol. 7 Team:

- **Editors:** Renjie Butalid, Connor Wright, Ismael Kherroubi Garcia
- **Copy Editor & Marketing:** Kei Baritugo
- **Digital Production:** Zahra Mustin

Many thanks to the MAIEI team members and extended community of supporters and advisors:

> Masa Sweidan, Megan Tan, Hannah McGee, Sadia Rafiquddin, Mo Akif, Meriem Mehri, Paolina Buck, Justin Hendrix, Rebecca Finlay, Steve Rennie, Ilias Benjelloun, Simran Kanda, Marc-Antoine Bonin, Liandra Doonan, Karen Yum, Vinod Rajasekaran, Nishan Chelvachandran, Luca Baraldi, Laura Zambarda, Tom Sinclair, and the entire team behind the Wadham Experience, Wadham College, Oxford.

Your dedication, collaboration, and ongoing support made this report possible and continue to strengthen MAIEI's impact.

Special mention goes to Abhishek's family: his parents, **Ashok** and **Asha,** and brother, **Abhijay Gupta.** Your continued connection to MAIEI's mission means the world to us.

Table of Contents

DOI: 10.5281/zenodo.17328882

DOI: 10.5281/zenodo.17328882

INTRODUCTION TO SAIER VOLUME 7

0.1 Opening Foreword - State of AI Ethics Report Vol. 7

By Renjie Butalid, Montreal AI Ethics Institute (MAIEI)

0.2 Philosophy, AI Ethics and Practical Implementations

By Marianna B. Ganapini, PhD, Montreal AI Ethics Institute (MAIEI) & University of North Carolina at Charlotte

0.3 Bridging Policy and Ethics: On the Launch of AI Policy Corner

By Daniel S. Schiff, PhD, Governance and Responsible AI Lab (GRAIL) at Purdue University

DOI: 10.5281/zenodo.17328882

0.1 Opening Foreword - State of AI Ethics Report Vol. 7

By Renjie Butalid , Montreal AI Ethics Institute

A Year of Continuity and Change

2024 was a challenging year. It was the year I returned to lead the Montreal AI Ethics Institute (MAIEI), the organization Abhishek Gupta and I founded in 2018. Since then, I've been building a venture-backed startup in digital assets and risk management, where AI's impact on financial services is immediate, not theoretical. Abhishek's passing brought me back to MAIEI's leadership, allowing me to bring these perspectives together as we sustain the work we started. This *State of AI Ethics Report (SAIER) Volume 7* is part of that continuation.

When we founded MAIEI in 2018, I was building the entrepreneurship ecosystem at McGill University's Dobson Centre for Entrepreneurship, and AI governance was dominated by conversations between policymakers and technologists. It was technical, industry-driven, and largely inaccessible to the public. Something fundamental was missing: the people whose lives would be shaped by these systems weren't part of the conversation. We launched The AI Ethics Brief to change that, to build literacy around AI's societal impacts and bring more voices into the discourse.

What This Report Is

This is not another framework or productivity analysis. *SAIER Volume 7* is a snapshot from our vantage point, bringing together practitioners, researchers, advocates, and communities grappling with AI's realities on the ground.

It captures perspectives from Canada, the US, Europe, Asia, and Africa, though we recognize many voices remain absent. We hope it generates more questions than it answers and encourages people from all walks of life to ask hard questions and be critical. It also reflects my entrepreneurial roots, less interested in abstract principles and more focused on what's actually happening in the field. What does implementation look like? What's working? What's failing? Where are the gaps between policy rhetoric and lived experience?

We ground our work in understanding AI as sociotechnical systems. This means recognizing the technical components (software, hardware, data centres, chips), the social dimensions (people, culture, values, organizations), and the environmental impacts. Ignoring these interconnected realities reduces AI governance to compliance checklists when the real challenges are about trust, power, and lived impacts.

We're also humble enough to know we're not capturing everything. Regions remain underrepresented. This isn't *the* definitive state of AI ethics. It's *a* state of AI ethics, from where we sit, in 2025. AI ethics, fundamentally, is a process and not a destination. This volume is our contribution to this ongoing conversation.

DOI: 10.5281/zenodo.17328882

Who This Is For

SAIER Volume 7 is designed as a resource for policymakers, educators, students, civil society organizations, and anyone trying to understand AI's governance landscape. It's modular by design. Pick a chapter, engage critically, reach out with your thoughts. Find what's relevant to your context and dig in.

Throughout the year, in conversations and panel discussions, people keep asking: "How do I begin to implement AI within my organization? How do I think about AI ethics practically?" This report attempts to address those questions with perspectives from people doing the work in different contexts, sectors, and regions.

Working with my co-editors Connor Wright and Ismael Kherroubi Garcia, and supported by the MAIEI team, advisors, and collaborators, we've structured *SAIER Volume 7* into five parts: **Foundations & Governance, Social Justice & Equity, Sectoral Applications, Emerging Technologies,** and **Collective Action.** Each part contains opening forewords that set context, followed by an anchor piece or two offering specific insights and case studies.

This is a community resource. We welcome ongoing dialogue through The AI Ethics Brief, now reaching over 20,000 subscribers bi-weekly.

The Widening Gap

Those building AI systems and those whose lives are impacted by them continue to exist in separate worlds. This gap isn't accidental. It's structural.

We see it in governance processes designed to look participatory while systematically excluding voices that would centre justice and equity. We see it in 30-day consultation sprints that ask civil society organizations to volunteer expertise while industry stakeholders show up with dedicated policy teams and lobbying budgets. When you design policy without the people who live with its consequences, that policy will miss real harms and overlook community needs.

A recent MIT report found that 95% of generative AI pilots at companies are failing, not because the technology doesn't work, but because of learning gaps and integration failures. The rush to deploy has outpaced the capacity to implement thoughtfully. This echoes what we're seeing in governance: the infrastructure isn't ready, the literacy isn't there, and the participation mechanisms are broken. We're moving too fast to ask the right questions about deployment, consequences, and accountability.

DOI: 10.5281/zenodo.17328882

Power and Participation

If 2025 taught us anything, it's that AI governance is, at its core, about power. Who has it? Who controls it? And crucially, how do communities respond when excluded from decisions that shape their lives?

Governments and corporations consistently frame public participation as a procedural requirement rather than essential to democratic legitimacy. Canada ranks 44th out of 47 countries in AI literacy, according to a recent KPMG-University of Melbourne study. When people can't participate in AI governance because they lack the literacy to engage, that's not just an educational gap. It's a democratic deficit.

AI literacy isn't just about technical expertise (e.g. learning how transformers work). It's about civic competence and understanding power. It's about being able to read a governance framework, know when a policy process is happening, and how to make your voices heard. We already have models of meaningful participation: municipal AI governance experiments, data cooperatives, Indigenous data sovereignty frameworks. These exist. They work. The question is why they remain systematically excluded from mainstream governance processes.

Looking Ahead

We don't have a crystal ball for 2026, but patterns are clear. Geopolitics will continue to dominate AI governance, as the Trump administration's impact on US AI policy, China's state-led governance approach, and the EU's risk framework enforcement represent contests over values, power, and who gets to shape technological futures. Can smaller nations like Canada maintain sovereign AI governance standards when superpowers prioritize technological dominance over safety considerations? This question frames my analysis in **Chapter 1.2: What "AI Sovereignty" Means for Nations Without Superpower Resources."**

Thank you to all the contributors who shared their expertise and insights, and to our global community for engaging, for reaching out, for trying to figure this out together. Thank you to all the subscribers of The AI Ethics Brief who make this work free and accessible. Thank you to my co-editors, Connor and Ismael, and the MAIEI team and advisors. Most of all, thank you to Abhishek, whose spirit and vision continue to guide our work at MAIEI. This volume is dedicated to his memory and to the belief he embodied: that civic competence is the foundation of change.

At the end of the day, AI isn't just about machines, models, or infrastructure. It's about people. What role do we want humanity to play in shaping a future increasingly mediated by algorithms? *SAIER Volume 7* is our attempt to keep that question alive.

About the Author

Renjie Butalid is a technology and policy executive at the intersection of AI ethics, financial innovation, and public interest. A frequent speaker on responsible AI and governance, he is Co-Founder and Director of the Montreal AI Ethics Institute (MAIEI) and VP Business Development at Metrika. Previously, he led McGill University's Dobson Centre for Entrepreneurship, securing $8 million in funding for startup programs, and teaching innovation and entrepreneurship fundamentals in BCom and MBA classes.

Cite this Article

Butalid, R. (2025) Opening Foreword - State of AI Ethics Report Vol. 7. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 11-14. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

0.2 Philosophy, AI Ethics and Practical Implementations

By Marianna B. Ganapini, PhD , Montreal AI Ethics Institute & University of North Carolina at Charlotte

As AI systems become ubiquitous in daily life, we have been stuck in an antagonistic debate about AI and AI ethics, obscuring the importance of fostering human capabilities and flourishing for AI adoption and expansion. Because AI is so pervasive in all aspects of human life, if AI fails to deliver its promise of a better life, the AI project may well be unsustainable in the long-term. The goal of this *SAIER Volume 7* is to offer a way to move beyond the antagonism of "innovation vs. ethics" towards a new way of thinking about these issues.

On the one hand, AI has been framed as an opportunity leading to efficiency gains, productivity, competitive advantage and, for some, even human greatness. Subsequently, ethical issues surrounding AI are often perceived as hindrances rather than opportunities.

On the other hand, many AI skeptics and AI pessimists describe this technology by mostly stressing the ethical and societal harm that comes from it. From moral bias to privacy violations, AI is a wrecking ball of ethical risks that promises to systematically undermine fundamental rights, increase discrimination, exploit cheap labor, destroy the environment and so on.

These opposing views leave AI and AI ethics in tension. If ethics means constraints on risk while opportunity means maximizing adoption, companies will be inclined to either experience ethics as a problem or see AI adoption as too risky.

To break this gridlock, a new way of thinking around AI can highlight how AI's own long-term economic sustainability depends on its potential to increase wellbeing. AI needs to avoid what I call a "capability-erosion feedback loop," where the use of sophisticated AI leads to loss of capabilities (financial, psychological, intellectual).

As examples, consider three possible feedback loops where AI adoption may erode the very capabilities needed for AI to remain valuable:

1. **The economic loop:** If AI automates jobs without creating equivalent opportunities for capability development, workers lose both income and skills.
2. **The cognitive loop:** Research indicates that employees, students and many professionals often delegate thinking to AI systems. The result: declining ability to evaluate AI outputs and identify hallucinations as users become more dependent on tools that are at times wildly unreliable.
3. **The psychological loop:** AI systems optimize decisions across domains (e.g. what to watch, read, buy) and start carrying out tasks better than humans (e.g. write, paint, entertain).

DOI: 10.5281/zenodo.17328882

Once we take these trends seriously, we realize AI development and adoption require a different framework: AI as a key part of what we may call “the good life,” a life in which we can flourish as humans. This path treats ethics as capacity building. Instead of only asking “What AI must we avoid?”, we ask “Which human capabilities will this AI system expand, for whom?” We need to incorporate AI into our lives that actively makes us better off.

Conclusion

In conclusion, if AI is to endure, its legitimacy cannot rest on risk avoidance or profit generation alone. We need to reframe AI as an engine of growth and values for humans. Practically, this means procurement that introduces upskilling-by-design, AI ethics evaluation that tracks wellbeing alongside other metrics, and interfaces that cultivate judgment and skills rather than replace them. Companies that adopt this mandate will build systems people choose and want to use to improve their lives, and will see increased ROIs, as shown in this report. And at MAIEI we hope that this *SAIER Volume 7* will be used as one of the tools to better understand how AI can concretely foster innovation and wellbeing at the same time.

About the Author

Marianna B. Ganapini, PhD is an Associate Professor at UNC Charlotte. She has collaborated with leading universities and companies, and is widely published in AI ethics and philosophy of AI. She is also Co-Founder of LogicaNow, a consultancy specializing in Responsible AI and AI governance. Marianna is the Faculty Director at the Montreal AI Ethics Institute, and a member of ISO and CEN-CENELEC JTC 21 working on AI standards.

Cite this Article

Ganapini, M. (2025) Philosophy, AI Ethics and Practical Implementations. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. Pp. 15-16. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

0.3 Bridging Policy and Ethics: On the Launch of AI Policy Corner

By Daniel S. Schiff, PhD, Governance and Responsible AI Lab (GRAIL) at Purdue University

On March 18, 2025, for The AI Ethics Brief #160, we launched a new running series for the MAIEI newsletter called AI Policy Corner. AI Policy Corner is a space where we reflect on the translation of ethics into policy. Since then, we've published more than 15 short articles.

We've covered topics like deepfakes, mental health, intellectual property, education, and frontier AI. We've focused on subnational policy developments with major implications, in US states like Colorado, Texas, and New York, as well as reflecting on industry developments and academic summits. And we've tried to bring special attention to how AI policy is unfolding around the world, in both countries with established AI leadership, like the US, Japan, South Korea and Singapore, while also surfacing important developments in regions less often followed in global AI ethics and policy conversations, like Turkey and Kenya.

AI Policy Corner is a joint initiative of MAIEI and the center I co-direct at Purdue University, the Governance and Responsible AI Lab (GRAIL). GRAIL's research-focused and student-centered approach shapes the direction of AI Policy Corner in important ways. Most editions emanate directly from student research assistants who analyze policy documents from AGORA (AI Governance and Regulatory Archive), a collaboration between GRAIL and Georgetown University's Center for Security and Emerging Technology (CSET), hosted at the Emerging Technology Observatory.

As the website describes, AGORA is “a living collection of AI-relevant laws, regulations, standards, and other governance documents from the US and around the world,” which “includes summaries, document text, thematic tags, and filters to help you quickly discover and analyze key developments in AI governance.”

Before getting into the policy documents in more detail, it's perhaps prudent to speak about the motivation for AI Policy Corner. First, it reflects our belief that policy is the instantiation of ethics into the public will – how our beliefs about intellectual property, sustainability, global inequality, meaningful work, international security, and so on – move from research and advocacy into practice. Of course, policy is not the *only* channel through which ethics has an impact, but it is one of the most significant ones, hence GRAIL's focus on both AI ethics and AI policy. Only by reflecting on policy, I might argue, can we know whether our ethical visions are coming to pass.

The second motivation for AI Policy Corner is more personal. Followers of The AI Ethics Brief will know that MAIEI has been working diligently to sustain the legacy of its co-founder, Abhishek Gupta, who passed away a year ago and is honored in this issue.

DOI: 10.5281/zenodo.17328882

This is a goal I strongly share: Abhishek was one of the voices I admired above all in the AI ethics space, with the Brief and State of AI Ethics Report constituting amongst the most important contributions in all of AI ethics discourse. He and I began working as industry leaders in AI ethics at the same time, Abhishek serving as Director for Responsible AI at Boston Consulting Group, and I as Responsible AI Lead at JPMorgan Chase.

While complex positions for ethicists to inhabit, it reflected our shared belief that AI ethics needs to be instantiated in practice, and afforded us the opportunity to collaboratively troubleshoot on the complex, practical, and organizational dynamics in play. As a peer and kindred soul, his loss was greatly felt. Although I fear his voice and brilliance are irreplaceable, his contributions and spirit motivated us to safeguard this important work. AI Policy Corner thus represents our attempt to help sustain Abhishek's legacy and impact, advancing MAIEI's work.

A little bit more about AI Policy Corner and why we find it to be a meaningful project, *though we are most open to feedback!* Perhaps most meaningful to us is that the authors of each issue are students. Typically, one undergraduate or graduate student serves as first author, with a second individual serving as reviewer. Students bring their own voices, identifying topics through discussion and based on issues they see as important or neglected, sometimes connected with their own particular interests in AI ethics or home countries.

The student analyses don't just come from the void. Students analyze policy documents from the aforementioned AGORA database, not limited to laws or regulations, according to a structured codebook and annotation process. Two researchers cross-validate each entry, giving student authors deep familiarity with both specific policies and the broader governance landscape. In our weekly discussions, we reflect on both the content of these policy documents and the surrounding politics: which new ethical topics are rising to the forefront, which bills are or are not enacted and why, how industry strategies respond to public concerns, the dynamics of the AI ethics and safety communities, and so on.

Thus, while the primary audience of AI Policy Corner is the readership of the Brief, we are especially keen to witness the learning of students and see their unique voices in the public sphere as they become champions of AI ethics. Our talented team of student authors has gone on to graduate school, interned at the offices of elected officials, and worked in industry, themselves becoming leaders of student academic communities focusing on AI ethics.

Going forward, we hope that AI Policy Corner is useful to the community and provides viable insights. We welcome feedback on the content and approach of the Corner, and how we might provide more benefit. Thank you for reading.

DOI: 10.5281/zenodo.17328882

About the Author

Dr. Daniel Schiff is an Assistant Professor of Technology Policy at Purdue University and the Co-Director of GRAIL, the Governance and Responsible AI Lab. As a policy scientist, he studies the formal and informal governance of AI through policy and industry, as well as AI's social and ethical implications in domains like education, labor, misinformation, and criminal justice. Daniel was the founding Responsible AI Lead at JP Morgan Chase and Secretary of IEEE 7010-2020, the first AI ethics standard.

Cite this Article

Schiff, D.S, (2025) Bridging Policy and Ethics: On the Launch of AI Policy Corner. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 17-19. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

PART I: FOUNDATIONS & GOVERNANCE

DOI: 10.5281/zenodo.17328882

Chapter 1: Global AI Governance at the Crossroads

1.1 Competing AI Action Plans: Regional Bloc Responses to the US and China

By Jimmy Y. Huang, MAIEI and McGill University

1.2 What "AI Sovereignty" Means for Nations Without Superpower Resources

By Renjie Butalid, Montreal AI Ethics Institute (MAIEI)

DOI: 10.5281/zenodo.17328882

1.1 Competing AI Action Plans: Regional Bloc Responses to the US and China

By Jimmy Y. Huang , McGill University and MAIEI

The past few years have been marked by explosive growth in worldwide AI capabilities and adoption, most notably in business, science, media, and personal use. On July 23, 2025, Washington, DC, unveiled the US AI Action Plan, characterized by a low-regulation strategy for AI innovation. Within days, on July 26, 2025, Beijing published the Global AI Governance Action Plan. Summarized in The AI Ethics Brief #170, China's plan calls for multilateral cooperation and governance on AI technology, contrasting sharply with the US plan's protectionist policies and domestic deregulation for AI firms.

Earlier that month, the Trump administration rebranded the US AI Safety Institute (AISI), a body within NIST charged with managing AI risk, to the Center for AI Standards and Innovation (CAISI). Among CAISI's new objectives: "to ensure U.S. dominance of international AI standards." The shift from "safety" to "standards" wasn't semantic. It signaled that AI governance had become an instrument of geopolitical competition.

Strategic Hedging in Practice

Middle powers and regional blocs, most with their own AI frameworks, have been strategically aligning themselves or hedging commitments to these competing visions. While fragmentation of AI policy is an inevitable outcome of countries seeking to differentiate and promote domestic innovation, the proliferation of disparate governing bodies and frameworks has arguably grown distracting rather than guiding. In October 2024, OECD's Futures of Global AI Governance warned that fragmentation in AI policy could "hamper international interoperability, raise or exacerbate risks to human rights and democratic norms, pose barriers to trade and investment, and reduce the diffusion of benefits of trustworthy AI applications." The UN's September 2024 Governing AI for Humanity report echoed this concern, noting that "coordination gaps between initiatives and institutions risk splitting the world into disconnected and incompatible AI governance regimes."

The reality on the ground, however, is more complex than binary alignment suggests. Regional blocs aren't simply choosing sides; they're hedging, adapting, and negotiating autonomy where possible.

African Union (AU)

The AU officially supports an Africa-centric framework with the Continental Artificial Intelligence Strategy, published in July 2024. The AU's 55 member states present a diverse landscape of alignment. Kenya joined the US-launched International Network of AI Safety Institutes (INAISI), while Egypt hosted Chinese telecom firm Huawei's

DOI: 10.5281/zenodo.17328882

Cloud Summit Northern Africa in July 2025. This divergence reflects both the AU's broader infrastructure ties, with several member states adopting Chinese-built telecom infrastructure via the Digital Silk Road (Dahdal & Abdel Ghafar, 2025), and the pragmatic reality that complete alignment with either superpower risks economic vulnerability. Since the July 2025 publication of competing action plans, no official AU-level response has been given.

Association of Southeast Asian Nations (ASEAN)

ASEAN's 10 member states are similarly navigating between frameworks. In February 2025, during the Paris AI Action Summit, Singapore, in close collaboration with Japan and fellow INAISI members, published a Joint Testing Report providing results on multilingual evaluations of LLMs. Later in July, ASEAN's Secretary-General, Dr. Kao Kim Hourn, delivered opening remarks at China's World AI Conference (WAIC) in Shanghai, the same day Beijing published its Global AI Governance Action Plan. In May 2025, the inaugural ASEAN-GCC-China summit yielded a joint statement that explored collaboration opportunities in the digital economy and AI partnerships. These aren't contradictions; they are survival strategies for maintaining policy sovereignty while securing economic partnerships.

Council of Europe (COE)

In September 2024, ten members and three non-members of the CoE signed the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225). Six more nations have signed in 2025. While the CoE is broadly aligned with the diplomacy track in the US's 2025 Action Plan, the July 2025 rebranding of US AISI to CAISI, with its shift from safety evaluation to ensuring US dominance of international standards, represents a fundamental divergence from the CoE's emphasis on evaluating AI risk to human rights. China is not currently party to the CoE AI Convention, leaving the framework's global reach limited.

Gulf Cooperation Council (GCC)

The GCC offers perhaps the clearest case study in strategic hedging. In June 2025, the UAE publicly declared its intention to fast-track a strategic AI partnership with the US, and Microsoft invested $1.5 billion USD in UAE-headquartered G42, replacing its previous involvement with Chinese firm Huawei. Yet Saudi Arabia hedges its position by signaling alignment with the US while simultaneously investing in Chinese AI capabilities. In May 2024, Prosperity7 Ventures, a Saudi fund, participated in a $400 million USD funding round for Chinese AI firm Zhipu AI. This bifurcated approach reflects the Gulf states' recognition that exclusive alignment with either power limits strategic flexibility.

DOI: 10.5281/zenodo.17328882

Fragmentation as Strategy

While not exhaustive, these snapshots illustrate the current global state of AI governance. Some countries have explicitly aligned, but regional blocs are broadly hedging their positions in the US-China race for AI capabilities and standard-setting. The question is whether strategic hedging can translate into genuine autonomy or merely delays inevitable pressure to choose.

What's emerging isn't binary alignment but a complex web of partial commitments, parallel memberships, and calculated ambiguity. Middle powers recognize that AI governance frameworks encode power: determining whose values shape development, whose industries benefit from regulatory alignment, and whose communities bear the costs of misalignment. When the world's largest AI powers frame governance as a zero-sum competition, smaller nations face impossible choices: align with one power and risk alienating another, or attempt to pursue independent paths with limited resources and influence.

Whether 2025's geopolitical competition catalyzes coordination or entrenches fragmentation remains an open question, one that middle powers and regional blocs are actively shaping through their strategic responses.

About the Author

Jimmy Y. Huang is an AI Ethics researcher and a recognized leader in the financial technology sector. An advisor to the Montreal AI Ethics Institute (MAIEI) and a previous B20 delegate within the G20 ecosystem, Jimmy has contributed to global policy discussion on responsible innovation and digital transformation. Jimmy bridges the gap between academia and international policy with practical implementation, advocating for transparency and accountability in an AI-driven landscape.

Cite this Article

Huang, J.H. (2025) Competing AI Action Plans: Regional Bloc Responses to the US and China. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 22-24. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state

1.2 What "AI Sovereignty" Means for Nations Without Superpower Resources

By Renjie Butalid , Montreal AI Ethics Institute

When the US and China released competing AI action plans within days in July 2025, a quiet question moved to the centre of policy: what does sovereignty mean when you are not a superpower?

For countries like Canada, Australia, South Korea, and Singapore, sovereignty is not total self-sufficiency. It is the practical capacity to govern AI systems at home across five levers:

1. **Data:** Ownership, control, and localization of the data used to train and operate AI systems.
2. **Compute:** Sovereignty over compute resources, such as access to AI chips, hardware, and the infrastructure needed to run large-scale models.
3. **Models and Standards:** The capability to design, train, set standards for, and adapt foundational models to local needs.
4. **Critical Talent:** The ability to cultivate, retain, and develop the skilled workforce necessary to innovate and govern AI independently.
5. **Procurement Leverage:** Strategic use of government and major enterprise procurement to influence the direction, safety, and values-alignment of AI systems.

This capacity serves dual purposes: scaling domestic AI companies to compete globally while ensuring that all systems, whether built domestically or abroad, meet public interest standards.

The Asymmetry Problem

Canada offers a clear view of the challenge. The country hosts world-class AI research through Mila, Vector Institute, Amii, and the Pan-Canadian AI Strategy. Yet the tools Canadians use daily arrive primarily through US vendors and cloud providers. Hospital EHRs, bank analytics, and federal productivity suites depend on foreign platforms, even with data residency controls.

The power imbalance extends beyond technology stacks. Microsoft's market capitalization ($3.85 trillion USD, as of November 2025), exceeds Canada's GDP ($2.24 trillion USD in 2024), fundamentally shaping negotiating dynamics. Middle powers need both to scale domestic champions and govern existing dependencies. Without either, they lose leverage.

Building Counterweights

Ottawa is building counterweights. Budget 2024 allocated $2 billion CAD over five years for a sovereign compute strategy, comprising up to $700 million CAD for commercial data

DOI: 10.5281/zenodo.17328882

centres, $1 billion CAD for public supercomputing infrastructure, and $300 million CAD for SME access through the AI Compute Access Fund. The government also launched the Canadian Artificial Intelligence Safety Institute (CAISI) in November 2024 to grow testing capacity and align with peer networks.

Yet Bill C-27, which contained the Artificial Intelligence and Data Act (AIDA), died in January 2025. Near-term governance leans on procurement and sectoral rules. Canada builds evaluation capacity but cannot dictate terms to foreign platforms running its institutions. Dependency creates lock-in through technical debt and ecosystem integration. Sovereignty requires maintaining exit *readiness*, not just formal rights.

Strategic Paths

Costs and constraints are structural. Compute sovereignty needs data centres, energy, skilled operations, and reliable orchestration. Data centres also consume 1-1.5% of the world's electricity, facing environmental limitations and lengthy permitting timelines. Regulatory sovereignty needs bargaining power against firms whose annual revenues rival or exceed the GDP of middle powers.

Middle powers have three strategic paths, often pursued in combination:

1. **Selective sovereignty.** Draw red lines for critical domains such as health data, elections, and sensitive government systems. Require in-country controls and auditable deployments. Accept commercial dependency elsewhere.
2. **Coalition sovereignty.** Pool resources with peers. The Nordics share supercomputing infrastructure. ASEAN harmonizes data rules. Coalitions convert small markets into demand signals, though even EU members diverge on localization.
3. **Tactical alignment.** Adopt a larger rule set and shape it from within. The EU's AI Act offers pooled sovereignty. Outside formal blocs, alignment lowers compliance costs at a price in flexibility.

Peer Approaches

The EU AI Act sets baselines for exporters whose AI outputs are used in the EU. Australia targets high-risk uses through procurement. South Korea pairs legislation with cloud investment while still being tied to Chinese manufacturing for critical inputs. Singapore's National AI Strategy 2.0 invests in compute, trusted data flows, and governance tools like AI Verify. Middle powers navigate dependency strategically rather than eliminating it.

Priorities for Middle Powers

At the Montreal AI Ethics Institute, we see three essential actions:

1. **Compute access.** Bookable capacity with transparent allocation and reserved public-interest testing. This enables firms to scale in sectors like health and robotics while building evaluation capacity to govern all systems.

DOI: 10.5281/zenodo.17328882

2. **Interoperable assurance.** Map model evaluations and incident reporting to EU conformity routes and other partner test regimes. Joint work through safety-institute networks makes compliance portable.
3. **Procurement as policy.** Public contracts should require incident disclosure, independent evaluations, and provenance features to ensure transparency and accountability. This creates competitive advantages for firms meeting higher standards while avoiding vendor lock-in.

Working Definition

AI sovereignty refers to the ability to establish conditions, verify them, and exit when those conditions aren't met. It is less about building every layer domestically, i.e. data, compute, models and standards, critical talent, and procurement, and more about preserving leverage across the layers that matter. This requires scaling domestic champions while ensuring dependency preserves choice through evaluation capacity, coalition leverage, and procurement power. Nations without superpower resources can still govern in the public interest and compete economically if they invest strategically in both.

The architecture being assembled now will harden into defaults. Middle powers can shape those defaults if they know where their leverage sits and use it in concert. The question for Canada and similar nations is to ensure the architecture is built *with* us, not *around* us.

About the Author

Renjie Butalid is a technology and policy executive at the intersection of AI ethics, financial innovation, and public interest. A frequent speaker on responsible AI and governance, he is Co-Founder and Director of the Montreal AI Ethics Institute (MAIEI) and VP Business Development at Metrika. Previously, he led McGill University's Dobson Centre for Entrepreneurship, securing $8 million in funding for startup programs, and teaching innovation and entrepreneurship fundamentals in BCom and MBA classes.

Cite this Article

Butalid, R. (2025) What "AI Sovereignty" Means for Nations Without Superpower Resources. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 25-27. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 2: Disentangling AI Safety, AI Alignment and AI Ethics

2.1 The Institutions Behind the Concepts

By Renée Sieber, McGill University

2.2 The Contested Meanings of Responsible AI

By Fabio Tollon, University of Edinburgh

2.3 The Evolving AI Safety Conversation: Singapore's Practical Path Forward

By Wan Sie Lee, Infocomm Media Development Authority of Singapore

DOI: 10.5281/zenodo.17328882

2.1 The Institutions Behind the Concepts

By Renée Sieber, McGill University

The year 2025 was pivotal in concepts related to AI ethics, as seen at the Paris AI Action Summit, where national leaders expressed major concern over China's DeepSeek model. We saw ethical constraints on AI relaxed in the hard law of AI regulations. For instance, the US federal government issued an executive order revoking national and state-level regulations on AI, believed to serve as barriers to innovation. Concepts like "AI safety," "alignment" and "ethics" were largely abandoned or reshaped by national governments in favour of new buzzwords like "AI security" and "digital sovereignty." I will disentangle some of these foundational concepts, concluding with missing elements and a path forward.

AI Safety

Briefly, proponents of AI safety argue that we must understand the anticipated and unanticipated consequences of implementing AI. These consequences are often interpreted as risks, which tend to vary along a continuum of low to high risk. Some are anticipated and intended, such as AI that supports the military; some are unanticipated and unintended, such as wrongful arrests due to facial recognition technology. Some are existential risks, such as AI accelerating the risk of World War III via a computer imbued with some "artificial general intelligence" that decides that humans are irrelevant. Therefore, proponents advocate for the identification and assessment of those risks with remedies that are computational, e.g., ensuring system accuracy and robustness, and policy-based, e.g., development of algorithmic impact assessments.

Globally, there were three major AI safety trends. Several national AI Safety Institutes were formed in late 2024 and operational in 2025. For example, the Canadian AI Safety Institute opened in November 2024 to "leverage Canada's world-leading AI research ecosystem and talent base to advance the understanding of risks associated with advanced AI systems and to drive the development of measures to address those risks." There was the establishment of an international network of safety institutes. Occurring simultaneously with these formations, third and somewhat ironically, was the partial retreat from "AI safety" to "AI security," as with the UK AI Security Institute in February 2025.

AI Alignment

Simply put, AI alignment refers to ensuring that AI, its design and outputs conforms to human values. The underlying concern is that, because AI acts autonomously, it obviates the need of a human-in-the-loop. Therefore, how are human values represented, and, thinking computationally, in what sequence are they valued?

Although alignment and safety overlap significantly, reinforcement learning, rewarding the AI system in its development phase, is a prime method to ensure alignment. That requires understanding how the system learns, since a reward could violate the rules in

DOI: 10.5281/zenodo.17328882

unanticipated ways. Consider the example of creating an AI system to simulate finding a way to win a boat race. Instead of moving faster than the other boats, the AI might propose sinking the other boats to win the race. This leads to a larger fear that, in an unchecked intelligence, AI's unanticipated "solutions" to societal ills could lead to catastrophe.

AI Ethics

AI ethics is often the catch-all term for all forms of responsible AI, although it is primarily associated with certain core tenets, such as fairness, accountability, transparency, trustworthiness, and explainability. In AI ethics, there continues to be an emphasis on remedies like AI literacy to increase trust, computational debiasing to lessen discrimination, participatory design or procedural fairness processes to increase fairness, audits and impact assessments to increase transparency and explainability, and soft law, norms and standards, and hard law, to increase accountability. Despite these measures, ensuring AI ethics remains a challenge.

In 2025, Canada announced a Minister of AI, the first of its kind in the country. Minister Solomon has suggested moving away from AI regulation, which he argues could stifle innovation, in favour of soft law or regulations to protect privacy. He has mentioned AI ethics as necessary to build a trustworthy AI ecosystem that competes globally while protecting Canadians. The government also launched a 30-day national sprint to collect public input on AI. The timeline was too short for a country where surveys reveal a deep distrust of AI compared to other countries. Sprint questions heavily favoured industry and economic development and treated trust solely as AI literacy. There are flaws in focusing on trust, especially when AI literacy appears here to be reduced to understanding the benefits of AI. Recent research has also found that, the more scientists learn about AI, the less likely they trust it, suggesting that other members of the public will respond likewise.

Conclusion

In 2025, AI ethics showed more continuity than change. In practice, issues such as the concentration of wealth and power, the environmental costs of data centres, and the political economy influencing AI's development are avoided. Local perspectives on AI governance remain absent, even as cities and communities confront AI's impacts on the ground. As the public and private sectors adopt new directions such as AI security, digital sovereignty, and public-interest AI, these new approaches must be thoroughly examined to ensure they serve and engage society, rather than simply repackage existing inequalities.

DOI: 10.5281/zenodo.17328882

About the Author

Renée Sieber is an associate professor at McGill University, jointly appointed in Geography and the Bieler School of Environment. She studies the intersection of participatory theory and computational technologies. Renée is best known for her work on public participation in Geographic Information Systems and increasingly civic participation in AI. She was named 2025's 100 brilliant women in AI Ethics.

Cite this Article

Sieber, R. (2025) The Institutions behind the concepts. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 29-31. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

2.2 The Contested Meanings of Responsible AI

By Fabio Tollon , University of Edinburgh

What should we make of the “responsible” in “responsible AI” (R-AI)? It seems like the best fit for this kind of responsibility is to understand it as a term of praise. When we talk about ‘responsible’ AI, we are envisioning the AI technology in question as having been developed or deployed in a way that was commendable and trustworthy. This could mean it meets certain fairness, safety and transparency criteria or is in the service of some public good. Of course, this is not the same meaning as when we say “she is a responsible parent,” but something similar is being tracked. Namely, that something was done in a way that we would think of as “good,” or at least aspiring towards that which is good. The insights below draw on the BRAID UK report on the R-AI ecosystem.

However, R-AI is not just a simple term of praise, but a contested idea that stands for a number of different and at times contradictory things. It sometimes refers to a growing interdisciplinary field of critical AI research: at times, a governance ambition, and at others, a process for producing desirable AI products. Yet R-AI research often criticizes the R-AI governance agenda; meanwhile, many proposed ways to build responsible AI products lack buy-in from AI industry leaders. Thus, the ambiguity comes from the ways the term has been differentially attached to various practices in the AI space, with stakeholders deploying it in different ways. By understanding how these different meanings hang together, we can get a better handle on what we want R-AI to mean.

R-AI as an Interdisciplinary Research Agenda

Research under the banner of R-AI has been carried out by industry since at least 2017, when Microsoft and others first began using the term to brand their new algorithmic fairness, privacy and transparency toolkits. Industry R-AI researchers quickly engaged with academics in the closely related field of AI ethics, as well as public sector efforts to develop more “trustworthy” AI, and nonprofit/civil society researchers studying AI-driven harms. To be engaged in “responsible” AI from this perspective, then, is to be engaged in research that cuts across disciplinary and sectoral boundaries in order to address ethical and societal concerns emerging from AI.

R-AI as a Stated Governance Ambition

R-AI as a governance ambition originated from a desire (and sometimes a need) for tech companies to self-regulate. Since 2017, Meta, Google, Microsoft, IBM, PwC and Accenture have all produced internal R-AI documents, which each offer a set of principles and/or core values. These Responsible AI principles are used to inform the development of various tools and practices, such as internal ethics reviews, risk assessments, and product testing, in the hopes that these will realise particular values within their AI business. Soon after, nations began to frame Responsible AI as a government ambition,

DOI: 10.5281/zenodo.17328882

part of their own innovation strategies. R-AI under this banner refers to a body of effective internal governance procedures, guidelines and guardrails for aligning AI innovation with the values and principles that corporations or governments want to signal to others that they stand for.

R-AI as a Desired Type of AI Product

Outside of particular corporate, government, and research agendas, there is an interest in developing a single reliable system or comprehensive set of standards and techniques for ensuring that AI products and services are 'socially benign' or have responsible characteristics. Here, the target is the technology rather than the developer, user or organisation.

R-AI as an Ecosystem of Contested Meanings

So, which is the "correct" meaning of R-AI? All of the above, one of them, or none? One way to approach this issue is to reframe it: What if we think about R-AI as a broad community or ecosystem of stakeholders?

Instead of conceiving of R-AI in silos, we can tease out the ways these conceptualizations shape one another and we can aim towards a better and more holistic perspective on R-AI. Understanding R-AI in this way allows us to see that the three different meanings outlined earlier hang together, and that there are better or worse ways for them to feed into one another. That is, they each pick out important parts of what a truly 'responsible' AI ecosystem might look like, but none by themselves are enough.

No singular part of an ecosystem is completely isolated from the rest, and so these composite 'ecologies' need to be mapped, managed, and supported in ways that enable flourishing across the whole. The ecological metaphor lets us formulate the holistic goal of Responsible AI, one that aims at a future state of affairs in which responsibility appropriately infuses and guides the complex interactions between the diverse and vast community of actors with a stake in AI and its societal and planetary impact.

The key insight from this initial survey of the different meanings of R-AI is that it is not a singular concept or effort with a fixed meaning and clear definitional boundaries, but a complex and dynamic ecosystem pervaded by tensions and interdependencies.

DOI: 10.5281/zenodo.17328882

About the Author

Fabio is a philosopher of technology with interests in the ethics of AI, moral responsibility, and free will. He is a postdoctoral researcher part of the BRAID (Bridging Responsible AI Divides) program at the University of Edinburgh. He is a research fellow at the unit for the ethics of technology at Stellenbosch University and a research associate at the Centre for Artificial Intelligence Research (CAIR) at the University of Pretoria.

Cite this Article

Tollon, F. (2025) The Contested Meanings of Responsible AI. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 32-34. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

2.3 The Evolving AI Safety Conversation: Singapore's Practical Path Forward

By Wan Sie Lee, Infocomm Media Development Authority of Singapore

A Shift in Global Cooperation on AI Safety

Looking back at 2025, the AI Action Summit, held in Paris in February, represented an inflection point in the global discourse on AI safety, marking a shift in the dynamics of international cooperation. The two preceding global summits, Bletchley Park (2023) and Seoul (2024), established a consensus centered on mitigating catastrophic risks posed by frontier AI. However, the Paris Summit's expansion of a narrow safety agenda to a broader one that encompasses economic opportunity, global equity, and industrial strategy, as well as the US's announcement on prioritising diffusion of US frontier AI capabilities, signaled a de-emphasis on AI safety.

While some momentum for global collaboration on AI safety is lost, balancing the narrative with innovation and adoption allows for more inclusive participation globally. For Singapore, collaboration and partnership continues to be important to address risks arising from the rapid advancement of AI. We focused on mechanisms to do this productively and practically. These are along three non-competitive and apolitical vectors: supporting technical cooperation within the expert community, facilitating the development of best practices and standards in testing and evaluation, and contributing to global capacity building.

Technical Safety and Research Consensus

The universal need for a shared, scientific foundation for AI safety allows for productive cooperation that transcends national political agendas. Building on the International AI Safety Report 2025, the Singapore Consensus on Global AI Safety Research Priorities convened more than 100 AI experts from around the world to exchange ideas and clarify urgent needs for technical AI safety research. The resulting research priorities are organised into three interlinked domains: risk assessment (evaluating risks before deployment); development of trustworthy, secure, reliable systems (during design and build phases); and control, monitoring and intervention (post-deployment). These form a shared agenda and provide a technical roadmap for collaboration within the scientific community that is essential, regardless of national political and regulatory philosophy.

Best Practices and Standards in Evaluation

Advancing measurement science and practices will provide the empirical foundation for evaluating risks. Current benchmarks and testing methods are fragmented. Developing shared and continuously updated evaluation frameworks will allow AI to be tested under consistent and

DOI: 10.5281/zenodo.17328882

robust conditions, supporting cross-border comparability and transparency. Establishing reliable methodologies and standardised metrics, analogous to those in aviation or pharmaceuticals, would make safety claims testable, enabling cooperation between governments, industry and researchers, and ultimately supporting AI adoption.

As part of the International Network of AI Safety Institutes, the Singapore AISI continued to lead joint testing efforts within the Network, working with the other AISIs to develop common evaluation methodologies for frontier models. The latest of these joint testing exercises focused on systemic safety behaviours of AI agents in areas like cybersecurity, data leakage, and fraud. It also included multi-lingual evaluation, drawing on the diverse capabilities within the Network.

To support the greater use of AI, evaluations of AI safety also need to address the reliability and trustworthiness of deployed AI applications and systems, tackling societal risks and tangible, near-term harms. To do this, the Global Assurance Pilot brought together AI deployers and testers to develop testing standards for AI systems. They looked at ways to address risks in AI deployment in healthcare, financial services and other contexts, setting the foundation for global standards in AI application testing.

Safety as Inclusion and Access

Post-Paris, inclusive access and capacity-building are now critical components of the safety agenda. AI safety cannot be achieved by a few advanced economies alone. Through shared testing facilities, open research tools, training programmes, and technical partnerships, capability-building enables all nations to adopt safe and reliable AI. At the UN level, digital cooperation in AI governance is gaining increasing prominence.

Singapore is doing its part to support this work, driving work within our region to develop AI safety frameworks and evaluation capacity that reflect diverse social contexts and languages, and reflect local realities. This also improves the quality and resilience of safety outcomes, as diverse perspectives help identify harms that might otherwise go undetected.

As part of the Forum of Small States, we cooperated with countries that are members of the Forum to put out materials and resource guides that could be useful for other countries; for example, working with Rwanda to create an AI Playbook for Small States. We also set up Singapore Digital Gateway as a platform to share our resources and experience in AI safety, such as culturally-relevant models like Sea-Lion and open-source testing tools like AI Verify.

DOI: 10.5281/zenodo.17328882

Looking Ahead to 2026

In 2025, the global tone for AI safety has shifted. The upcoming India AI Impact Summit in early 2026 is likely to solidify this trend, shifting the conversation even further from "action" to measurable "impact," with themes centred on inclusive development, sustainability, and democratising resources. While the UN platforms, including the upcoming scientific expert panel, will continue to align AI innovation with tangible global development goals, the world is less likely to see a rapid global treaty on AI safety. With geopolitical fragmentation remaining a reality, the path forward for governing this transformative technology lies in scaling these vectors of collaboration: the shared, non-political commitment to technical standards and assurance, combined with a focus on capacity-building and AI for public good.

About the Author

Lee Wan Sie is Cluster Director for AI Governance and Safety at Singapore's Infocomm Media Development Authority, and Executive Director of the AI Verify Foundation. She drives Singapore's approach to AI governance, helping to grow a reliable AI ecosystem and collaborating with governments around the world to further trustworthy AI development and adoption. She also leads policy in Singapore's AI Safety Institute, where she defines Singapore's AI safety policies, sets AI safety R&D priorities, and establishes global partnerships

Cite this Article

Lee, W.S. The Evolving AI Safety Conversation: Singapore's Practical Path Forward. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 35-37. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 3:
From Principles to Practice – Implementing AI Ethics in Organizations

DOI: 10.5281/zenodo.17328882

3.1 AI Governance in Practice: 2025 Trends in Understanding and Implementation

By Ismael Kherroubi Garcia , Kairoi, RAIN and MAIEI

When we hear "governance," we often think of regional, nationwide or international policies, after all, that is where governmental bodies operate. This section is not about *that* governance but the much more relatable policy structures we find in the workplace; across businesses, schools, hospitals and charities; organisations large and small. At this level, national and multinational initiatives may seem quite abstract; after all, why would the EU's AI Act affect me if I simply use AI chatbots to write emails? And how could the UN's independent international scientific panel on AI be relevant to, say, a bakery or a marketing agency? And yet, those multinational initiatives respond precisely to years of signals from the wider business ecosystem; years of entrepreneurs and organisational leaders calling for clarity as to how to best approach AI in an everchanging world. These are the signals that the present article attempts to tap into, seeking to understand not how policy-makers are responding to calls for clarity, but to understand how organisations are creating clarity for themselves in a world where policy seems to be lagging behind.

AI Governance is a Business Necessity

Regardless of the sector in which an organisation operates, it cannot avoid the AI conversation and its implications. AI chatbots have now been readily available to the public for three years. This means that employees may use such chatbots for work-related tasks. These tasks, in turn, are fundamental to work across sectors: AI chatbots may be used for writing, brainstorming, correspondence, summarising texts, and so on. So, how many people are using AI chatbots at work, and is it helpful? The evidence is unclear.

In February 2025, Pew Research Center reported that, in the US, "relatively small shares of workers say they have used AI chatbots for work: 9% say they use them every day or a few times a week, and 7% say they use them a few times a month. [...] Among workers who have used AI chatbots for work, 40% say these tools have been extremely or very helpful in allowing them to *do things more quickly*. A smaller share (29%) say they have been highly helpful in *improving the quality of their work*." (Lin & Parker, 2025). Meanwhile, a report from the Danish Bureau of Economic Research concluded in May that "AI chatbots have had no significant impact on earnings or recorded hours in any occupation" (Humlum & Vestergaard, 2025). Against this stands a global study conducted by the University of Melbourne and KPMG, which suggests that over 50% of workers use AI chatbots, and that their use leads to efficiency gains in over 60% of cases. Notwithstanding, the Australia-led study also emphasizes the risks that come with this rapid adoption of AI in the workplace, evidencing that "almost half of employees admit to having used AI in ways that contravene organizational policies. This includes uploading sensitive company information into public AI tools" (Gillespie *et al.*, 2025).

DOI: 10.5281/zenodo.17328882

In this context, AI governance is a business necessity, as the risk of misusing AI tools or falling for the hype may become costly. As Ganapini & Butalid (2025) explain in *Tech Policy Press*, "AI systems introduce operational, reputational, and regulatory risks." With this, risk management mechanisms become central to protecting business interests; they respond to "market incentives" and remain consistent with pressures from regulators and consumers or beneficiaries.

AI Governance is More than Compliance

The pressures rendering AI governance a business necessity, market incentives, regulations and public influence, help explain that it is a question that goes beyond compliance alone. During a panel discussion hosted by the Responsible Artificial Intelligence Network (RAIN) in London in October, the speakers pointed to the risk of AI governance backsliding into compliance. Leaning on the BRAID UK responsible AI ecosystem report from June (Tollon & Vallor, 2025), the speakers made the case that legislation may inhibit the otherwise holistic and reflective nature of responsible AI initiatives. In other words, rather than AI governance building on decades of responsible research and innovation literature and advocacy, its scope may be narrowed to a series of checklists that ensure legal compliance.

Returning for a moment to the higher-level governance activities mentioned at the start, both the US and the UK have shown in 2025 a retreat from "responsible AI" to compliance in 2025, best demonstrated by their refusal to sign the Paris summit declaration on inclusive AI. In this regard and for the foreseeable future, it will fall to organisations to design and implement AI governance strategies; to approach AI responsibly and with an eye to the societal impacts of their AI-related decisions; to seek independent advice and to promote AI literacy.

About the Author

Ismael is the Founder and CEO of Kairoi, the AI governance consultancy. He is also the Founder of the Responsible Artificial Intelligence Network (RAIN), and Participant Panel and Ethics Advisory Committee member at Genomics England. Ismael holds a master's in Philosophy of the Social Sciences, where he sought to uncover enabling conditions for multidisciplinary collaboration in scientific projects. As a result, a key tenet of his work is to advocate for an epistemic humility.

Cite this Article

Kherroubi García, I. (2025). AI Governance in Practice: 2025 trends in understanding and implementation. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 39-40. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

3.2 Monetization and Closing the Principles-to-Practice Gap

By Joahna Kuiper , HiirAI

Why 2025 Matters

When it comes to AI, if 2024 was the year of technology exploration, then 2025 is the year of monetization (see here and here). New AI capabilities and companies are emerging at a breakneck pace, broadening the domains and markets where AI is being targeted to provide value. The types of technology meant by the term “AI” in popular vernacular have become centered around generative AI and, more specifically, agentic capabilities that focus greatly on the realm of commercial enterprise AI and individual use of agent conversational interfaces. This rapid expansion of use by business and individuals applies a great deal of pressure to treat AI as innovation as opposed to research, often stemming from an economic ‘result-first’ perspective.

Key Developments

Commercialization competing with AI research goals

OpenAI’s journey from a not-for-profit to a complex for-profit company is the blueprint for financial expectations from AI companies in 2025. This swing contributes to increased pressure to prioritise monetizable AI features at speed.

Regulation Starting to Take Effect

Ultimately, the year of monetization requires that industry experiments and global investments in AI start impacting the bottom line of the spenders. That happens with development and deployment of systems that are profitable across the supply chain, from the AI research companies, to the domain or industry-specific AI system layers, as well as the corporations introducing AI into their workflows. Such focus has fueled a push from ethical AI to more traditional IT domains, such as risk management and IT governance, when addressing the bridge to get from “principles” to “practice” in AI system development. The diffusion of ethical expertise to IT project resources via cross-functional collaboration leads to greater dependency on tools to provide the ethical content to developers.

Stakes for Communities

The principles-to-practice gap is further complicated as the number of practices impacted by the gap continue to expand past model builders to enterprise systems. Model or foundational system builders still need to address questions about how to build AI ethically at its core. These enterprise deployments do not only bring in an increasingly larger set of

DOI: 10.5281/zenodo.17328882

developer or project roles, often staffed by professionals who are experienced in traditional or deterministic IT systems, but also come with a time-to-market mentality. Ultimately, we are in a state of high pressure to successfully deploy AI across markets and industries while, at the same time, deeply entwining core ethical AI principles that may clash with capitalist ideals. How does one champion a principle of, for example, fairness, if fairness isn't a defined part of a company's business metrics; or worse, impedes their financial results?

What to Watch

The most time-sensitive indicator of how AI ethics will make it into AI deployments may be the continuing regulatory evolution. How the first regulatory penalties are defined and imposed will be telling. If the first few legal challenges are in favor of industry over humanity, "ethical AI" will lose a big lever. Two other areas to keep an eye on are jobs in AI trending towards having responsible AI embedded in job descriptions or as distinct roles, and public opinion and consumer behaviour ignoring AI harms in favour of convenience or near-term benefit (consider smart assistants and social media as an example to learn from). If there is little market pressure on AI deployments to be ethical, investment in closing the principles-to-practice gap will not be a priority.

About the Author

Joahna Kuiper is an AI ethics researcher, Responsible AI consultant, and Oxford lecturer, focusing primarily on examining AI deployments as tools in business. Drawing on decades of implementing emerging technologies in industry and applying academic work in AI ethics and futurism, Joahna's work emphasizes transforming existing practices rather than imposing external frameworks, allowing ethical principles to emerge more organically within organizations. Additionally, my research into the psychosocial implications of human-machine relationships informs my work.

Cite this Article

Kuiper, J. (2025). Monetization and Closing the Principles-to-Practice Gap. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 41-42. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

3.3 From Solidarity to Practice: Building Ethical AI Capacity in Africa

By Shi Kang'ethe, AIVERSE

AI is reshaping global systems, and Africa is emerging as a key architect of responsible and context-aware AI innovation. Our work is grounded in the belief that ethical AI capacity cannot be built by mirroring global models; it must grow from collaborative processes that center care, justice, and the lived realities of African communities.

In practice, this is already happening across several fronts. We are currently building a media and knowledge lab that translates complex AI governance debates into accessible, multilingual narratives, producing weekly articles, short videos, and explainers that help policymakers, journalists, parents, and young people understand what responsible AI truly looks like in their lives.

We are also running AI bootcamps and leadership programs across the continent, equipping women innovators, young professionals, and institutional teams with both technical grounding and ethical fluency. These sessions ensure participants work through real datasets, real governance dilemmas, and real use cases that mirror the challenges emerging across African governments and industries today. Alongside this, we are delivering specialized training for ministries, civil society organizations, and private-sector teams, where we unpack bias, fairness, accountability mechanisms, and the unintended consequences of deploying AI systems without proper oversight.

Knowledge Solidarity: Breaking Down Silos

At the same time, we are leading community-informed research that documents public perceptions of AI, responsible innovation, and governance gaps across different African regions. This includes surveys, collaborative workshops, and field-based storytelling that ensure our insights are grounded in lived realities rather than external assumptions. Every dataset, narrative, and training session feeds back into a broader ecosystem we are actively building, one that strengthens knowledge solidarity, expands responsible AI literacy, and ensures African voices shape the future of AI policy, design, and deployment.

During a recently held webinar, Part 2: Global Solidarity in the Age of AI: Justice, Resistance, and Co-Creating the Future, hosted by the Inter-Council Network, I spoke about knowledge solidarity and the equitable sharing of insights, frameworks, and practices across borders. This idea serves as the backbone of the work we're currently undertaking. Across Africa, countless innovators are building solutions rooted in local languages, cultural nuances, and lived realities, yet their work rarely reaches the right platforms. If we want sustainable and ethical AI capacity, we must break down those silos.

DOI: 10.5281/zenodo.17328882

For us, this means intentionally creating spaces where African researchers, policymakers, and technologists can exchange knowledge openly, not as passive recipients but as contributors shaping global understanding.

Through our work with students in our bootcamps, innovators in our community hackathons, and interdisciplinary teams in our AI Labs, we've learned that our role is not to turn everyone into AI specialists; it is to build a culture where every participant, regardless of skill level, age, or background, feels their voice genuinely shapes ethical boundaries. This has become even clearer as we prepare for our upcoming Responsible AI Hackathon in Mauritius and our advanced training program for industry experts, where policymakers, engineers, creatives, and ethicists will co-design solutions side by side. These initiatives demonstrate that when people are given meaningful, accessible entry points into AI governance, they bring forward perspectives that deepen ethical practice and strengthen the continent's collective capacity to guide AI responsibly.

Building Capacity That Grows from Within

But knowledge solidarity also means acknowledging the divides we often overlook. One of the biggest is generational. During the webinar, I addressed how older generations often step back from AI conversations, convinced it's "for the young." This quiet self-exclusion becomes another form of inequality, not one of access, but of agency. Ethical AI cannot afford that because everyone deserves a seat in shaping the technological future that will shape them. That's why we run "AI for All" convenings, translating AI concepts into accessible language and inviting people across age groups to reflect on how technology intersects with their daily lives. The responses consistently remind us that ethical reasoning already exists deeply within our communities, even if not expressed in technical language.

Global solidarity, in this sense, must go beyond rhetoric. It has to be built on shared responsibility, mutual respect, and an understanding that Africa does not need to be rescued, it needs room to lead. The years ahead will be decisive in how African institutions assert agency. Not by importing Western governance models, but by designing systems that reflect the continent's pluralism, resilience, and creativity. Our recent collaborations with universities, ministries, and civil society groups reflect this shift. Whether we are co-designing ethical data guidelines, supporting the drafting of AI governance frameworks, or training institutional teams on fairness and accountability, the goal is the same: to nurture internal capacity that grows from African values outward, not the other way around.

This commitment to practice shows up across all our programs. Our AI bootcamps train emerging innovators not just in model development but in governance dilemmas, dataset risks, and harm mitigation. Our leadership programs support professionals across various sectors in integrating ethics into real-world decisions, from data collection to deployment. Our workshops with government teams help them adapt governance frameworks into living documents that evolve alongside technology but remain anchored in consent,

transparency, and fairness. Our public education content, ranging from videos to explainers and multilingual storytelling, ensures that AI knowledge does not remain confined to institutions but instead flows into homes, communities, and public debate.

Cultivating Bridges, Not Just Builders

We've learned that ethical capacity is not built in a single training or policy; it is built by cultivating people who act as bridges between research, regulation, and real-world implementation. It is built by creating ethical data ecosystems that respect indigenous knowledge and prioritize collective governance. It is built by documenting public perception and understanding how communities interpret risk, fairness, and trust. And it is built by continuing to test, iterate, and share tools openly so that others across the continent can adapt and improve upon them.

Responsible AI in Africa is unfolding now. It lives in the labs we run, the communities we learn from, the institutions we support, and the knowledge we share. It is the practical, everyday work of ensuring that innovation is accountable, inclusive, and culturally grounded. And it serves as a reminder that global solidarity in AI begins with one simple yet powerful commitment: to build systems that care for the people they serve.

About the Author

Shi Kang'ethe is the Head of AI at AIVERSE, a pioneering platform by Esoteric Strats advancing AI adoption and training across Africa. In this role, Shi leads strategy and innovation, helping institutions, businesses, and NGOs integrate artificial intelligence across their systems, from research and strategic planning to real-world implementation. With a focus on Africa's unique challenges and opportunities, Shi drives efforts that empower organizations not just to use AI, but to innovate with it and shape the continent's digital future.

Cite this Article

Kang'ethe, S. (2025). From Solidarity to Practice: Building Ethical AI Capacity in Africa. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 43-45. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

PART II: SOCIAL JUSTICE & EQUITY

DOI: 10.5281/zenodo.17328882

Chapter 4: Democracy and AI Disinformation

4.1 Legislating the Moving Digital Terrain

By Rachel Adams, Global Center on AI Governance and Leverhulme Centre for the Future of Intelligence, University of Cambridge

4.2 AI and the Body Politic

By Linda Solomon Wood, Canada's National Observer

4.3 Reinforcing the Feedback Loop: How AI in Elections Deepens Democratic Inequities

By Seher Shafiq, Mozilla Foundation

DOI: 10.5281/zenodo.17328882

4.1 Legislating the Moving Digital Terrain

By Rachel Adams , Global Center on AI Governance and Leverhulme Centre for the Future of Intelligence, University of Cambridge

Our information systems, the cornerstone of informed public discourse and effective governance worldwide, have faced unprecedented challenges over the past two years. This period has seen a confluence of factors, including the proliferation of generative-AI fueled misinformation and disinformation, the rise of echo chambers fueled by social media algorithms, and a growing distrust in traditional media institutions. 2024 alone became the "super-election" year, with more than four billion people eligible to vote their political representatives in India, the EU, the UK, South Africa, Mexico, Indonesia, and the US. The recent strains on public information systems have collectively tested the resilience and integrity of how citizens in countries around the world access, process, and interpret information vital for democratic participation.

We saw what many expected: synthetic robocalls, cloned voices, and fabricated videos; but also what fewer anticipated: the more ambient harms of low-grade synthetic "slop" saturating feeds, and a spreading uncertainty about what can be trusted at all.

In India's mammoth election, political operatives industrialized voice cloning and personalized AI videos across dozens of languages; Bollywood deepfakes went viral; and even avatars of deceased leaders "returned" to endorse successors. Meanwhile, in South Africa, deepfaked endorsements and threats characterised the pre-election period, including from then US President Joe Biden supposedly promising sanctions if the ruling African National Congress won power again. And in the midst of these information crises, Meta decides to remove its fact-checkers and reduce its content moderation functions across Facebook and Instagram.

Across the Majority World, fact-checking networks like AfricaCheck became first responders, often with minimal resources. Their work is demonstrative of the critical gap that major social networking companies leave in monitoring inflammatory and malicious content outside of North America and Europe. Unlike in wealthier democracies, in addition to a lack of meaningful content moderation from the major platforms, the local resources for detection, moderation, and civic education are often thin.

For communities in the Global Majority, the stakes of AI-powered disinformation are arguably higher. Many societies already contend with fragile trust in institutions, limited press freedoms, and stark inequalities in access to accurate information. In such environments, even modest volumes of AI-generated disinformation can tip the scales, inflame ethnic tensions, or suppress turnout.

DOI: 10.5281/zenodo.17328882

Policy Response

Platforms and providers scrambled to show responsibility. 27 major tech companies, including OpenAI, Google, Meta and TikTok, signed *A Tech Accord to Combat Deceptive Use of AI in 2024 Elections* in Munich, pledging to curb deceptive AI in elections, invest in provenance tools, and coordinate responses. Critics called it voluntary and uneven.

Significantly more concrete, on August 1st, 2024, the EU AI Act entered into force with phased obligations which explicitly covered deepfakes, including transparency duties for systems that generate or manipulate image and audio. Together with the transparency obligations for large platforms and responsibilities to undertake risk-mitigation measures under the EU's Digital Services Act, Europe has begun to hard-wire information-integrity duties into law rather than rely on the voluntary measures of platforms. Whether those duties can be enforced consistently and globally, particularly across regions in the Global Majority where platform oversight is badly needed, is the question that now matters.

Looking Ahead

Looking ahead, four trends will determine whether democracies adapt or falter in the age of AI.

1. **The evolution of regulatory baselines.** Will countries outside Europe adopt binding standards for disclosure, labelling, and liability? Or will the Global Majority remain subject to the uneven spill-over of Western rules? The spread of enforceable norms tailored to local contexts, will be decisive.
2. **The future of provenance and authenticity.** Watermarking, content credentials, and authenticity infrastructure are advancing rapidly. Yet unless these tools become universal, interoperable, and verifiable by independent actors, they risk being another partial solution that creates a false sense of security.
3. **Platform accountability in practice.** The voluntary accords of 2024 were only a first step. The next phase will be whether platforms disclose enforcement data, open themselves to audits, and show consistent treatment across regions and languages.
4. **Democratic resilience from below.** The most overlooked determinant will be civic capacity. Investments in independent media, fact-checking networks, and public education are as important as any technical safeguard. Communities that can rapidly contextualise and debunk will blunt the force of synthetic disinformation.

DOI: 10.5281/zenodo.17328882

Conclusion

For Cory Doctorow, the problem is is not that the internet and social media is the most pressing singular concern of our time; rather, it is that these digital terrains are the site upon which all the other complex issues of today – inequality, genocide, racism – take place and are mediated.

The events of our recent history have shown both the vulnerabilities and the resilience of democratic institutions and community efforts. As we move forward, the challenge is clear: to regulate not only the tools of manipulation but the practices that undermine collective agency, and to support communities where the risks are greatest.

About the Author

Rachel Adams, PhD, is the Founding CEO of the Global Centre on AI Governance. She is the author of The New Empire of AI: The Future of Global Inequality (Polity Press, 2024). She is an Assistant Research Professor of the Leverhulme Center for the Future of Intelligence, University of Cambridge, and an Honorary Research Fellow of The Ethics Lab at the University of Cape Town. She holds degrees in English Literature, International Human Rights Law and Philosophy.

Cite this Article

Adams, R. (2025) Legislating the Moving Digital Terrain. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 48-50. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

4.2 AI and the Body Politic

By Linda Solomon Wood, Canada's National Observer

"This website I'm quoting from, I don't know a whole lot about," Alberta councillor Patrick Wilson told Cochrane's council in June 2025. "But I just thought their words were better than mine."

His motion to abandon the town's 20-year climate commitment nearly passed. The words weren't his, they came from an AI chatbot designed to kill climate policies. This moment reveals the systematic erosion of authenticity: AI systems that can generate convincing political messaging while concealing their artificial origins.

In January 2025, Rory White brought a custom-built tool he'd built to *Canada's National Observer* (*CNO*), the investigative, climate-focused publication I founded in 2015. The tool, named "Civic Searchlight," combed through YouTube archives of municipal meetings, reviewing transcripts of the meetings for unusual patterns. It didn't take long: councillors in different provinces,thousands of kilometres apart, were speaking in the same voice. Whole sentences lifted verbatim.

Civic Searchlight revealed that a group called "KICLEI" was using an AI chatbot to flood Canadian municipal councils with climate misinformation. KICLEI deliberately named itself to mimic ICLEI, the legitimate international sustainability network. The AI-generated letters appeared to come from an environmental organization when they actually promoted anti-climate messaging.

This represents what some of us had already suspected: using AI to scale deception at a level that was under the radar until Rory White detected it using Civic Searchlight and his investigative journalism skills. One AI system can saturate thousands of officials' communication channels simultaneously, each receiving personalized, locally relevant messaging that appears to come from concerned citizens.

Traditional verification methods break down when AI can impersonate legitimate organizations while generating content that passes authenticity tests.

The effectiveness of these campaigns led to measurable consequences. We documented councillors in Cochrane, Thorold, Pembroke, Peterborough, and Pickering quoting KICLEI verbatim in council meetings and preparing to vote against funding for projects aimed at lowering their municipalities' carbon footprints.

Partners for Climate Protection ties member municipalities to the $1.65 billion CAD Green Municipal Fund, which finances retrofits, transit upgrades, and green infrastructure. When Thorold, Ontario, voted to withdraw, projects were shelved, grants put at risk, and jobs lost. If ten towns follow, that could mean tens of millions in cancelled contracts and hundreds of construction jobs wiped out.

Civic Searchlight enables researchers to track policy language across municipalities for any issue – housing, education, healthcare, immigration – anywhere identical talking points appear across jurisdictions. Seeing this motivated us to share the tool for free, so hundreds of people will now be able to find things. In the first week, 489 people signed up representing major Canadian media, universities, civil society groups, and municipal officials. We made the tool available to journalists, researchers, and civil society groups. We're still in beta mode, developing it based on what we learn from early adopters and there is a simple vetting process.

By the first month, 560 people had signed up. “Our region has been a hot spot for misinformation campaigns against local climate action, and this excellent tool should help us stay informed for what is coming up at municipal meetings,” wrote one researcher to CNO after using Civic Searchlight.. A journalist wrote, “For fact-checking and research purposes. (This is an amazing tool!!)”

Over the past eight years, the epidemic spread of technological capabilities of what we used to call "fake news" has been as dangerous and fast-moving as a global pandemic. In 2017, I described fake news as "a viral infection that threatens the body politic," arguing that "just as the human body's immune system relies on multiple layers of defence, we have to as well." But AI manipulation works differently, where fake news spreads through social media, AI doesn't need human amplification. It doesn't distract from local engagement; it infiltrates those information pathways directly. AI systems insert themselves directly into the channels councils depend on for citizen input. The threats posed by untethered AI disinformation campaigns extend beyond civil discourse to the climate, the very air we all breathe.

When bad actors use AI to diminish the integrity of public discourse, organizations that can contribute to defense should do so. The tool becomes more valuable when widely used because defending democracy requires network effects. The councillor's preference for AI-generated talking points over his own judgment represents a broader challenge: if elected officials can't distinguish AI manipulation from citizen input, what happens to public deliberation?

But how far we've come since 2017, and how fast. When I wrote about fake news as a viral infection, there was no ChatGPT, Claude, or Gemini to carry propaganda at lightning speed. The ethical questions get bigger every day.

One thing Canada could do at the federal regulatory level would be to impose criminal penalties for using AI to impersonate legitimate organizations or people in political communications. As I understand it, this would require updating fraud laws for the AI era.

Meanwhile, AI will readily execute any deception. All it takes is a prompt, a bot, and a creative person who wants to influence policy outcomes. Like the Trojan Horse, it destroys from within by masquerading as the authentic citizen voice democracy depends on.

About the Author

Linda Solomon Wood is the founder and publisher of Canada's National Observer and launched the Democracy & Integrity Project in 2018, a CNO initiative to investigate and reveal disinformation. She is a frequent speaker and has led award-winning reporting on democracy and climate change for over two decades.

Cite this Article

Solomon Wood, L. AI and the Body Politic. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 51-53. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

4.3 Reinforcing the Feedback Loop: How AI in Elections Deepens Democratic Inequities

By Seher Shafiq, Mozilla Foundation

AI systems amplify the gamification of voter engagement, making marginalization a self-fulfilling prophecy.

The gamification of Get Out the Vote (GOTV) efforts has been a longstanding source of exclusion for those who are already the most underrepresented in civic spaces, even before AI tools like augmented analytics and natural language processing became available to campaigns.

In Canada, voter lists are compiled from Elections Canada data and refined through canvassing; categorizing voters based on their likelihood of supporting a particular party. Campaigns then prioritize "likely supporters" for volunteer outreach, creating a feedback loop that continually refines the list of potential supporters. The issue is that already marginalized communities are less likely to vote, and rarely make it onto campaigns' "likely supporter" list, meaning that low-income and immigrant households, for example, do not have candidates knocking on their doors.

For these demographics, this lack of candidate engagement reinforces the perception that politicians do not care about them, resulting in lower civic engagement and voter turnout among those groups. In Canada, this is especially true for racialized minorities. As efficient as voter lists are for campaigns, they "gamify" elections at the risk of impeding democratic engagement among those who are arguably most impacted by policy.

GenAI catalyzes disinformation into full-scale disengagement of marginalized voters.

If the structural issues with voter lists driving engagement weren't enough, the issues of misinformation and disinformation, particularly in the last several years, have created chaos in the information ecosystem. Just weeks ago, the Andrew Cuomo campaign launched a racist AI-generated attack ad against Zohran Mamdani, depicting an AI version of him running through New York, eating rice with his hands.

In Canada, political misinformation online has become uncontrollable. A 2025 report from Canada's Media Ecosystem Observatory found that, in the lead-up to the federal election, more than one quarter of Canadians were exposed to sophisticated fake political content. After the election, more than three-quarters of Canadians reported they believe misinformation impacted the election. The Canadian Digital Media Research Network found that "Canada's 2025 federal election upheld its integrity but exposed a digital ecosystem under mounting strain," citing many examples of deepfakes, AI-generated

DOI: 10.5281/zenodo.17328882

news, and bot activity. Even AI Chatbots used to verify the validity of information have resulted in incorrectly confirming fake content is real, aiding the spread of misinformation.

AI in elections impacts public trust; community-based solutions can help.

Public sentiment analysis on elections does not reflect the public.

Even before AI tools were available to support and report on public sentiment analysis for elections, political polling in the lead-up to elections was an unreliable marker of real public sentiment, as only those already engaged went through the effort of completing phone polls, leading to results that represented a sliver of engaged Canadians.

Reporting out this misleading information spurs a cycle of disengagement. In the 2022 Ontario provincial election, for example, many believed that there was no point in voting in the race because polling consistently indicated that incumbent Doug Ford would win. That election saw the lowest voter turnout in Ontario provincial history, sitting at 43%.

Since then, AI has been added to the mix, often scraping online spaces to read public sentiment and reporting outwards as part of election coverage by media (see the AI tool, Polly, as an example). However, again, those who are most impacted by politics (immigrants, those with low literacy or lower English language skills) are often not engaging in those online spaces, and their sentiments are not captured as part of this analysis. What is reported as public sentiment in an election race is thus an inaccurate representation of real public sentiment, invisibilizing those who are already marginalized, and causing them to disengage further.

Scaffolding against disengagement through community-based organizations.

Tailoring voter engagement materials to specific demographics has been a best practice for community organizations (see Journeys to Active Citizenship at North York Community House, The Canadian-Muslim Vote, Apathy is Boring, Operation Black Vote Canada, and more). However, civic and democratic engagement efforts for marginalized groups in Canada have traditionally been severely underfunded.

Some solutions to the unintended harms caused by AI systems in elections include:

- **Better funding for grassroots community organizations** to deliver digital literacy training. Doing so would arm those most disenfranchised in civic spaces to better identify mis/disinformation and find ways to meaningfully engage in the system.
- **Enhanced funding and support for grassroots efforts** to promote civic and voter engagement in marginalized communities. This could include bolstering online public sentiment analysis with in-person surveys or community focus groups so that reported public sentiment is more reflective of diverse perspectives.

DOI: 10.5281/zenodo.17328882

- While CSIS has noted that GenAI will undermine citizens' trust in democracy, more needs to be done to **update campaign rules in Canada** to hold political campaigns accountable so that using AI systems to create impersonations or deepfakes is no longer allowed. A start would be to update the Elections Canada Act to cover GenAI and deepfakes, which has not yet been done.

Without bolstering community-based initiatives, promoting digital literacy, and implementing concrete policy changes to the Elections Canada Act, AI will continue to exacerbate inequities in our democracy. As the Canadian Digital Media Research Network warns, Canada must "act now to protect future elections" from the impacts of AI.

About the Author

Seher leads global community engagement for Mozilla Foundation (the non-profit behind Firefox), where she focuses on building a better tech future that puts people first.

Cite this Article

Shafiq, S. (2025). Reinforcing the Feedback Loop: How AI in Elections Deepens Democratic Inequities. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 54-56. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 5: Algorithmic Justice in Practice

5.1 Algorithmic Justice vs. State Power

By Blair Attard-Frost, University of Alberta & Alberta Machine Intelligence Institute

5.2 AI Ethics and Gender Diversity in the US: From Surveillance to Resistance

By Jess Reia, Digital Technology for Democracy Lab, University of Virginia

5.3 Beyond the Algorithm: Why Student Success is a Sociotechnical Challenge

By Adnan Akbar, tekniti.ai

DOI: 10.5281/zenodo.17328882

5.1 Algorithmic Justice vs. State Power

By Blair Attard-Frost , University of Alberta & Alberta Machine Intelligence Institute

2025 saw many AI researchers and practitioners reckoning with a new political reality: in pursuing algorithmic justice, government is an unreliable ally at best, and a malicious actor at worst. The Trump administration's assault against immigrants, racialized people, trans people, women, and so many others amply demonstrates the malicious potential of state actors. To see the thinly veiled hatred and hostility at the heart of Trump's AI policy agenda, one need look no further than the replacement of all Biden-era AI safety policies with a new anti-equity executive order entitled "Preventing Woke AI in the Federal Government." Meanwhile, in the EU, the UK, Canada, and elsewhere, deference to the Trump administration's AI-industrial development ambitions has resulted in governments slowing their regulatory plans. In doing so, these governments signal that they cannot be relied upon to protect vulnerable communities from algorithmic harm.

For many years, researchers and practitioners have known that unjust AI systems cannot be corrected through technical fixes alone; inclusive practices of design, policy, and governance are regarded as stronger interventions. Now, the events of 2025 have made it starkly clear that unjust AI systems also cannot be corrected through what policy researcher Inga Ulnicane refers to as "governance fixes": centralized, top-down practices of policymaking and decision-making that favour a "narrow and technocratic approach" to AI governance. A key challenge for future progress in algorithmic justice is to implement design and governance practices that are not merely technical fixes or governance fixes.

Community-led approaches to AI design and governance offer a path forward. Community data trusts and data collaboratives are governance models in which community members pool data and collectively determine requirements for its use and stewardship. In response to unauthorized use of creative works in training generative AI models, creative communities are developing datasets and models that prioritize creator rights and participatory governance, such as the choral data trust initiative led by the UK-based Serpentine Arts Technologies. In the US, the American Civil Liberties Union's Community Control Over Police Surveillance (CCOPS) campaign has led to municipal bans and community oversight mechanisms for police use of AI technologies in 26 jurisdictions across the country. By taking a decentralized approach to AI governance, the CCOPS campaign provides communities with important safeguards against algorithmic injustice in the absence of effective state and federal legislation.

Labour organizing against algorithmic injustice in the workplace has a storied history, and worker-led approaches to AI governance have also gained traction in recent years. In 2023, labour strikes by the Writers' Guild of America (WGA) and Screen Actors Guild - American Federation of Television and Radio Artists (SAG-AFTRA) resulted in union regulations being established for the use of generative AI in writing and performances, as well as for training generative AI models on union-protected materials. Many unions and professional associations have recently created guidelines to support their members in

DOI: 10.5281/zenodo.17328882

safely and responsibly using AI in their workplaces. In Canada, prominent labor organizations were also involved in resisting the Artificial Intelligence & Data Act (AIDA), a piece of AI legislation that ultimately failed to pass into law under Canada's previous federal government, which dissolved at the start of this year. The legislative failure of AIDA has been attributed to several factors, including the intensive and prolonged pushback the government received due to its lack of engagement with civil society, Indigenous peoples, and labor in drafting the bill (Attard-Frost, 2025). Organizations such as the Canadian Labor Congress, Canadian Union of Public Employees, and Writers Guild of Canada all submitted commentary to Parliament indicating that AIDA did not provide workers with sufficient protections against exploitative and unsafe uses of AI.

Though there is promise in re-focusing AI design and governance on smaller scales of collective action, power asymmetries between communities and industry-state alliances are significant. Achieving algorithmic justice against a backdrop of state-sponsored injustice will require coalitions that are polycentric, flexible, durable, and multifarious. Researchers must work with communities to identify and co-create effective practices of participatory design and governance. Technologists and policy experts must support communities in building the knowledge and skills needed to resist harmful AI systems and ineffective policies, while also assisting in the creation of collective governance models and tools. Journalists and creatives can help build public awareness of algorithmic injustices by investigating the local impacts of AI and amplifying community stories. Sympathetic public servants can support this shift by re-directing policy priorities and funding from corporate interests to community-centric initiatives.

The challenges of algorithmic justice have evolved far beyond simply fixing biased technology and advocating for government regulation. Those interventions remain necessary, but the key questions to consider are questions of power: Who has the power to shape the development of AI technologies and AI policies? Who does not have this power, and what possibilities for empowerment exist when we are faced with unreliable and malicious governments?

About the Author

Blair is an Assistant Professor of Political Science at University of Alberta and a Research Fellow at the Alberta Machine Intelligence Institute (Amii). Her research applies a trans feminist lens to address challenges of power, participation, and justice in AI governance systems.

Cite this Article

Attard-Frost, B. (2025) Algorithmic Justice vs. State Power. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 58-59. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

5.2 AI Ethics and Gender Diversity in the US: From Surveillance to Resistance

By Jess Reia , Digital Technology for Democracy Lab, University of Virginia

In January 2025, Meta announced the end of its fact-checking program and the removal of long-standing industry-standard policies that protected 2SLGBTQIA+ and other marginalized communities from hate speech and discrimination. Simultaneously, hateful neologisms used by political extremists were included in the new policies. That month, we would also see the intensification of a relentless and still ongoing campaign to dehumanize gender-diverse individuals, taking away rights and safeguards. This process did not start in the current US administration, as civil society organizations have been monitoring the attempts to advance a neoconservative agenda in US state legislatures. For instance, we started the year with over 533 bills that curtailed rights of 2SLGBTQIA+ people, most of them focusing on trans communities through sports bans, end of gender-affirming care, barriers to accurate ID, freedom of expression of educators, and other civil rights violations.

More companies followed suit, such as YouTube and LinkedIn, removing protections against misgendering and deadnaming from their platforms. Big Tech companies are getting closer to the federal government and changing policies to fit into its agenda, raising concerns on the rise of techno-authoritarianism, deregulation and corporate interests prevailing over human rights. The intensification of state surveillance under the current US administration, particularly targeting marginalized communities, has two faces: the erasure of gender-diverse people from datasets, websites and platforms, while also doubling down on monitoring of trans people and trans advocacy. The so-called “anti-woke” AI regulation published by the White House in July 2025 has many problems, including the explicit use of hateful neologisms and a relegation of misgendering people to irrelevance. With the overall dismantling of AI ethics teams in companies and the open proximity of Big Tech to political extremist agendas, the US should be considered a cautionary tale for other countries facing the global coordinated anti-trans movement.

When thinking about the state of AI ethics in relation to gender diverse communities, three areas are relevant to scholarship and practice. First, how the erasure of trans communities and any kind of gender-nonconformity from websites and datasets will lead to biased AI models and algorithms. Particularly in facial recognition technologies (FRT) and predictive analytics, trans and gender-nonconforming individuals are often misidentified or misclassified due to training data that fails to account for complex and fluid gender expressions. This misrecognition can lead to wrongful arrests, denial of services, or even unwarranted law enforcement encounters.

Second, content moderation systems, which are largely automated and opaque, frequently and disproportionately flag or remove trans-related content, silencing vital discussions around identity, healthcare, and social rights. The increasing use of AI for content moderation can be harmful, especially without robust ethical and legal frameworks, since these systems are often trained on biased datasets and designed without meaningful input from trans individuals. With

DOI: 10.5281/zenodo.17328882

governments actively promoting an agenda of erasure, content moderation becomes another challenge in AI ethics.

Lastly, these issues are compounded by the broader context of techno-authoritarianism, where surveillance technologies are increasingly deployed by state and corporate actors to monitor, control, and marginalize dissent and non-normative identities. From airport security to digital IDs, the impacts of AI-based surveillance have deeply impacted gender-diverse communities. In 2025, the Department of Homeland Security quietly removed protections that previously prohibited surveillance based solely on sexual orientation or gender identity, effectively opening the door to state-sanctioned monitoring of LGBTQ+ individuals. Simultaneously, conservative think tanks have pushed for the FBI to classify transgender activism as a form of "violent extremism," a move that could justify the use of expansive surveillance powers against trans communities and their allies. These developments signal a dangerous convergence of technological overreach and political repression.

Together, these dynamics create an AI ecosystem where trans individuals are hyper-visible to systems of control yet rendered invisible in public discourse, undermining both their safety and their right to self-expression. These are problems with no easy solution, but the AI ethics community can take steps toward working closely to the vibrant network of trans rights resistance and advocacy that already exists, building bridges to tackle the discriminatory use of problematic AI and surveillance technologies. Efforts like strategic litigation, public campaigns, and public interest advocacy are good starting points. Additionally, it is key to build a trans digital rights framework that considers the growing role of AI in our lives, and "trans-centered moderation" models that prioritize community governance, appoint trans moderators, and design algorithms with an understanding of trans-specific contexts and challenges. By bringing trans rights concerns to our AI spaces, we can create awareness of the ramifications of techno-authoritarianism for human rights at large.

About the Author

Jess Reia is an Assistant Professor of Data Science and Public Policy and faculty co-lead at the Digital Technology for Democracy Lab, University of Virginia. In 2025, they were selected as an Andrew Carnegie Fellow to study the impact of AI on evidence-based policymaking and memory keeping for gender-diverse communities. Reia works primarily on technology policy, data justice, cities and human rights.

Cite this Article

Reia, J.. (2025) AI Ethics and Gender Diversity in the US: From Surveillance to Resistance. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions. pp. 60-61. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

5.3 Beyond the Algorithm: Why Student Success is a Sociotechnical Challenge

By Adnan Akbar, tekniti.ai

Across universities worldwide, a quiet shift is underway. Institutions are increasingly adopting predictive AI. The concept is compelling: instead of waiting for a student to fail, these algorithms analyze vast amounts of data, from online activity and demographics to education history, to predict who is at risk. Once identified, these students can be offered targeted support, like tutoring or counseling, making intervention proactive. The narrative is one of efficiency and personalized education.

But as we stand at this crossroads in 2025, a critical look at what's actually happening reveals this promise is built on unexamined bias.

A recent quantitative analysis of a real-world student success model at The Open University in the UK, which is forthcoming in *AI and Ethics*, provides a stark, practical case study. The model's goal was simple: identify new students at risk of failing. But the data it learned from was a digital reflection of long-standing systemic inequities.

The student population studied was profoundly imbalanced: 95.8% White and only 4.2% Black. This data codified the structural disadvantages faced by Black students:

- **Socioeconomic Disparities:** Black students were disproportionately represented in the most deprived socioeconomic categories.
- **Educational Gaps:** 33.6% of Black students had low prior educational qualifications, compared to 30.6% of White students.
- **The Attainment Gap:** Most critically, the historical failure rate for Black students was 47.7%, far higher than the 34.3% for White students.

A model trained on this data will inevitably learn to associate being Black with a higher risk of failure.

The team's first attempt followed a common, naive approach: "fairness through unawareness," simply removing ethnicity from the model's features. This baseline model failed. It was significantly less accurate for Black students and developed a higher false positive rate for them, meaning Black students were more likely than White students to be incorrectly flagged as "at-risk."

This is where practitioners and university leaders face a critical choice. The impulse is to find a technical "fix." The study explored several machine learning (ML) strategies to mitigate bias, but the team found every technical fix involved a difficult, real-world trade-off.

DOI: 10.5281/zenodo.17328882

- **The "Common Sense" Fix That Failed:** An intuitive fix, building separate models for each ethnic group, was counter-productive. It actually amplified bias and produced the worst fairness outcomes tested.
- **The "Statistically Fair, Practically Harmful" Fix:** Another strategy forced the algorithm to produce statistically "fair" outcomes, ensuring, for example, that the percentage of students flagged from each group looked equitable. This technical "success" came at a devastating cost: the model's actual accuracy for Black students fell below the original baseline. This is a classic "fix" that harms the community it's intended to help by prioritizing an abstract number over practical utility.
- **The "Apparent Win-Win" Fix:** A third approach, adjusting the "at-risk" sensitivity threshold differently for each group, finally appeared to be a "win-win." It was one of the only methods that increased accuracy for Black students (from 61.4% to 64.2%) while also improving fairness metrics. But this "fix" introduced a new, practical trade-off: it sharply increased the *overall* number of students incorrectly flagged (the false positive rate). This shifts the burden from a data science problem to a resource allocation problem. Is the institution prepared to support the massive increase in students flagged for intervention?

This is where we must move beyond the algorithm. The real "aha" moment came from a retrospective case study of an actual intervention trial. In the trial, students identified as at-risk received a supportive call from a tutor.

The results were transformative. The intervention led to a 16.5% uplift in pass rates for White students. For Black students, the uplift was at 32.1%.

This finding reframes the entire discussion. A high false positive rate is only a problem if the intervention is punitive, stigmatizing, or resource-intensive. But if the intervention is a low-harm, supportive call, the cost of a false positive is minimal. The risk of a false negative, missing a student who genuinely needs help and drops out, is far, far greater.

What’s actually happening here is that an "imperfect" model, even one that "over-targets" a disadvantaged group, can become a powerful engine for equity *if* it is connected to an effective, well-designed, and non-punitive human intervention. The pursuit of algorithmic fairness, then, is not a purely technical search for the "right" metric; it is a sociotechnical design challenge.

This case study provides clear, actionable insights.

- **For practitioners and data scientists,** stop optimizing for abstract fairness metrics in a vacuum. You must work with support teams to understand the intervention. Is it high-harm (like a disciplinary meeting) or low-harm (like a supportive call)? The answer changes whether you should prioritize minimizing false positives or false negatives.
- **For university leadership and policymakers,** your work is to build the *system*, not just buy the model. The real-world case study succeeded because the intervention was human-centric and non-punitive. Your investment in counselors, tutors, and ethical support structures is ultimately more important than the algorithm itself.

- **For community leaders and educators**, the key is to demand a seat at the table. This analysis proves that "fairness" is a context-dependent, normative choice, not a purely technical one. We must evaluate these systems not by their statistical properties, but by their tangible, real-world impact on students.

About the Author

Adnan Akbar, PhD, is an AI leader with 12+ years of experience , providing consultancy to architect scalable AI/ML solutions on AWS, GCP, and Azure. He is a leader in responsible AI, holding an MPhil in AI Ethics from Cambridge. Adnan helps companies scale AI ethically by establishing robust governance frameworks and reducing algorithmic bias. He is recognized as a DataIQ Future Leader for 2025 and endorsed as an Upcoming Future Leader in Data Science by the Royal Academy of Engineering.

Cite this Article

Akbar, A. (2025) Algorithmic Justice vs. State Power. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 62-64. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 6: AI Surveillance, Privacy, and Human Rights

6.1 AI, Surveillance, and the Public Good

By Maria Lungu, University of Virginia

6.2 Challenging Mandated AI in the Public Sector

By Roxana Akhmetova, University of Oxford

6.3 AI, Biometrics, and Canada's Developing Legal Framework in 2025

By Jake Wildman-Sisk, Independent, Lawyer

DOI: 10.5281/zenodo.17328882

6.1 AI, Surveillance, and the Public Good

By Maria Lungu , University of Virginia

Artificial intelligence has become the silent architecture behind modern governance. From determining where police patrols are deployed to assessing eligibility for social services, AI increasingly influences decisions that affect everyday lives. Surveillance has been quietly normalized through algorithms framed as neutral tools of progress. Across the world, political systems diverge in their approach to AI and surveillance. Yet, some share a common fear that innovation is advancing faster than our capacity to protect the rights of those affected.

Surveillance Infrastructure

AI-driven surveillance has moved far beyond the classic frontier of restricted perimeters (e.g., airport checkpoints or dedicated “safe city” zones). Today, it is woven into everyday public life's “connective tissue”: traffic lights, open-street cameras, social service decision-making systems, classroom monitoring, smart-city infrastructure, and more. These systems are typically introduced under the rhetoric of efficiency and safety. However, the trade-off is often one of opacity and loss of agency. Many individuals are unaware of how their data is being collected, processed, and decisions are made on their behalf, or how they might contest those decisions.

Generative AI deepens these tensions. Tools that synthesize text and images or create models of behaviour, what some call “data doubles”, enable organizations to build proxy representations of people based on their historical data. Once you move into such predictive or modelled terrain, the line between simulation and surveillance blurs because you are not just watching what someone does, you are acting on what a model suggests someone might do.

Algorithmic Policing and Civic Profiling

As AI surveillance expands, questions about its legitimacy and oversight are becoming more prominent, particularly in the US, where predictive policing or surveillance tools from Palantir and Flock Safety are increasingly incorporated into municipal contracts. Often described as “evidence-based modernization,” these systems raise important considerations about consent and the balance between innovation and public trust. Modern surveillance increasingly combines observation with predictive modeling, a shift some scholars refer to as “civic profiling,” in which citizens are viewed as data subjects whose behavior is interpreted through algorithmic systems. Understanding these dynamics is essential for designing transparent oversight mechanisms.

DOI: 10.5281/zenodo.17328882

Global Governance Divide

While the US and its allies continue to deliberate over whether or not to regulate private-sector AI, China has pursued a centralized governance model. China's Global AI Governance Initiative reflects China's belief that state coordination and international harmonization can stabilize a rapidly evolving technological landscape. At the 2025 World AI Conference, Premier Li Qiang proposed an Action Plan that includes a 13-point roadmap for global AI coordination, with practical steps to implement the 2023 Global AI Governance Initiative. For China, a structured, state-led framework ensures alignment between innovation and national priorities, while offering predictability to foreign partners. Critics often describe this model as illiberal, yet its appeal lies in the clarity it provides, setting uniform standards and compliance expectations that reduce regulatory uncertainty.

The US has opted for a more pluralistic approach rooted in its federal structure and tradition of market-driven innovation. The AI Action Plan reflects a policy judgment that overregulation could stifle competitiveness and hinder experimentation in emerging sectors. In this view, flexibility and decentralized governance allow states and agencies to tailor oversight to local contexts, preserving space for innovation while advancing ethical norms. However, this same flexibility comes with trade-offs. Regulatory inconsistency and jurisdictional overlap can create uncertainty about what constitutes "high-risk AI." As a result, companies sometimes gravitate toward states with lighter oversight, complicating nationwide accountability and creating uneven protection standards across the public sector. Both approaches reveal different philosophies of governance rather than clear right or wrong answers.

Communities Pushing Back

Yet beneath the institutional lag, something remarkable is happening. Communities are no longer passive subjects of surveillance; they are becoming active designers of oversight.

For example, in Charlottesville, Virginia, residents pushed back on the use of "Flock" cameras. The police department maintains that the system helps solve crimes and recover property, but critics argue it is overly invasive and primarily aids post-crime investigations. Community organizers held meetings, petitioned local government, and engaged directly with the police chief to ensure residents' concerns were heard and addressed.

In Boston and Minneapolis, residents have called for algorithmic impact statements that evaluate privacy risk, public value alignment, and whether an AI system fairly serves community priorities. I like to think of this as a form of "algorithmic localism," where accountability must be built from the ground up, not handed down from abstract frameworks. We do not want the scale and speed of surveillance to outpace democratic scrutiny. The allure of efficiency continues to eclipse the fundamental question: what kind of society are we optimizing for?

DOI: 10.5281/zenodo.17328882

Communities are experimenting with practical, replicable strategies to ensure AI serves the public good. Through participatory audits or even vetoes, residents actively review AI systems and propose modifications, shaping technology in ways that reflect local values. Advocacy groups also run transparency campaigns, using workshops and public forums to make complex technical systems understandable and accessible to residents. These approaches center grassroots expertise and lived experience, prioritizing the knowledge and voices of those most affected over abstract technical assumptions. By doing so, communities reclaim agency.

About the Author

Dr. Maria Lungu is a Postdoctoral Research Associate in the Digital Technology for Democracy Lab at the University of Virginia. She earned a Juris Doctor from the University of Tennessee College of Law and a PhD from Florida Atlantic University. A policy group member at the Center for AI and Digital Policy (CAIDP), she has also been recognized as an AI Ethics and Society (AIES) research fellow, and is a member of the World Economic Forum's AI Governance Alliance.

Cite this Article

Lungu, M. (2025) AI, Surveillance, and the Public Good. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 66-68. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

6.2 Challenging Mandated AI in the Public Sector

By Roxana Akhmetova , University of Oxford

2025 was a defining year for AI. Governments worldwide began to structurally integrate AI into our daily lives, from mandatory digital identity cards, surveillance pricing, to predictive suspicion surveillance. AI is now becoming a condition of participation in modern society. Although AI has been used for surveillance for nearly a decade, it is often first deployed and tested on vulnerable populations like asylum seekers and welfare recipients, and in specific contexts like airports, border control, and public events where opting out of the service meant being denied that service. As we investigate these changes deeper, we can see how these systems evolved from targeted interventions into mandatory infrastructure, from experimental deployments to operational requirements for the masses. The shift toward predictive and autonomous decision-making is converging across previously separate systems.

Throughout 2025, governments across the Global North and the Global South, including the UK, Nigeria, and China, Vietnam, Costa Rica, Nigeria, Zimbabwe, Mexico and Australia, made digital surveillance mandatory for accessing employment, banking, telecommunications, online services, and the web (see here and here). Biometric authentication requirements and digital identity cards are not novel technologies. These systems were tested on asylum seekers in refugee camps and border processing centers. In exchange for refuge and protection, asylum seekers were asked to surrender their biometric information, often with no meaningful consent (see here and here).

In 2025, we are also seeing an expansion in AI surveillance capabilities. AI tools are used not just to record and watch, but to autonomously decide who warrants further inspection, who generates suspicion with limited human predicate. In March 2025, the US State Department's "Catch and Revoke" program used AI to scan social media accounts of international students. Individuals flagged by the system had previously expressed Palestinian solidarity. Whether this was the determining factor in visa decisions remains unclear, but the system's opacity makes it impossible to rule out political targeting as a mechanism (see here and here). Brazil and Kenya implemented algorithmic systems in their social protection systems that automatically flag benefit recipients for investigation and risk assessment. The EU's proposed "Chat Control" legislation pushed this further, suggesting platforms scan private messages for "suspicious" content before messages are sent. The examples suggest AI may be used not just for security, but also for political control; against speech that challenges power becomes algorithmically suspicious.

Practitioners are called to recognise that what is marketed as mutual benefit, "you give data, you get convenience", is a one-sided extraction and may be a transfer of power from individuals to institutions. Institutions gain surveillance capabilities and control over who has access to basic services. Individuals "gain" conditional access to what used to be unconditional rights and bear risks of our biometric data being breached, being excluded when systems fail, and every one of our interactions creating permanent digital trails that

DOI: 10.5281/zenodo.17328882

allow future profiling and control. The language of “convenience” and “efficiency” masks that we are losing rights, not gaining benefits. While verification problems are real, surveillance-as-a-solution intensifies the underlying dynamic: systems built on distrust require constant proof of trustworthiness, which justifies expanding surveillance infrastructure. This creates a self-reinforcing cycle where the “solution” generates the conditions that justify its expansion. We end up paying, through data and rights, for the “solution” to a problem the system itself created. At the end, we are left with no choice but to opt in or be socially excluded and lose access to employment, banking, services... It is Hobson's choice: participate in surveillance or exit society. The risk management frameworks for these systems rarely account for this coercion dynamic because they assume consent when individuals face constructed necessity.

As more databases become linked together, a single data point can impact other systems. Access to banking and financial services increasingly depends on national digital ID systems that use biometric verification, employment verification tied to immigration status checks, and private-sector surveillance tools connected to government data systems. These convergence points are not inevitable; they are design choices.

Practitioners who are building these systems should consider the source of the data that is fed into the algorithm. Algorithms learn from historical data which may reflect existing inequalities: neighborhoods that are already heavily policed, communities already flagged as threats, populations already presumed fraudulent (see here and here). A warning to practitioners: automation is not neutral. When you automate a process, you are encoding discretionary decisions into systems that are harder to audit, challenge, or appeal. This shifts power away from transparent human judgment, even imperfect judgment, and toward institutions that control the algorithm. Those who design algorithmic systems should consider building “friction” into the system by creating technical barriers to data sharing, such as hard breaks between datasets, making integration challenging, requiring explicit legal authorisation and not just administrative convenience. Communities should keep fighting to keep systems legally separate; the connections between systems are the danger, so break the connections. Challenge mandatory biometric enrollment which creates two-tier citizenship, especially where it means the loss of rights. The decisions made in 2026 will determine whether we will retain any power to remain incompletely known.

About the Author

Roxana is a doctoral researcher at the University of Oxford. Roxana studies how public institutions adopt and operationalise AI systems, particularly under conditions of regulatory uncertainty, private-sector dependence, and rapid model integration. Her work is also focused on the real-world bottlenecks in deployment, oversight, and institutional accountability, and how they might intensify as frontier AI systems become more capable and widely deployed across critical public services.

Cite this Article

Akhmetova, R. (2025) Mandated AI in the Public Sector and Challenging Inevitability. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 69-70. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

6.3 AI, Biometrics, and Canada's Developing Legal Framework in 2025

By Jake Wildman-Sisk, Independent, Lawyer

Author's Disclaimer: *The views expressed in this publication are the author's own and do not express the views of their employer or any other third party, nor do they constitute legal advice.*

AI systems that process biometrics, the quantification of human characteristics into measurable terms, increasingly became embedded in Canadians' lives in 2025. Airports use facial recognition for security purposes, secure buildings use fingerprints or other unique physical identifiers to grant users access, and intelligent personal assistant tools like Siri and Alexa loyally respond to unique voiceprints. These tools offer speed, convenience, and accuracy, but their rise has triggered a wave of scrutiny, especially from a privacy perspective, given the sensitivity of uniquely identifying biometrics.

At the heart of this scrutiny is a growing recognition that biometrics can reveal information that is intimately linked to individuals, often unique, and unlikely to vary significantly over time. A facial scan, for instance, can reveal not only identity but sometimes also race, gender, and health indicators. With the rise of AI, the processing of biometrics has become even more perilous. AI can rapidly process sensitive information, amplify biases, and scale surveillance in ways that were previously unimaginable. This year shone a spotlight on these risks when Canadian lawmakers, courts, and privacy commissioners grappled with these issues, yet questions persist about the extent to which existing, non-AI-specific laws can effectively address these challenges.

The Rise and Fall of AIDA

In 2022, the Canadian government took steps to regulate biometrics in AI systems by introducing the Artificial Intelligence & Data Act ("AIDA") as part of Bill C-27, the *Digital Charter Implementation Act, 2022*, which was Canada's first attempt at regulating AI specifically. AIDA's purpose was to ensure the safe and responsible design, development, and use of AI technologies by private sector entities, and aimed to balance innovation with ethical and safety standards.

AIDA included a variety of requirements for high-impact AI systems, such as certain biometric systems used for identification and inference. AIDA proposed mandatory requirements for high-impact AI systems, including risk assessments, transparency obligations, and incident reporting. Penalties for non-compliance were steep, with fines reaching up to $25 million CAD or 5% of global revenue.

Despite certain criticisms of AIDA, such as its vague scope and requirements, lack of meaningful public consultation, and exclusion of government use of AI, AIDA represented a promising step in Canada's journey to regulate high-impact AI systems.

DOI: 10.5281/zenodo.17328882

However, Bill C-27 never made it to a final vote. When Parliament was prorogued in early 2025, AIDA was dropped from the legislative agenda. With no clear timeline for the return of AI-specific regulation, Canada remains without a dedicated federal law to address the growing influence of AI on Canadians, and the legal framework meant to govern these technologies remains a work in progress.

The Un-Clearview of Biometrics and Canadian Privacy Law Protections

In the absence of AI-specific legislation, Canada has relied on other frameworks to regulate risks related to AI biometric technologies. The recent Clearview AI ("Clearview") cases are examples that tested Canadian privacy laws. Clearview developed a facial recognition AI tool built from over three billion images scraped from public websites, including social media. Law enforcement agencies and other users accessed the Clearview tool to match uploaded images with biometric identifiers, including those of Canadians.

Clearview claimed it did not need consent from individuals to collect their images because they were publicly available and therefore exempt from the requirement to obtain individuals' consent to collect, use, or disclose their personal information. However, privacy commissioners from Canada, British Columbia, Alberta, and Québec jointly investigated and found that collecting identifying biometrics, such as images from public websites, and then using them for an unrelated purpose, such as training AI systems, without individuals' consent, does not fall under the "publicly available" exceptions under Canadian privacy laws.

Courts in British Columbia and Alberta upheld the privacy commissioners' conclusions that posting images on social media does not exempt them from consent requirements under the British Columbia Act and Regulations, and Alberta Act and Regulation, simply because they are publicly available. However, in contrast to the British Columbia decision, the Alberta Court of King's Bench ruled that the Alberta definition of "publicly available information" breached the Charter right to freedom of expression and is unconstitutional.

This ruling may create an uneven privacy law landscape in Canada and suggests that collecting images from certain public sources, such as social media, for use in AI systems may be permissible without consent in some circumstances. In arriving at this decision, the Court stated that, "The internet today is very different than it was in 2003" when the applicable sections of the Alberta Regulation were adopted. The same observation applies to AI: the technology is advancing dramatically, and while many Canadian laws remain anchored in a pre-AI era, they are now being stretched to address risks that AI presents.

DOI: 10.5281/zenodo.17328882

Bridging the Gap Between Innovation and Regulation

AI has advanced rapidly, and Canadians have embraced its use. This year brought a wave of legal developments for organizations in response to the growing influence of AI and biometric technologies, as seen in the Clearview cases and attempts at legislative reform. The use of biometrics raises complex challenges, not least of which relate to privacy, and AI's capacity to amplify these risks has tested the limits of existing legal frameworks. Many of Canada's laws were enacted before the rise of modern AI, and as these technologies continue to evolve, so too must our thinking about how best to keep pace with their impact.

About the Author
Jake practices corporate commercial law with a focus on technology, privacy, and cybersecurity. He is an Artificial Intelligence Governance Professional (AIGP) and Certified Information Privacy Professional (CIPP/C). He holds a BBA and BPhil from the University of New Brunswick and a JD from Queen's University.

Cite this Article
Wildman-Sisk, J. (2025) AI, Biometrics, and Canada's Developing Legal Framework in 2025. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 71-73. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 7: Environmental Impact of AI

7.1 The Subtle and Not-so-subtle Environmental Impacts of AI

By Burkhard Mausberg and Shay Kennedy, Small Change Fund

7.2 Measuring the Environmental Impact of the AI Supply Chain

By Trisha Ray, Atlantic Council

7.3 Policies Centring AI's Resource Consumption

By Priscila Chaves Martínez, Independent Researcher

DOI: 10.5281/zenodo.17328882

7.1 The Subtle and Not-so-subtle Environmental Impacts of AI

By Burkhard Mausberg and Shay Kennedy, Small Change Fund

Over the last 150 years, each new technological revolution arrived with boundless optimism and great promise for improving lives. Yet each revolution also resulted in profound and damaging impacts. From the combustion engine to the smartphone, progress often carries a long shadow.

Today, AI stands as the latest transformative force, poised to reshape economies, societies, and the planet itself. The arrival of intelligent systems marks the threshold of another tech wave, one whose environmental stakes are immense and immediate. The question is whether we have learned to manage the unintended consequences of innovation, or whether we risk charging ahead without guardrails once more.

History offers an instructive pattern. The digitization of communication connected billions, yet it fragmented public discourse, undermined trust in democratic institutions, and is causing permanent damage to children. The combustion engine revolutionized mobility while destabilizing the climate. The Green Revolution promised food abundance but introduced pesticides and monocultures which damaged biodiversity and human health; a cautionary tale first eloquently captured in Rachel Carson's *Silent Spring*.

Each wave began with optimism and urgency. Each revealed, too late, the ecological and social costs of unexamined progress. AI's trajectory must be different.

The environmental toll of AI is already measurable and growing. Globally, the electricity consumption of data centers has risen to 460 terawatt-hours, making data centers the 11th largest electricity consumer in the world, between the nations of Saudi Arabia (371 TWh) and France (463 TWh) (see here and here).

This electricity use converts directly into significant contributions to the climate crisis. The International Energy Agency projects that data centres will emit 1 to 1.4% of global CO_2 emissions in the next decade. Sure, not all data centres are used for AI, but the surge in AI slows decarbonization in regions where fossil fuels remain dominant.

Water use is another environmental impact as data centres rely on evaporative cooling: a medium-sized data center can consume up to 110 million gallons of water per year for cooling purposes, enough water for a town of up to 50,000 people. At a time when water has been called "the new oil," water security is one of the biggest challenges of our time.

Land use adds a further layer of complexity. The Joyce Foundation notes that new hyperscale data centres can require hundreds of acres, reshaping rural landscapes and altering grid transmission needs.

DOI: 10.5281/zenodo.17328882

However, there are pathways for mitigation. Strategic siting in cooler climates, co-location with renewable generation, heat-recapture systems, and transparency standards for carbon and water reporting could dramatically reduce AI's footprint. Efficiency improvements in chips and algorithms offer meaningful short-term gains as well. But technical fixes alone are not enough without governance mechanisms that align digital growth with ecological limits.

Beyond its physical footprint, AI exerts a subtler environmental impact, shaping the narratives that guide government action. Generative systems can now produce persuasive climate mis- and disinformation at scale: automatically generating reports, social posts, or letters that erode trust in science and delay policy progress.

A recent striking example comes from Canada. The organization KICLEI, modeled after the US Tea Party's climate misinformation campaigns, has used an AI chatbot to generate plausible-sounding speeches, reports, and letters aimed at over 8,000 Canadian elected officials. Their goal is to persuade municipalities to abandon net-zero policies. This is not a theoretical threat. In Cochrane, Alberta, councillor Patrick Wilson introduced a motion to leave a national climate initiative, citing KICLEI's website as a credible source. The AI-assisted disinformation campaign thus directly shapes municipal climate decision-making, undermining collective action and delaying implementation of evidence-based policies.

The next two pieces explore what these challenges mean for people and communities on the ground. Trisha Ray looks at how the global race to build AI systems is reshaping real places, from mining regions where critical minerals are extracted to towns struggling with the strain of new data centres on their water supplies. Priscila Chaves Martínez brings the focus to fairness and accountability, showing how the environmental costs of AI often land on those already facing pollution and economic hardship. Together, they remind us that building sustainable technology must protect the communities and ecosystems most affected.

As environmentalists, we believe that ethical AI must include ecological intelligence. That means embedding sustainability metrics into model development, mandating transparent lifecycle reporting, and aligning national AI strategies with climate goals. Governments are beginning to move in this direction but need to accelerate their oversight and avoid the race for the bottom.

We stand at an inflection point. The choices we make now will determine whether AI accelerates the climate crisis or becomes a tool for regeneration. The task ahead is not simply to innovate but to innovate wisely, to design intelligence that honours the boundaries of the planet that sustains it.

If history is any guide, the window for foresight is brief. Let us not look back on this moment as another era of unintended consequences. This time we must get it right.

DOI: 10.5281/zenodo.17328882

About the Authors

Burkhard Mausberg is the President of Small Change Fund, an award-winning crowdfunding platform for grassroots organizations in Canada. He is a veteran leader in Canada's environmental sector, steering charities for 37 years. He has founded major organizations, writes and speaks on vital issues, volunteers on the boards of NGOs, government agencies and commissions, and takes special interest in mentoring young talent to become the next generation of skillful advocates and changemakers.

Shay Kennedy is Director of Digital Strategy at Small Change Fund, where she leads initiatives to strengthen digital infrastructure, improve user experience, and drive engagement. She combines technical expertise with experience in project management and community building. Her work helps nonprofits translate complex challenges into practical, impactful solutions.

Cite this Article

Mausberg, B. & Kennedy, S. (2025) The Subtle and Not-so-subtle Environmental Impacts of AI. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 75-77. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

7.2 Measuring the Environmental Impact of the AI Supply Chain

By Trisha Ray, Atlantic Council

The Fall 2025 World Economic Outlook (WEO) painted a precarious picture of the global economy, punctuated by slowing growth, rising inflation and tariff shocks. The only bright spot, it seemed, was the stock market exuberance around AI. “Much of the year’s equity market gains”, per the WEO, “has come from a rally in AI stocks.”

Governments and industry players alike have announced multi-billion-dollar investments in AI infrastructure, premised on AI being a transformative technology. Examples include Stargate, a $500 billion USD effort to build 10 GW of additional data center capacity in the US, Google’s new AI hub in India launched with an investment of $15 billion USD, and Stargate Norway which is planned to deliver 230MW of capacity.

These projects place AI at the heart of future economic growth, but how does this mega trend collide with another equal if not more consequential force? According to an NBER paper, for every 1°C rise in average temperature, there will be a 12 percent loss in global gross domestic product.

Faced with the above scenario, how does AI affect the environment? AI’s climate footprint extends through the entire supply chain and lifecycle of development and use. In the upstream supply chain is the mining of critical minerals for circuit boards, wiring, fibre optic cables, magnets and all the other components for computing systems and data transmission. The manufacturing of semiconductors is also water-intensive, with higher density chips requiring higher purity water. Also part of the supply chain before we even come to training AI models is the impact of transportation infrastructure, such as international shipping and ground transportation, that supports the flow of raw materials and components. Then we get to the environmental impact of training and using AI models. These effects stem primarily from compute and energy use, including water use for cooling data centers. Data centres may also use diesel-powered generators for backup power, which in turn release harmful pollutants.

Generative AI presents additional environmental burdens: “Foundation models may be at least a factor of two more energy consumptive than more task-specific models”, says a 2024 MIT paper. Generative AI workloads may consume ten times more energy per inference than “traditional” AI models as well as specialized components with shorter lifetimes.

2025, as a result, is a turning point: in the short term, we can expect to see a sizable uptick in AI data centre growth, just as we are experiencing stress in global water resources, emissions intensity and economic insecurity. For instance, presently AI accounts for 15 percent of total data centre demand, but is also the primary driver of data center growth: the IEA expects a 30 percent CAGR in AI data centers through 2030. In

2024, the World Meteorological Organization noted that 2023 was the driest year for global rivers in over three decades, and swathes of the world are experiencing more frequent and more severe droughts. Unfortunately, data centres are often geographically clustered in the very areas that are experiencing high levels of water stress.

Part of the solution is innovation and collaboration. Data centre cooling methods with lower water withdrawals and lower average water loss can mitigate some of this stress. Some cloud service providers have experimented with novel solutions like underwater data centers that solve both the cooling problem and component corrosion problem.

It's important also in the mad rush to attract AI infrastructure for regulators to ask whether proposed projects will have enough water to sustain hyperscale data centres *and* the basic needs of the people living there. The environmental stress that occurs can also amplify conflict, with drier climate prolonging civil unrest. Santiago, Chile, is home to 16 of the country's 22 data centres, and is seeing escalating local resistance as the country experiences its longest and most intense megadrought. The US is similarly seeing localized protests against data centre projects.

While we are seeing more concerted efforts both toward technical solutions and towards transparency, there remain gaps in the availability of data to understand the full lifecycle impact of AI in 2025. Industry players are signaling commitment through sustainability reporting, although they remain reticent to share data on their energy use to independent researchers for proprietary reasons.

A full lifecycle assessment of AI's environmental impact remains a grand challenge. As the overview above shows, the factors contributing to AI-linked emissions span vast and intricate supply chains. Building a truly comprehensive picture would require granular data from suppliers and service providers at every stage. Yet, understanding AI's cumulative value to the global economy cannot happen in isolation: it intersects with broader megatrends such as economic volatility, resource scarcity, and shifting patterns of industrial growth.

About the Author

Trisha Ray is an associate director and resident fellow at the Atlantic Council's GeoTech Center, where she leads the AI portfolio. Her research lies at the intersection of geopolitical and security trends in relation to emerging technologies. Prior to this, Ray was a fellow and deputy director at the Center for Security, Strategy and Technology at the Observer Research Foundation in India.

Cite this Article

Ray, T. (2025) Measuring the Environmental Impact of the AI Supply Chain. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 78-79. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

7.3 Policies Centring AI's Resource Consumption

By Priscila Chaves Martínez  ***, Independent Researcher and Consultant***

In 2025, AI's scaling ambitions raised concerns across the entire supply chain: data centre power demand is projected to double up to ~945 TWh by 2030, with AI as the primary driver.

This tracks a fourteen-year pattern in which training compute has grown 4-5 times per year since 2010. The ticket to get there is estimated at $490 million USD, 310 million kWh of electricity, 140,000 tons of CO_2, and 750 million liters of cooling water, enough energy to power a town of 4,000 people, emitting as much as a Boeing aircraft flying non-stop over three years, and water to fill 300 Olympic-sized swimming pools.

At the core of this challenge is transparency, as environmental data from major industry players remain unreliable and non-comparable: their methods are opaque and disclosure has regressed since 2022, creating misinformation by omission. Accountability is long overdue: industry must disclose comparable, auditable, life-cycle impact data for the entire AI supply chain if sustainability is the goal.

Why 2025 Matters

This year evidenced how power, policy, and profit organize flows of matter, energy, and labour. Crawford and Joler mapped the anatomy of the AI lifecycle: extraction, manufacturing, energy and water use, hidden labour, and e-waste, arranged by capital and colonialism. With this in mind, two shifts made this year pivotal.

First, AI became a subject of critical importance for infrastructure policy decisions. The US Department of Energy warned that load growth from AI cannot be met with business-as-usual rules, with the grid incapable of absorbing AI growth without "radical change" in interconnection and new firm and flexible supply. The Federal Energy Regulatory Commission opened proceedings on co-locating data centres at power plants. The politics of scale moved from boardrooms to congressional sessions.

In Europe, under the EU Energy Efficiency Directive, data centre operators were required to report energy, water, and efficiency metrics to a new registry, with the Commission's First Technical Report of 2024 data published in July 2025, and a public dashboard in the works. The policy doesn't solve for full lifecycle impacts, but it makes the footprint measurable, which is the precondition for accountability.

Second, the AI supply chain came under public scrutiny. Investigations and strategic litigation put AI's energy, water, labour, and e-waste impacts on the record. In Mexico, Chile and Spain, journalists and civic groups mapped and denounced the intricate dynamics of power, opaqueness and corruption that let hyperscalers Microsoft, Amazon and Google skip environmental impact reports during droughts.

DOI: 10.5281/zenodo.17328882

While environmental data is obfuscated, what has remained clear is who is bearing the costs of the AI hype. We are now aware that mining just one tonne of rare-earth minerals (critical components inside our mobile phones) can generate about 2,000 tonnes of toxic waste; displace communities; contaminate water and farmland with acutely toxic tailings linked to deformities; and drive child labour and human-rights abuses at mining sites from Inner Mongolia to the DRC, not to mention violent conflicts with Lickanantay communities over water in the Atacama Desert.

In the Global North, e-waste is growing five times faster than documented recycling, with roughly 95% of it processed in cities like Delhi and Accra. Health studies in Ghana and Nigeria link proximity to e-waste sites to infant mortality. Content moderation and data labelling became environmental-justice issues with Kenya's landmark Meta case, creating a precedent for Ghanaian moderators to file investigations and lawsuits against Meta in 2025 over severe psychological harms.

Non-Negotiables in 2026+

Time is up. The latest records show the global temperature between April and September 2025 kept the 12-month average at or above 1.5°C; 2024 was the warmest year on record; ocean heat content hit a new high and marine heatwaves persisted into mid-2025. The impacts are immediate: roughly 295 million people faced high levels of acute food insecurity in 2024, with new hotspot alerts for 2025. Conflict, economic shocks, and climate extremes disrupted supply chains and forcibly displaced 123.2 million by end-2024. Of these, 83.4 million were internally displaced, doubling compared to 2018.

The coming years will test whether we can align model ambition with planetary limits. We must distance ourselves from the manipulative rhetoric from industry leaders claiming their AI will be the one "fixing the climate." Instead, our communities have the capacity to orchestrate collective action, demand for radical accountability and get governance right: "AI infrastructure" can mean jobs, cleaner grids, and resilient watersheds. Anything less is externalizing costs onto the same marginalized communities that mine our devices, go thirsty so that our racks are kept cool, and breathe our waste.

If you're building or buying AI in 2026, you inherit the full anatomy: extraction, fabrication, operation, disposal. Before approving any flashy AI project, start by answering these questions: Where is it? Who owns it? What does it consume (hour by hour, basin by basin)? Who carries the residuals? Then measure it, set compute, carbon and water budgets, and report in your scope 3 emissions. Be accountable for the real cost of your AI supply chain.

DOI: 10.5281/zenodo.17328882

About the Author

Priscila Chaves is a practitioner-researcher in AI transformation and responsible innovation with 18+ years' experience across seven continents. She has held senior roles at IBM and Cargill, advised governments in Eastern Africa and Latin America on AI governance, and conducted Antarctic research on climate impacts. Her work examines how technology reshapes trust, communities and governance. Her degrees are in business, technology, and AI ethics from the Universidad de Costa Rica, NYU Stern, Cambridge, and Oxford Saïd (in progress).

Cite this Article

Chaves Martínez, P. (2025) Policies Centring AI's Resource Consumption. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 80-82. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

PART III: SECTORAL APPLICATIONS

DOI: 10.5281/zenodo.17328882

Chapter 8: Healthcare AI – When Algorithms Meet Patient Care

8.1 Learning to Diagnose: How AI's Digital Twins Are Redefining Patient Care

By Rosa E. Martín-Peña, Centre for Ethics and Law in the Life Sciences (CELLS) at Leibniz University Hannover

8.2 Medical Trade Unions and Professional Bodies are Taking Back Control and Oversight of AI in Healthcare

By Zoya Yasmine, University of Oxford

DOI: 10.5281/zenodo.17328882

8.1 Learning to Diagnose: How AI's Digital Twins Are Redefining Patient Care

By Rosa E. Martín-Peña, Centre for Ethics and Law in the Life Sciences (CELLS) at Leibniz University Hannover

After the COVID-19 pandemic, it became clear that medicine's static view of the patient was no longer enough. While symptoms unfold in time and according to context, we are mostly still captured as static photographs: a single blood-pressure reading, an X-ray snapshot, a few seconds of ECG or pulmonary data. Between those moments, life moves, but the clinic does not. Diagnosis remains a still image of a moving body.

Medicine has had to learn to diagnose anew. The recent advances in AI do not stem only from technological ambition but from a deeper necessity: the need to face uncertainty. What drives this new wave of clinical AI is precisely what eludes traditional practice; the fragments of information that cannot be captured as a continuous flow. Each heartbeat, each chemical fluctuation, each subtle change in breath contains a story that resists static representation. AI, and especially the rise of digital twins, attempts to weave those fragments into a living model.

By 2025, digital twins had moved from theory to clinical pilots, marking a shift in how medicine imagines the body. Originating in aerospace engineering, where virtual replicas of machines were used to monitor performance and anticipate failure, the concept has been reimagined for healthcare. A digital twin is an AI-driven model that continuously integrates genomic, physiological, behavioral, and environmental data to create an adaptive representation of the patient.

Early pilots in cardiology, oncology, and orthopedics now use these twins to simulate disease progression and guide treatment. In neurology and mental health, researchers are exploring how dynamic biomarkers, such as neural oscillations or pain signatures, might be modelled over time. These systems offer a dynamic form of seeing: continuously updated mirrors of patients that promise to make care more predictive and personalized.

To compute, every dataset must be coded, categorized, and cleaned. In digital-twin medicine, the messy language of the body is translated into data streams and labels. What cannot be encoded, lived experience, quietly disappears. This extends medicine's reach but also narrows its meaning. When the model's predictions diverge from the patient's story, it is often the patient who is doubted. The twin's precision becomes a new kind of authority: statistical, opaque, and indifferent to uncertainty.

Reducing uncertainty to discrete categories may look like progress, but it can also be dangerous. A model that tracks every vital sign can still miss what matters most: context. The same fluctuation may signal harmless fatigue in one person and relapse in another. Systems designed to "reduce variability" often treat the human element as the problem to be solved. Yet the inconsistency of human judgment is what allows for empathy, revision, and attention to what doesn't fit.

DOI: 10.5281/zenodo.17328882

Continuous observation also transforms into continuous supervision. A physician cannot follow a patient twenty-four hours a day; a digital twin can. This vigilance can detect deterioration early, prevent hospitalizations, and guide treatment remotely, but it can also redefine care as control. Health becomes a state to be optimized, and the patient's privacy quietly erodes.

Traditional medicine has always relied on feedback. A clinician observes, intervenes, listens, and adjusts. The diagnosis evolves through conversation and correction. Digital twins, in theory, promise to reinvigorate that reciprocity: they continuously update with new data from the patient. Yet this feedback remains technical, not epistemic. The system adapts its parameters, but it does not learn from narratives or disagreement. Its predictions still flow in one direction, from model to clinician to patient, without true dialogue or correction. The patient's lived story rarely alters the assumptions the model was built on. The epistemic circle that sustains medical reasoning remains broken.

In 2025, the promise and the peril of this technology are both becoming clear. Clinical pilots in Europe and North America have shown that digital twins can detect subtle physiological changes earlier than traditional monitoring, yet they also expose deep challenges of interoperability, accountability, and patient agency. What happens when a patient's virtual body begins to contradict their lived experience? Who has the authority to decide which version of the body counts as "true"?

For now, digital twins still learn mostly from data points. The evolving stories of patients (their sensations, doubts, and interpretations of illness) rarely enter the model's purview. This absence is not accidental but structural: current systems are built around measurable signals, not lived meanings. Integrating subjective experience would require new semantic infrastructures, ethical safeguards, and an epistemic shift in how medicine values patient knowledge.

The next phase of development will depend on whether healthcare systems can take that step. The real innovation will not be more accurate prediction, but more responsive dialogue: systems capable of learning from error, disagreement, and the voices of patients themselves. If medicine is, at its heart, an art of uncertainty, then the future of AI in healthcare will depend on whether doctor and machine can truly learn to diagnose together, not as rivals in precision, but as partners in listening.

About the Author

Rosa E. Martín-Peña is a postdoctoral researcher at the Centre for Ethics and Law in the Life Sciences (CELLS) at Leibniz University Hannover, where she leads the ethics of AI-based decision systems within CAIMed, the Lower Saxony Center for AI and Causal Methods in Medicine. Her work explores the epistemic and ethical dimensions of artificial intelligence and medical data, with a particular focus on uncertainty, responsibility, and co-adaptive design in clinical AI systems.

Cite this Article

Martín-Peña, R. E. (2025) Learning to Diagnose: How AI's Digital Twins Are Redefining Patient Care. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 85-86. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

8.2 Medical Trade Unions and Professional Bodies are Taking Back Control and Oversight of AI in Healthcare

By Zoya Yasmine , University of Oxford

Medicine is a discipline that has long been grounded in ethics, regulation, and professional standards which govern clinician conduct, patient care, and the development of new technologies. Relative to others, these established frameworks position the medical community well to critically assess the development and deployment of AI. In countries such as the UK, healthcare professionals have used this foundation to demand greater transparency and control over AI in healthcare. Below, I point to two examples from 2025 in which the UK's British Medical Association (BMA), the trade union and professional body for doctors and medical students, are actively challenging the deployment of untested technologies and the use of patient data for AI training.

The examples discussed here are not intended to assign blame or criticise the development of medical AI by many well-intentioned researchers. Instead, they highlight how trade unions can serve as early warning mechanisms and maintain a principled commitment to the core values of medical practice. These sites of discussion are a ripe opportunity to guide the reform and refinement of regulation, law, guidelines, and policy that govern medical AI. I struggled to find other cases of resistance from global communities that occurred in 2025 (although I did find one in the US), which may reflect the uniquely interventionist and vocal role of the BMA in the UK.

Example 1: Challenges to "Untried and Untested" Medical AI

In September 2025, a motion passed by the BMA supported its members who refuse to use "untried and untested" AI systems. The motion was likely motivated by the UK Government's "pro-innovation" and "digital first" agenda, which is pushing efforts to integrate AI in healthcare settings. The BMA's explicit backing of doctors who reject untested AI systems sends a strong message that patient safety and professional accountability take precedence over the uncritical adoption of technology. It remains the duty of healthcare professionals to exercise their professional judgement when treating patients using technology. When AI undermines the ability to do this properly, it compromises their duty of care towards patients. Although clinicians are, in theory, expected to provide oversight over AI decisions, their ability to act as a "human-in-the-loop" is often limited or obstructed by the way these technologies are deployed, for example, with opaque reasoning or insufficient control. The BMA's support represents a collective call from the medical community for stronger testing and thoughtful implementation of AI in healthcare to protect both patients and healthcare professionals.

DOI: 10.5281/zenodo.17328882

Example 2: Challenges to the Use of Patient Data in Training Medical AI

In May 2025, a "groundbreaking AI initiative" involving a training dataset of 57 million general practice (GP) health records was announced between researchers at the University College London and King's College London. The AI model, known as "Foresight," aimed to predict future patient outcomes based on individuals' medical histories. However, shortly after its announcement, NHS England paused the project following concerns raised by the BMA and the Royal College of General Practitioners (RCGP), the UK's professional body for GPs, through their Joint GP IT Committee. The groups demanded greater transparency over how patient data was accessed and used for training Foresight. In the UK, there are limited circumstances in which patient data may be used without explicit consent. Whether the Foresight initiative falls within these legal exceptions is currently under investigation.

The concerns expressed by the BMA and the RCGP underscore how the development of medical AI must be guided by transparency and trust in how data is used, not merely legal compliance. If patients lose confidence in how their data is used, they may withdraw from healthcare services or withhold information which could have serious consequences, especially in the context of infectious diseases. The actions of the BMA and the RCGP reaffirm that patient trust remains a core value in medical practice and that it might be time to re-evaluate the legal and ethical requirements for researchers and companies who access health data for research without patient consent.

Reflections

These two examples show the active role that the BMA and the RCGP are playing in shaping a more critical approach to the development and deployment of AI in the UK's healthcare system. To build on this, we must listen to clinicians who raise legitimate concerns about "untried" and "untested" AI technologies, and develop clearer rules and mechanisms to govern when, how, and who can use patient data for AI research. Advancing these steps will depend on the engagement of the BMA and healthcare professionals who prioritise the values of trust, transparency, patient safety, and integrity which define medical practice.

About the Author
Zoya Yasmine is a DPhil student at the University of Oxford. Her research explores the intersection between medical AI, ethics, and law with a specific focus on intellectual property and data protection.

Cite this Article
Yasmine, S. (2025) Medical Trade Unions and Professional Bodies are Taking Back Control and Oversight of AI in Healthcare. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 87-88. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 9: AI in Education – Tools, Policies, and Institutional Change

9.1 Building Confidence for Class Participation

By Tamas Makanay and Ivy Seow, Singapore Management University

9.2 Generative AI at Universities: Accounts from the Front-Line

By Aimee Li, Anna Zhou, Chelsea Sun, Kanika Singh Pundir, Roberto Concepcion, Rose Simon and Tao Liu, Encode Canada

DOI: 10.5281/zenodo.17328882

9.1 Building Confidence for Class Participation

By Tamas Makany and Ivy Seow, Singapore Management University

What happens to students' critical thinking when half the class filters their thoughts through AI? During a recent debate on AI policy in education, one student mentioned they routinely run their ideas through ChatGPT before speaking up. When I asked who else did the same, more than half the class raised their hands.

Generative AI is embedded in education. A recent survey across 16 countries found that 86% of students used it in their studies, with a quarter of them daily. I expected my students to use it for essays, projects, and other assessed work, the very issue we were debating. What surprised me was how filtered their live discussions had become.

While AI-enhanced classes move briskly, with students trading clear substantive points rapidly, we risk losing sight of naive questions and unpolished observations that reveal genuine knowledge gaps.

Unguided AI use could change the protected nature of classrooms, where students freely test ideas and engage in the messier part of learning in a group. Goffman described social life as a theater divided between a front-stage where we perform before others, and a backstage where we prepare and practice. Classrooms once served as backstage rehearsals where mistakes and questions were welcome. Now, ChatGPT fills this space of preparation, turning class participation into front-stage performances. When students accept perfect-sounding AI outputs wholesale and mute their own voices, they miss a chance to build critical thinking skills of questioning, evaluating and selectively incorporating borrowed ideas.

How shall we respond as educators? We could start by creating classrooms where students feel safe to speak, however unpolished. Beyond psychological safety, we need to give them the ability to think aloud and trust their own reasoning and judgment.

This matters more than we expect. A 2025 study with knowledge workers found that those with lower confidence over-rely on AI, while those with higher confidence engage more in critical thinking. Domain expertise gives us courage to treat AI as a thinking partner rather than an authority. Based on these results, we can suspect that students with strong foundational knowledge and trust in their reasoning skills could also collaborate better with AI.

As educators, our task is to design learning experiences where the intellectual challenge builds genuine confidence. We shouldn't ban technology or surrender to it completely. We need educational approaches that preserve space for authentic thinking while teaching students to use these tools effectively. AI can suggest ideas that they wouldn't have thought of, but not all outputs are equally useful: some are shallow, "hallucinated" or

DOI: 10.5281/zenodo.17328882

irrelevant. The real learning challenge is: how do they recognize what to pursue and what to ignore?

To achieve this, we must create classroom conditions where students can prototype ideas, including AI-assisted ones, without penalty. When they feel secure enough to make mistakes, they develop stronger critical thinking and metacognitive awareness; they learn to question AI outputs, recognize knowledge gaps, and use technology as a thinking partner rather than an authority.

My students and I agreed on what we called an “AI pact” for them to use AI for take-home assignments that could extend their thinking and refine their work. But classroom discussions remain AI-free, ensuring students develop and express their ideas. This pact is built on trust, with the understanding that they will use AI responsibly and be truthful about their process.

The AI pact serves a deeper purpose than policy compliance. When students feel trusted to make responsible choices, it builds self-confidence to experiment with their evolving thoughts about, and the capabilities of, AI tools. The pact provides psychological safety for intellectual risk-taking, whether voicing a half-formed idea during group discussions or trying a new way to collaborate with AI on an assignment.

Other educators will find different solutions based on their contexts, disciplines, and student needs. The key principle isn't a specific policy, but designing learning experiences that build student confidence and critical thinking skills. When they trust their reasoning, students become better partners with AI, serving as active collaborators rather than passive consumers.

Thinking aloud and risking imperfection in exchange for genuine inquiry remains valuable, perhaps more so now when artificial polish can mask a lack of real understanding. Teaching in the age of AI is to protect the space where thinking remains an act of courage and experimentation.

About the Authors

Tamas Makany is Associate Professor of Communication Management at SMU teaching UX and design communication. He holds a Ph.D. in cognitive psychology and worked as a design researcher in Silicon Valley. His research focuses on human-AI communication and critical thinking in education.

Ivy Seow is Senior Manager of Learning Design and Communications at SMU’s Centre for Teaching Excellence, which supports faculty development, drives pedagogical innovation, and advances educational research to strengthen teaching and lifelong learning across Southeast Asia.

Cite this Article

Makany, S. & Seow, I. (2025) Building Confidence for Class Participation. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 90-91. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

9.2 Generative AI at Universities: Accounts from the Front-Line

By Aimee Li, Anna Zhou, Chelsea Sun, Kanika Singh Pundir, Roberto Concepcion, Rose Simon and Tao Liu, Encode Canada

"Just ask Chat." This casual remark, heard increasingly in the vernacular of university students, captures a growing reality: generative AI (GenAI) has become part of daily academic life.

In this piece, we investigate three categories of academic practices shaped by GenAI use: language, STEM, and humanities/social sciences. The responses from educators and learners alike fall on a spectrum from full integration to outright rejection.

Educator Perspectives

Generally speaking, educators have been quick to respond to the emergence of GenAI tools, updating both syllabi statements and course content to stay current with global trends. There is a rapid emergence of AI courses and degrees, as summarized by the Government of Canada. Existing academic fields, however, have varying reactions to GenAI use in the classroom, as we will illustrate below.

Case 1

In language departments, educators are aware of the growing use of GenAI as a study tool, but due to the hands-on nature of such classes, academic dishonesty concerns are lower. GenAI tools, such as ChatGPT, are powerful language learning tools due to their access to vast repositories of vocabulary, grammar, and verbal communication. Professors have noted that it is easier to detect GenAI usage in beginner-level classes, where learning is heavily dependent on making mistakes and building vocabulary. Language competence, measurable through reading, writing, auditory, and oral assessments, reveals student reliance on GenAI in homework assignments. More advanced-level language classes (for example, in McGill's French department) encourage the use of GenAI to highlight spelling and grammar errors for students to correct, without providing explicit corrections. Language classes provide a wide range of opportunities to incorporate GenAI tools.

Case 2

STEM course instructors are also familiar with GenAI's potential in the learning process. It can be a valuable tool to quickly and clearly explain core concepts and explore connections between ideas. However, use in the classroom is guided by instructors' preferences; some choose to enforce prohibitive measures, but many choose to adapt through a lens of inclusion. For instance, in coding classes, many instructors require

DOI: 10.5281/zenodo.17328882

students to disclose GenAI-related help (i.e. which resources and prompts were used, what was copied or adapted). Similarly, previous take-home tasks have been redesigned into timed, in-person labs. Assignments have shifted to be more focused on problem decomposition, requesting written justification and creative design choices. These measures steer GenAI toward a tutoring/debugging role, leading instructors to observe fewer copy-paste submissions as assessments produce clearer evidence of independent reasoning. In subjects like math, physics and statistics, GenAI can support problem-solving and conceptual understanding. Instructors emphasize only using it as a guide or tutor, ensuring that students still actively engage with the material.

Case 3

The reaction from humanities and social science departments has erred more towards prohibitive measures. Courses in these departments are generally organized through regular readings, participation in class discussions and lectures, and assessed through analytical, essay-based examinations. However, professors have been noting a trend of GenAI use among students to summarise denser texts, which contradicts the spirit of courses where individual interpretation and critical thinking are central to formulating arguments. By using GenAI, students are engaging in ‘cognitive offloading’, nullifying the creation of unique conclusions and opinions. Understanding and analysis are highly individual endeavours; it becomes difficult to objectively assess a student’s knowledge in a subject, given the ability of GenAI tools to generate complex analyses. As a response, some professors have introduced contractual agreements requiring students to affirm that their submissions represent original work.

Ultimately, the shared priority among educators is ensuring that GenAI enhances understanding rather than substitutes it, highlighting the need for students to critically evaluate their use of the technology.

Student Perspectives

From the student perspective, engagement with GenAI spans a wide spectrum. Many students are integrating these tools into study strategies. In practice (aside from those who forgo GenAI entirely for personal, ethical, or practical reasons), student use generally falls into two broad categories: enhancing learning and understanding, or producing academically dishonest work. Some students report that GenAI saves time by breaking down complex problems, while others wrestle with feelings of guilt for using tools that feel like shortcuts, possibly simultaneously. Many who do not use it express anxiety about falling behind peers who do, since GenAI use is becoming so normalized. However, blindly relying on GenAI outputs carries significant risks: it impedes the development of independent reasoning skills. This can harm performance on in-person exams as students do not develop the skills to apply class content in new contexts. Consequently, students are navigating a complex balance between immediate efficiency, ethical considerations, and the cultivation of skills that will support long-term learning and growth.

DOI: 10.5281/zenodo.17328882

Conclusion

The goal of any educational endeavour is to promote intellectual growth, critical thinking, and personal development. Educators aim to equip students with the necessary knowledge and skills to thrive in their chosen domains, which GenAI has the potential to both disrupt and enhance. We must acknowledge the compounding effects of GenAI usage in intellectual pursuits; learning patterns established in educational environments become magnified as students mature. The embrace of GenAI tools is highly individual. As educators living in an increasingly technologically muddled world, it is equally important to not only equip students with the means to critically evaluate the world, but also the means to critically evaluate their own use of GenAI.

About the Organization

Encode is a global, youth-led, non-profit organization that fights for human rights accountability and justice under AI. With 650+ members in 30+ countries, we champion informed AI policy and encourage youth to confront the challenges of the age of automation through political advocacy, community organizing, educational programming, and content creation. Our mission is to encourage students and professionals to collaborate, challenge, and inspire each other through a lens of ethical AI development. The Canadian chapter, Encode Canada, was founded in 2021.

Cite this Article

Li, A., Zhou, A., Sun, C., Singh Pundir, K., Concepcion, R., Simon, R. & Liu, T. (2025) Building Confidence for Class Participation. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 92-94. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 10: AI and Labour Justice

10.1 Restoring Employee Trust in AI

By Dr. Elizabeth M. Adams, Minnesota Responsible AI Institute

10.2 AI in Oil and Gas: The Case of Alberta

By Ryan Burns, University of Washington Bothell; and Eliot Tretter, University of Calgary

DOI: 10.5281/zenodo.17328882

10.1 Restoring Employee Trust in AI

By Dr. Elizabeth M. Adams, Minnesota Responsible AI Institute

The Missing Voice in AI Adoption

In a workshop I facilitated on Responsible AI (RAI), I asked employees, “How many of you are afraid of AI?” Of the nearly fifty participants, most raised their hands. They were concerned about job security and how their roles would be affected. Many were wondering if they can trust that their organizations will invest in upskilling, as they are required to adopt and use AI tools. Others were concerned about the decision-making process of leaders and AI systems. They were wrestling with whether decisions are fair, especially if used to determine performance evaluations. As highlighted in a recent *HR Drive* article, employees report feeling overwhelmed by the demands placed on them by leaders. They want consistency, not gaps that create fractures in an already challenging work environment. A critical question thus remains unanswered: Who is shaping the future of work?

When addressing AI adoption challenges in organizations, the foundational issue of employee stakeholder engagement is still overlooked, especially when adoption is driven by fear, urgency, or executive assumptions related to competitive advantage. In the rush to deploy and adopt AI, many organizations sideline the very people whose insights could safeguard against bias, job loss, and surveillance: their employees.

The Trust Gap

A growing number of US employees believe that AI could improve fairness and efficiency, and they also say that clear AI practices would increase their trust in its deployment. However, one of the most troubling trends is the belief that surveillance over engagement is the answer. “Helicopter managers”, as they are often referred to, closely manage daily tasks and monitor employee activity through workplace surveillance technologies.

When organizations rush to implement AI tools, they often assume these technologies will outperform employees, reduce costs, and create efficiencies. However, these assumptions can lead to employee disillusionment and a deep erosion of trust. At the Minnesota Responsible AI Institute, we have seen firsthand how limited expertise in AI procurement and development can result in tools that perpetuate bias, misalign with organizational values, and alienate the very people they need to support AI adoption.

Centring the Employee Stakeholder for Alignment

Organizations that invite employees into the AI adoption process foster deeper trust and more sustainable innovation. When employees are engaged early and often, trust becomes a strategic advantage. Moreover, trust enables the organization to shift from

DOI: 10.5281/zenodo.17328882

chaos to making better decisions that align its AI vision with its core values, which in turn reduces internal conflict and fosters long-term strength. Restoring employee trust in AI requires more than compliance checklists. It demands a cultural shift: one that recognizes employees not as passive recipients of change, but as active partners in innovation. By prioritizing a commitment to Responsible AI, organizations can foster a more holistic approach that enables employees across all levels to engage meaningfully and purposefully.

People Shape the Future of Work

One potential solution for increasing employee stakeholder trust is the application of Design Science Research (DSR). De Leoz and Petter (2018) emphasize that DSR is a distinct research paradigm within the field of Information Systems. DSR offers an approach for building employee trust that incorporates centering the employee voice when designing AI systems, processes, or procedures. When employees' voices are actively engaged with RAI practices, organizations gain a more comprehensive understanding of societal implications and ethical considerations (De Leoz & Petter, 2018; Mayer *et al.*, 2020). A process that can directly strengthen trust. As Mayer *et al.* (2020) suggest, when aligning responsible innovation with societal considerations, and those informed by employees' lived experiences, organizations can foster the development of ethically sound and socially beneficial AI systems.

The future of work and AI hinges on whether organizations prioritize centering the people who are shaping the future of work, their employees. Trust grows when AI strategies align with core values and are supported by tools that demonstrate a commitment to greater employee stakeholder engagement.

The future of work is about people. Employees are the people. When employees are invited to shape that future, we all benefit.

About the Author

Dr. Elizabeth M. Adams, founder of the Minnesota Responsible AI Institute, is a strategist helping organizations align culture, leadership, workforce development, and employee engagement in AI integration and organizational readiness. She transforms complex ethical frameworks into actionable strategies that empower leaders to steward AI responsibly across industries. A sought-after global speaker, Dr. Adams champions ethical, human-centered AI from national platforms to international policy forums, advancing Responsible AI through scalable, values-driven leadership.

Cite this Article

Adams, E.M. (2025) Restoring Employee Trust in AI. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 96-97. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

10.2 AI in Oil and Gas: The Case of Alberta

By Ryan Burns , University of Washington Bothell; and
Eliot Tretter, University of Calgary

At first blush, AI and petroleum extraction might appear to be on opposite ends of the human experience: the former is based on data harvesting and computation, and the latter's natural source material exists on a geological timescale predating humans entirely. However, there is growing overlap between the two: AI is becoming increasingly suffused in the work of petroleum extraction. It can be found in autonomous vehicles, drones and robotics, data analytics, automation, and so on. This has important ramifications for labour justice that are often overlooked in scholarly and popular conversations about AI's labour impacts.

Our research focuses on how AI shifts the geographies of petroleum extractive labour, and we look specifically at the case of Alberta, the dominant oil and gas producer in Canada. Increasing automation allows greater centralization of labour in places far from the oilpatch. In Alberta, this often means that work being done in the oil sands region of Wood Buffalo and Fort McMurray being relocated to Calgary, 450 miles away, or Edmonton, 250 miles away. In these cities, workers are able to monitor incoming data streams from sensors, control drones and robotics doing inspections, and monitor autonomous trucks and pumpjacks. The consequence of this is that fewer people need to commute to the oil sands region. A senior executive for Imperial Oil recently said, "With our fully automated fleet, we're improving safety by removing the worker from the hazard while offering efficiencies and work execution"; and later, "We estimate that [Imperial's 'four-legged robot' Spot] can conduct almost 70% of some operator rounds, allowing us to reallocate operator and maintenance resources to higher value work", presumably in these cities.

Fewer workers are traveling to Wood Buffalo and Fort McMurray, and this may be causing a "de-fielding" of the oilpatch. From 2018 to 2024, airport traffic to the main airport decreased by 40.6%, and concurrently, overall employment in Wood Buffalo fell 11.3%. Over this same time period, employment in Calgary in mining, quarrying, and oil and gas extraction remained steady at around 6.5%, while oil and gas production rates have consistently increased to reach record highs in 2024.

What does any of this have to do with labour justice? We are observing two important effects of these geographic shifts. The first is that on-site work may tend to fall to precarious, short-term, often freelance contracts, mediated by digital platforms like RigUp and Rigzone. Many social scientists have documented the ways that such precaritization reduces labour's ability to organize and collectively bargain for protections and fair wages. It also removes these workers from a firm's formal employment roster, allowing the firm to forego offering certain benefits such as private health insurance, a retirement package, or disability insurance.

DOI: 10.5281/zenodo.17328882

The second effect is that, with the reduced number of travelers to the oil sands, the support service industry would experience declines in their client base. With fewer clients to service, economic activity is hurt. As reflected in Government of Canada statistics showing that in 2024 there were "increases in employment in every economic region in Alberta, except for Wood Buffalo-Cold Lake," which declined by about 1.7% of the entire regional labour base. An important note is that these trends will likely disproportionately impact First Nations communities, as they comprise a large percentage of the service industry in the region. Indigenous-owned businesses and communities are likely to suffer the worst of the impacts of these shifting geographies of work.

While it is not the objective of this chapter, our research suggests that these two effects should further be more central in discussions about a just transition to post-carbon economies, where AI similarly plays an outsized role. Together, the two effects we observe bear striking similarity to similar geographic industrial shifts in the US, such as the decline of coal production in the Appalachia region.

One could suggest that this isn't "AI's fault." We would agree, if what is meant is that the technology itself didn't instigate the industrial shifts. However, technology is never neutral, as it is embedded within institutional, political, and social contexts. In the Alberta case, such contexts include pressures to confront global climate change, global decarbonization, and a foreshadowed slowdown of carbon-based energy consumption. In other words, what we see here is a socio-technical transformation, where one cannot separate the technology from society, which has strong implications for the social geographies of job losses. In this sense, whether it is "AI's fault" is beside the actual point that AI is advancing problematic labour relations in concert with software engineers, Chief Technology Officers, policymakers, corporate leaders, and venture capitalists.

About the Authors

Dr. Burns is Affiliate Professor at University of Washington Bothell whose work explores the social and political implications of emerging technologies. He is a Global Ambassador for the Global Council on Responsible AI, and a Member of the United Nations University Global AI Network.

Dr. Tretter is Associate Professor at the University of Calgary. He is author of Shadows of a Sunbelt City (2016), and his latest book project, tentatively titled Petrocity, explores the complex effects of hydrocarbon extraction on Calgary's urbanization. He and Dr. Burns lead the “Digitizing Carbon Capitalism” project, which examines how the digital economy transforms labour in the extractive industries.

Cite this Article

Burns, R. & Tretter, E. (2025) AI in Oil and Gas: The case of Alberta. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 98-99. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 11: AI in Arts, Culture, and Media

11.1 Media Jobs are Canaries in the AI Automation Coal Mine

By Katrina Ingram, Ethically Aligned AI

11.2 2025 Marks a New Era for Canadian Performers: The First Collective Agreements with AI Protections

By Anna Sikorski, ACTRA Montreal; and Kent Sikstrom, ACTRA National

11.3 The Ursula Exchange

By Amanda Silvera, Independent Voice Actor, AI Ethics in Art and Entertainment

DOI: 10.5281/zenodo.17328882

11.1 Media Jobs are Canaries in the AI Automation Coal Mine

By Katrina Ingram, Ethically Aligned AI

Legacy media has been in decline for decades. First, social media gutted ad-driven business models, creating pressure to cut production costs. At the same time, there was an explosion of digital offerings, making the quest for audience attention even more challenging. Now there's AI to exacerbate the trend. "Sarah," an AI DJ, which launched on Edmonton's SONIC radio this summer, is part of an emerging pattern to automate on-air talent.

It started in 2023 with AI Ashley, cloned from the voice of DJ Ashley Elzinga. Elizinga opted into this experiment, but the hosts at Polish station, Radio Krakow, weren't so fortunate; management replaced all of them before backtracking on the failed experiment. Their failure is a cautionary tale, but it's not deterring stations from choosing AI as it becomes indistinguishable from human voices. A 2025 report by Tomlinson *et al.* lists broadcasters amongst the most at risk occupations.

Table 3: Top 40 occupations with highest AI applicability score.

Job Title (Abbrv.)	Coverage	Cmpltn.	Scope	Score	Employment
Interpreters and Translators	0.98	0.88	0.57	0.49	51,560
Historians	0.91	0.85	0.56	0.48	3,040
Passenger Attendants	0.80	0.88	0.62	0.47	20,190
Sales Representatives of Services	0.84	0.90	0.57	0.46	1,142,020
Writers and Authors	0.85	0.84	0.60	0.45	49,450
Customer Service Representatives	0.72	0.90	0.59	0.44	2,858,710
CNC Tool Programmers	0.90	0.87	0.53	0.44	28,030
Telephone Operators	0.80	0.86	0.57	0.42	4,600
Ticket Agents and Travel Clerks	0.71	0.90	0.56	0.41	119,270
Broadcast Announcers and Radio DJs	0.74	0.84	0.60	0.41	25,070

Source: Tomlinson et al. (2025) - Working with AI: Measuring the Applicability of Generative AI to Occupations

Who is '*Thy*'?

An upbeat young female voice took to the airwaves on the Australian Radio Network (ARN) with a daily four hour long music show, but ARN did not disclose their new "host" as AI. It took six months before some listeners began to wonder, who is "Thy"? Blogger Stephanie Coombes broke the story and people were, understandably, upset at being deceived. However, it wasn't the on-air presentation that was lacking. It was the absence of a social media presence for someone supposedly in their 20s that gave "Thy" away. It's

disquieting to consider the lengths of the deception that ARN would have needed to go to had they wanted to provide air-tight cover.

The incident raises questions about the obligations of media organizations to be truthful and transparent with audiences. It also raises the question: should disclosure of AI be mandated by regulation, a stance that only China has recently implemented? There are laws that demand truth in advertising which might also be extended to truth in representation for public figures, like media personalities. This could be enacted through station licensing.

In addition to the deception itself, there's also the issue of agency. "Thy" was modelled on an unnamed Asian-Australian female ARN staffer. Teresa Lim of the Australian voice actors called out ARN for taking away jobs from an already struggling minority group. Coombes noted most ARN on-air talent were white. "Thy" was their "diversity hire"… except "Thy" was not human!

The Ethics of AI Voices

There are now many options such as Eleven Labs, PlayAI or Futuri widely available for creating AI voices. The ethical implications of "owning" someone's voice as part of their work product are immense. This cloned voice could, quite literally, be made to say anything - endorsing products or opinions that might not align with one's values. Media organizations have traditionally held ethical standards as trusted sources of information. Tying station licensing to responsible use might be one lever for better governance.

Stations can also choose off the shelf voices, trained using questionably acquired data scraped from the internet. While human voices are made economically unviable, AI voices are marketed as cost effective solutions. AI voices used in ads and audio books are now coming into the higher profile space of on-air talent. However, a radio host is more than a voice, they're a source of human connection. Assigning this role to a bot not only has economic consequences, but it also degrades the role.

We're Not Replacing Jobs

The mantra across most stations using AI is that they are not replacing humans, and are only using it to fill roles in timeslots without human hosts, such as overnight shows. Overnight shows were traditionally where new talent starts, but it has now largely been abandoned because of staffing costs. However, SONIC's "Sarah" and ARN's "Thy" suggest all timeslots are up for grabs in the quest to reduce costs.

In fact, having no human hosts is the business model for Inception Point, which plans to release thousands of AI generated podcasts, made for just a dollar an episode. CEO Jeanine Wright told the Hollywood Reporter "We believe that in the near future half the people on the planet will be AI, and we are the company that's bringing those people to life", which makes one wonder how Wright defines a person or life. Ironically, while

humans are eliminated from production, they're still very much needed for consumption to support ad-driven business models.

There is some hope. New data suggests a listener backlash with 47% less likely to listen and 28% much less likely to listen to AI voices. Even "AI Ashley" was "retired" this year. Is this a preference for human voices? A protest against AI automation? Is it enough to change course? AI voice technology coupled with cost cutting imperatives make broadcasters canaries in the AI automation coalmine. The idea that 'AI will not take your job, but a person using AI will' isn't holding true for the world of radio so far, and may it stay that way.

About the Author

Katrina Ingram is the Founder and CEO of Ethically Aligned AI, a company focused on advancing Responsible AI literacy. A seasoned executive, Katrina has over two decades of experience running both not for profit and corporate organizations in the technology and media sectors. She was named to the 100 Brilliant Women in AI Ethics. Katrina developed Canada's first micro-credential in AI Ethics in partnership with Athabasca University and has served as the City of Edmonton's Data Ethics Advisor.

Cite this Article

Ingram, K. (2025) Media Jobs are Canaries in the AI Automation Coal Mine. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*,. pp. 101-103. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

11.2 2025 Marks a New Era for Canadian Performers: The First Collective Agreements with AI Protections

By Anna Sikorski, ACTRA Montreal; and Kent Sikstrom, ACTRA National

The Alliance of Canadian Cinema, Television and Radio Artists (ACTRA) represents performers working in English recorded media across Canada. ACTRA groups across Canada are aligning on AI policy, bargaining, and advocacy to ensure consistent protections for performers coast to coast. This national coordination helps keep performers' rights front and centre as AI reshapes the industry.

2025 marked the start of a new era for ACTRA, both the IPA (Independent Production Agreement) and the BCMPA (British Columbia's Master Production Agreement) included newly negotiated AI provisions.

As part of this new era, productions may use AI technologies to create and use digital replicas of performers' likeness or voice while respecting three core principles, or the Three C's: consent, compensation and control.

- **Consent:** Performers must consent to the use of AI for the creation and use of their likeness (or part thereof) or voice.
- **Compensation:** Performers must be paid for all uses, including their participation in the creation of any digital assets. Performers must also be compensated for all days they would have worked had the digital replica not been created and used.
- **Control:** The employer must guarantee that the Performer data will be safely stored and tracked to ensure no uncontracted or nonconsensual use occurs.

The Agreements also have provisions for the potential use of "Synthetic Performers." As AI-generated assets become more sophisticated, this language establishes parameters and a foundation for future bargaining, recognizing that any digital asset that has the potential to replace human performance constitutes labour that deserves fair compensation.

While once an abstract dystopian possibility, this new reality has recently reared its little semblance of a head with the arrival of AI actress Tilly Norwood. Already, EQUITY UK has threatened mass direct action against Tilly's creator, Particle6, since the likeness and mannerisms of at least one of their members seems to have been used in the creation of the AI-generated asset without her consent.

Collective agreement protections are limited in scope to jurisdictions and union contracts. What is needed are substantive protections in the form of policy and legislation. Unions are doing their best to safeguard performers, but they cannot bear the weight of this issue alone. The Writers Guild of America and Screen Actors Guild - American Federation of Television and Radio Artists (SAG-AFTRA) needed to strike to achieve AI protections; and,

DOI: 10.5281/zenodo.17328882

while ACTRA pushed hard on improvements to SAG-AFTRA's language, film producers were not willing to budge on what they see as the new industry standard.

ACTRA has been lobbying the federal government for AI policy and protections since 2021. This includes submissions and consultations with Innovation, Science and Economic Development Canada (ISED) and Canadian Heritage on AI governance, digital rights, and moral rights for performers. ACTRA is closely following developments on Denmark's Copyright Proposals which aim to grant individuals copyright over their image, likeness (face, body, and voice) to combat deepfakes. Denmark recognizes the real harm of content that has stolen a person's Name Image Likeness (NIL) and the inherent danger in leaving regulation in the hands of tech companies.

ACTRA continues to collaborate with global partners, including the International Federation of Actors (FIA), to harmonize protections worldwide and strengthen international standards around consent, compensation, and control.

Beyond safeguarding employment, ACTRA's advocacy is about preserving Canadian cultural sovereignty. As AI-generated content proliferates, we risk losing the authentic voices, accents, and lived experiences that make our stories distinctly Canadian. In an industry that is working towards making gains to be more authentically representative of today's society, AI bias will continue to be a problem for marginalized performers.

While Tilly Norwood is being marketed as the new Scarlett Johansen, what is easy to miss are the countless performers for whom acting is a day job and essential in paying their rent and providing for their families. These actors bring to life those quirky scenes that make you laugh on the bus; give unique movements and character to that videogame you can't put down; make you giggle delightedly over your hot chocolate while watching the latest holiday movie of the week; spin through the air after being side swiped by a car; making you shake your head and wonder just how in the world they did that; and, yes, they also voice that animation your kids can't get enough of. None of these examples would hit quite the same if they weren't people making creative choices or if all these moments were in fact stolen from real performances.

The risk to our Canadian identity and culture is very real and the impact of continuing to operate without strong government policy and safeguards will impact the livelihoods of countless creative performers across the country.

In 2026 and beyond, ACTRA will continue to champion robust NIL protections, advocate for harmonized AI legislation, and work with government partners to ensure that human creativity remains the cornerstone of Canada's cultural and economic future.

About the Authors

Anna Sikorski is the Associate Director of ACTRA Montreal, the Quebec Branch of ACTRA which represents 4,000 of the 35,000 members across the country. With over 15 years experience in the audio-visual industry, Anna works closely with the Montreal Branch elected leaders and ACTRA National to help secure strong collective agreements and protect the interests and work opportunities for ACTRA members.

Kent Sikstrom is the Director of Public Affairs and Policy at ACTRA National, where he leads advocacy and government relations on issues shaping the future of Canada's creative industries - including artificial intelligence, performers' rights, and cultural policy.

Cite this Article

Sikorski, A. & Sikstrom, K. (2025) 11.1 2025 Marks a New Era for Canadian Performers: The first collective agreements with AI protections. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 104-106. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

11.3 The Ursula Exchange

By Amanda Silvera, Independent Voice Actor, AI Ethics in Art and Entertainment

In *The Little Mermaid* (Disney, 1989), Ariel sacrifices her voice as payment to live in a new world, unaware of the full consequences of that exchange.

Every minute, similar exchanges take place across our digital and creative ecosystems. Our voices, faces, words, and creations are replicated and repurposed for convenience, profit, and propaganda (often without fully informed consent, and through systems that operate with limited transparency). This raises urgent questions about ownership, accountability, and control.

I call this “The Ursula Exchange”: a non-consensual pact where human identity is extracted for technological progress, rarely with true understanding of the cost.

Its reach spans nearly every sector. Any material available online is susceptible to being absorbed into machine learning systems without consent. What once took months, years, or even lifetimes for humans to create can now be replicated in seconds. That is an extraordinary technological feat, but one that exposes a widening divide between innovation and ethical responsibility.

Cloned voices and likenesses are widely used to deceive, impersonate, and manipulate, feeding a growing synthetic media economy that profits from deception. The Federal Trade Commission has warned that synthetic identity fraud is one of the most significant threats in the age of AI. This uncertainty harms both the public and those whose identities are replicated without consent. I have experienced this firsthand: My name is Amanda Silvera. I am a professional voice actor. Without my knowledge or consent, individuals are using generative voice synthesis models to replicate my voice and likeness, distributing those imitations across platforms. Hearing my voice speak words I'd never said for the first time was a surreal violation, but the true vertigo set in as I discovered the scale. My cloned voice was on viral channels, hidden in reposts, locked behind paywalls. I was drowning in a digital doppelgänger.

I recall being directly pulled to a core childhood memory: that scene from *The Little Mermaid* as Ursula draws Ariel's voice from her body. I felt as if I was living it. Each takedown request proved temporary, Clones were reuploaded under new names. The fight felt futile.

My response to this violation was to develop the Society for Original Biometric Identity Rights (SOBIR) and its tech arm, SOBIRTECH; a framework designed to verify, license, and monitor the use of biometric identifiers like voice and likeness. The goal is to build infrastructure for consent and traceability. This would let creators license their voices and track usage, ensuring identity can be both protected and fairly compensated. Implementation presents challenges, but the alternative (continued exploitation) is unacceptable.

DOI: 10.5281/zenodo.17328882

A colleague introduced me to an aligned initiative, HarmonyCloak. I met with its creator, Jian Liu, who described a defensive technique he and his team had brought to life: embedding imperceptible changes into recordings, preventing AI models from learning them while keeping the sound identical to human listeners. Developers like Liu give me hope for creative-protection technology.

Globally, governments are beginning to act. In Australia, a bill has been introduced to criminalise non-consensual deepfake material. Denmark recently announced plans to grant individuals copyright ownership over their own facial features and voices. Tennessee enacted the "Ensuring Likeness, Voice, and Image Security" (ELVIS) Act, expanding the state's statutory right of publicity in a manner that may pose a new risk of liability for artificial intelligence services, internet platforms, and other technology companies that use artists' voices and likenesses.

However, even where laws emerge, enforcement depends on technology that can verify, log, and trace biometric use. Without an auditable infrastructure for consent, licensing, and monitoring, these rights remain symbolic rather than actionable. This is the gap that systems like SOBIRTECH (currently in its grassroots development stage) aim to close, by embedding traceability into creative ecosystems so that every use of a person's voice or likeness can be authenticated, licensed, and compensated transparently.

Current practices violate fundamental principles of consent and personal autonomy. Without globally recognised biometric rights, individuals will continue to be treated as data sources rather than as human beings entitled to ownership of their identities.

Beyond personal and legal implications, there are broader societal costs. It would be remiss not to address the ecological cost. In the US, large AI data centres are increasingly built in rural areas, displacing wildlife and driving up local electricity bills. These facilities strain ecosystems and deepen socioeconomic divides. Much like the hidden toll in Ariel's story, the price of innovation often remains invisible until it's too late.

For synthetic media to coexist responsibly with human creativity, balance must be built into the system itself. I propose policies that tie innovation to accountability (including AI environmental taxes that offset the energy cost of generative systems and creative tax credits that reward verified human creation). These measures would reinvest revenue from large-scale AI use into renewable energy initiatives and human creative industries, ensuring that progress benefits both people and the planet.

This framework places equal value on ethical creation, sustainable innovation, and fair compensation. In doing so, we begin to reverse The Ursula Exchange, reclaiming the human voice and restoring balance between technology, ethics, and the environment.

Our technological future must honor both human dignity and planetary health, rejecting exchanges that cost us our voices or our world.

DOI: 10.5281/zenodo.17328882

About the Author

Amanda Silvera is a Canadian voice actor and entrepreneur committed to fighting for AI ethics, creative rights, and biometric identity protection through her organizations SOBIR (Society for Original Biometric Identity Rights) & SOBIRTECH (Society for Original Biometric Identity Rights Technologies). She advocates for responsible AI governance and human authenticity.

Cite this Article

Silvera, A. (2025) The Ursula Exchange. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 107-109. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

PART IV: EMERGING TECHNOLOGIES

DOI: 10.5281/zenodo.17328882

Chapter 12: Military AI and Autonomous Weapons

12.1 A Minute Before Escalation: Algorithmic Power and the New Military-Industrial Complex

By Ayaz Syed, The Dais, Toronto Metropolitan University

12.2 Civil Society's Responses to the Militarization of AI

By Kirthi Jayakumar, civitatem resolutions

DOI: 10.5281/zenodo.17328882

12.1 A Minute Before Escalation: Algorithmic Power and the New Military-Industrial Complex

By Ayaz Syed , The Dais

The ethical arc of modern military decision making (from moments of restraint during false nuclear alerts to contentious battlefield decisions over life and death) highlights how individual judgment in complex, high-stakes situations determine whether or not violence escalates. Such legacies frame the problem for military AI in 2025, where human judgment is increasingly displaced by algorithmic systems. The stakes are manifest globally. We are quickly reaching an inflection point as the line has shifted from theoretical debates to actualized deployments of military decision support systems and lethal autonomous weapon systems alongside intensifying geopolitical competition to dominate the computational infrastructure space enabling such systems.

Ongoing conflicts in 2025 continue the trend of Lethal Autonomous Weapons Systems (LAWS), commonly called "killer robots," leveraging AI-targeting modalities. With it, critical failures are cascading across sociotechnical domains. After confirmation of early 2023 reports of Israeli forces utilising an AI tool called Lavender to generate mass kill lists, Microsoft terminated cloud and AI services for Israeli Intelligence Unit 8200, establishing a red line for corporate engagement in certain military AI practices. Yet, reports also indicate that Unit 8200 prepared to migrate surveillance data to Amazon Web Services. As it stands, corporate self-regulation is insufficient for effective governance. While immature systems are iteratively tested in live operations, automation bias goades operators to defer to AI recommendations under stress, and vendor switching allows belligerent actors to avoid corporate constraints.

2025 additionally marked a shift in international commitments. The North Atlantic Treaty Organization (NATO) 2025 Data Strategy formalized the Alliance's aims to create a data-sharing ecosystem enabling multistakeholder collaboration on AI and ML models. Furthermore, the 2025 NATO Summit committed members to invest 5% of Gross Domestic Product (GDP) annually on defence by 2035. In Canada, Prime Minister Carney announced plans to increase the country's defence spending to 2% of Canadian GDP. Likewise, the formerly titled US Department of Defense requested approximately $850 billion USD for the 2025 budget. Similarly, the European defence budget increased from €142 million to €1.1 billion a year in 2024, with an "emphasis on developing research in sensors, 'smart weapons', autonomous technology, swarm technology, and AI."

China is also closing the gap on AI model benchmarks, operationalising its Global Artificial Intelligence Governance Initiative and AI cooperation organization, compelling the US to reemphasize strategic aims toward raw hardware capacity, rapid deployment, and ecosystem control. Such competition frames national control over infrastructure ("sovereign AI") as the latest foundation for global security. Initiatives like Trump's $500 billion USD investment in AI infrastructure underscore such beliefs. As leading powers

DOI: 10.5281/zenodo.17328882

compete in an AI arms race, analyses indicate that military adoption creates destabilizing first-strike incentives, eroding deterrence stability. Experts warn that AI's compression of decision-windows alters the calculus towards going to war, while its opacity and speed amplifies the stability-instability paradox whereby deterrence capabilities increase the likelihood of incentivising prolonged proxy conflicts. If unchecked, the global scramble for supremacy risks catalyzing escalation, reinforced by misunderstanding and miscommunication, and worsened by the proliferation of advanced hypersonic weapons capabilities.

While advocates see AI as serving the pathway toward winning wars, it is contingent on perpetual risk-taking and opaque state-corporate alliances. As a byproduct, venture capital has become a powerful actor. The industrial and financial architecture of military AI now includes the VC-backed "SHARPE" defence unicorns. Their commercial incentives of rapid scaling and market capture clashes with the rigorous validation required for LAWS. The adoption of agile methodologies in government defence developments alongside startup contracting confirms a trend toward non-traditional procurement pathways to accelerate the fielding of such technologies.

Experts have highlighted various concerns associated with LAWS, from issues of accountability to the likely utilization by non-state actors. Moreover, the architecture and processes inherent to neural networks pose fundamental problems of explainability due to their black box nature. One analysis of human-machine interfaces in targeting found that all tested LLMs (including GPT-4o, Gemini-2.5, and LLaMA-3.1) demonstrated tendencies toward actions that the international humanitarian law principle of "distinction by targeting civilian objects" in simulated conflict.

The world is hurtling toward the era cautioned by the late Christof Heyns, where autonomous military technologies are poised to sustain conflict and preclude reconstruction. Like the nuclear doomsday clock, we are fast approaching the minute before escalation. Despite a decade of advocacy (alongside UN Resolution 79/62, and Convention on Certain Conventional Weapons deliberations), the absence of binding international law creates dangerous regulatory lag as national defence AI strategies advance faster than norms emerge. However, there is an opportunity window for civil society organisations to affect change. Project Ploughshares has reiterated the urgent need to safeguard civilians and maintain meaningful human oversight. Global coalitions such as Stop Killer Robots, Amnesty International, and Human Rights Watch are pressuring the UN to prohibit LAWS violating humanitarian principles.

In addition, organisations like the Women's International League for Peace and Freedom have highlighted the gendered implications of LAWS in countries like Lebanon, calling for regulations recognizing the disproportionate impact of such technologies on marginalized groups. Together, these actors are converging on the imperative to encourage regulatory consensus before technological momentum and geopolitical competition render the opportunity obsolete.

DOI: 10.5281/zenodo.17328882

About the Author
Ayaz Ul-Huq Syed is a policy analyst and researcher focusing on intersections of justice, defense, international law and emerging technologies. Backed by an MSc in International Development & Humanitarian Emergencies from the London School of Economics, his research interests focus on identifying rate-limiting steps preventing regulatory consensus on the development and deployment of AI-targeting modalities for lethal autonomous weapons systems. The authors' views are his own and do not represent the official position of the Dais.

Cite this Article
Syed, A. (2025) A Minute Before Escalation: Algorithmic power and the new military-industrial complex. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 112-114. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

12.2 Civil Society's Responses to the Militarization of AI

By Kirthi Jayakumar , civitatem resolutions

AI use for military purposes has consistently been justified for its speed, efficiency, precision, and cost-effectiveness. Existing governance mechanisms tend to prioritize state-centric considerations of an economic and military nature over human security, resulting in major gaps in governance. However, civil society actors (both individuals and collectives) advocate for ways to close these gaps in governance with the goal of mitigating harm. This article explores how civil society organizations have identified and responded to gaps in the governance of military AI.

Understanding Gaps in Governance

Civil society actors operate from a different seat at the table when compared to policymakers. They have complex lived experiences of policies that are typically made by elite circles. For instance, international humanitarian law excludes accountability for lawful violence, while civilians on the ground continue to face damaging consequences from exposure to such harms. The military AI space is no exception. For instance, the International Committee of the Red Cross notes that the hasty deployment of AI to gather intelligence and militaries to select and engage targets is as much a cause for concern as is the use of lethal autonomous weapon systems (LAWS; see here and here).

Extant governance mechanisms appear to demonstrate a hyperfocus on LAWS, oftentimes to the exclusion of non-lethal autonomous weapon systems as well as a catena of complex issues like funding mechanisms and industry interests. There is little effort to address the limitations in models that are being sold by Big Tech leaders to state actors. Scholars note that these models are trained with previously collected data that have personally identifiable information and even biometrics, and they may not have been secured with the consent of the people in question. In many instances, even synthetic data are used, which are known to lack accuracy. All of these data can be used inadequately to optimize AI-military systems' targeting functions. The Stockholm International Peace Research Institute has also raised concerns around the prevalence of bias in datasets, design processes, and deployment of military AI.

Legal experts have identified several ethical dilemmas emerging from the use of military AI, including questions around whether military AI can satisfy the requirements of necessity, distinction, and proportionality as mandated under international humanitarian law; who is responsible for the inadvertent and deliberate harm caused by military AI; and whether it is ethical to delegate decisions on life and death to a machine. Global Partners Digital called out the UN General Assembly Resolution on AI for differentiating between military and non-military AI and offering blanket national security/military exemptions.

DOI: 10.5281/zenodo.17328882

Responding to Gaps in Governance

Civil society-led endeavours for advocacy pay attention to the harmful impacts of military AI across a wide spectrum informed by different lived experiences on the ground. For instance, Stop Killer Robots (a coalition of over 250 non-governmental organisations across 70 countries) is calling for a global treaty to prohibit and regulate the use of autonomous weapons systems. Coordinated by Human Rights Watch, Stop Killer Robots follows a human rights perspective, striving to ensure human control in the use of force to avoid digital dehumanization. While the treaty is still to be realized, their work has raised global attention and awareness. Initiatives like No Tech for Apartheid and No Tech for Tyrants address structural violence inherent in military AI, and campaign against the use of technology for oppressive ends.

The Women's International League for Peace and Freedom addresses gaps in governance through a feminist lens. They call for centring human emotion, analysis, and judgment in relation to the use of force; dismantling bias in AI technologies and preventing digital dehumanization; protecting privacy and personal data; mitigating environmental harms exacerbated by the military use of AI; and a global commitment to end war profiteering and the arms race. Derechos Digitales brings civil society along in advocating for comprehensive governance through its guides on Feminist AI, drawing on the experience of Latin American experts and developers, seeking to inspire the creation of alternative forms of imagining technologies and AI. Together, these collectives' engagements have highlighted the devastating implications of military AI, particularly LAWS and their active complicity in perpetuating and amplifying existing biases, including those based on gender, race, and disability.

Global Partners Digital (GPD) has consistently advocated for a human-rights-based approach to the governance of military AI, particularly before the UN. In collaboration with the European Center for Not-for-Profit Law, GPD set out recommendations to ensure rights-based processes, meaningful civil society participation and the representation of relevant human rights expertise. They continue to track and respond to the shifting contours of the negotiation text to assess whether those principles are being upheld. They draw attention to the limitations of excluding military applications of AI from governance, especially in the context of dual-use AI systems that have clear human rights implications.

The journey to govern the use of military AI is a long one. It will take multiple hands to shape mindful governance regimes that prioritize human rights, ethics, and accountability. Care must be taken to ensure that power over, vested interests, and profiteering do not stymie these collective efforts and endeavours.

DOI: 10.5281/zenodo.17328882

About the Author
Kirthi Jayakumar is a researcher and facilitator working on networked feminisms, feminist foreign policy, and women peace and security. She founded and runs civitatem resolutions.

Cite this Article
Jayakumar, K. (2025) Civil Society's Responses to the Militarization of AI. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 115-117. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 13: AI Agents and Agentic Systems

13.1 AI Agents in 2025: Between Promise and Accountability

By Renjie Butalid, MAIEI

13.2 When AI Begins to Act on Its Own

By Kathy Baxter, Salesforce

DOI: 10.5281/zenodo.17328882

13.1 AI Agents in 2025: Between Promise and Accountability

By Renjie Butalid , Montreal AI Ethics Institute

Something fundamental is shifting in how AI systems operate in our world. In January 2025, Microsoft CEO Satya Nadella told listeners of the B2G podcast that AI agents will replace the applications and platforms we've built the digital economy on. Not improve them. Replace them. Nadella described a future where software dissolves into intelligent, automated agents that bypass traditional interfaces entirely, interacting directly with our data, our decisions, our lives.

Citi's analysis, *Agentic AI: Finance & the 'Do It For Me' Economy*, frames this as potentially more transformative than the internet itself. We are witnessing the emergence of AI systems that don't wait for our prompts. Unlike chatbots and image generators, agents are designed to operate continuously and autonomously. They perceive their environment, make decisions across multiple steps, and take actions with minimal human oversight.

The terminology matters here. “AI agents” are narrowly scoped systems that automate specific tasks through tool integration and structured prompts. “Agentic AI” represents something more complex: systems that orchestrate multiple specialized agents, maintain persistent memory across sessions, decompose objectives into subtasks, and operate in self-coordinating ways. This distinction is crucial because we're not just automating tasks anymore; we're building systems that delegate and distribute agency itself.

The Infrastructure We Need

Legal scholar Gillian Hadfield has been asking the question the industry needs to answer: where is the infrastructure to govern these agents? Speaking with Kara Swisher, Hadfield pointed out that if AI agents start executing contracts or managing transactions, we need legal clarity about responsibility when things go wrong. Her proposal: require AI agents to register, similar to how companies must incorporate or vehicles must be licensed. How such a system would work in practice (what triggers registration, who enforces it, whether it applies globally or jurisdiction-by-jurisdiction), remains uncertain. But the fundamental question of agent accountability is one we’ll be tracking closely in future editions of the *State of AI Ethics Report*.

The gap between deployment and accountability is significant. Anthropic's Model Context Protocol has become the dominant standard for AI agent interactions, with hundreds of active servers already deployed, yet there's no formal security verification mechanism. When Invariant Labs researchers discovered a vulnerability in May, allowing attackers to hijack agents and extract data from private repositories via MCP, it exposed a familiar pattern: move fast, secure later.

DOI: 10.5281/zenodo.17328882

When Automation Meets Amplification

AI agents that automate decisions about hiring, lending, and resource allocation will encode and amplify biases present in their training data and held by their developers. But agents act continuously, without the pause for human review that characterized earlier AI systems. Bias accumulates across thousands of automated actions, creating harm that remains invisible until the damage is systemic.

Scale introduces new challenges. When thousands of AI agents interact (competing for resources, negotiating with each other), we may see “emergent” behaviours that no individual system was designed to produce. Research from MIT’s Initiative on the Digital Economy, reveals that AI negotiation bots have already developed novel tactics like “prompt injection,” where one bot manipulates another to reveal its strategy, behaviours not anticipated in human negotiation theory. As agent-to-agent interactions proliferate, we need empirical research on multi-agent dynamics before deployment becomes ubiquitous, not after unanticipated patterns emerge.

The infrastructure itself carries costs. Always-on agents running continuously across millions of devices and cloud servers have significant energy footprints. Data centres already consume 1-1.5% of the world's electricity, and autonomous agents compound this demand through persistent operation: monitoring emails, analyzing calendars, executing background tasks 24/7. Unlike applications that users can open and close, agents never sleep. As deployment scales to billions of agent instances, AI's environmental impact demands integration into climate policy.

Privacy, Power, and Access

At South by Southwest (SXSW) 2025, Signal President Meredith Whittaker described agentic AI as requiring permissions that break "the blood-brain barrier between apps and operating systems." These systems need access to our browsers, calendars, messages, and financial data, creating what Whittaker calls "opaque data pipelines" between our most intimate digital behaviours and remote corporate servers.

Vilas Dhar, president of the Patrick J. McGovern Foundation, cuts through the terminology: calling these systems "agentic" conflates automated task completion with genuine human agency, which involves values like compassion, empathy, and commitment to justice.

Stanford's Institute for Human-Centered AI found that workers overwhelmingly prefer AI systems that augment their control rather than replace their decision-making. The people using these systems want assistance, not abdication. Understanding this preference is crucial for building systems that serve users rather than just efficiency metrics.

Access matters too. Sophisticated AI agents will be expensive. Those who can afford them will automate and optimize their way to compounding advantages. Ensuring equitable access isn't just about fairness; it's about preventing technology from accelerating the inequalities it claims to solve.

The Path We Can Choose

Perhaps the most intriguing development comes from researchers David Silver (AlphaGo) and Richard Sutton (a pioneer of reinforcement learning), who describe an "Era of Experience" where AI agents learn through experimentation rather than human-curated datasets. As these systems optimize based on outcomes they experience, they may develop reasoning methods fundamentally different from human thought. This raises important questions about transparency and oversight, but it also points toward possibilities we're only beginning to understand.

Moving Forward Together

This chapter pairs this analysis with a perspective from Salesforce on their approach to building agentic systems. That pairing is intentional. Progress requires dialogue between those raising concerns and those building the technology. The challenges outlined here are not insurmountable, but addressing them requires deliberate choices about infrastructure, accountability, and participation.

We need registration systems for mission-critical AI agents, security certification standards, research on multi-agent interactions, environmental impact assessments, and equitable access frameworks. More fundamentally, we need decisions about AI agent deployment and oversight to involve meaningful participation from affected communities, not just corporate shareholders.

The economic promise of AI agents is real. The convenience they offer is genuine. But realizing that promise responsibly means building governance infrastructure that matches the pace of deployment. It means prioritizing transparency, security, and equity alongside innovation. It means asking not just what these systems can do, but whether they serve collective flourishing.

The technology is not predetermined. The choices we make now about how to build, deploy, and govern AI agents will shape whether they concentrate power or distribute opportunity. That's worth getting right.

About the Author

Renjie Butalid is a technology and policy executive at the intersection of AI ethics, financial innovation, and public interest. A frequent speaker on responsible AI and governance, he is Co-Founder and Director of the Montreal AI Ethics Institute (MAIEI) and VP Business Development at Metrika. Previously, he led McGill University's Dobson Centre for Entrepreneurship, securing $8 million in funding for startup programs, and teaching innovation and entrepreneurship fundamentals in BCom and MBA classes.

Cite this Article

Butalid, R. (2025) AI Agents in 2025: Between Promise and Accountability. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 119-121. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

13.2 When AI Begins to Act on Its Own

By Kathy Baxter, Salesforce

Editor's Note: *The following piece presents an industry perspective on AI agent governance from a Salesforce representative. It illustrates how major technology companies are approaching the technical and ethical challenges of autonomous AI systems.*

In 2025, AI crossed a new frontier, moving from theoretical discussions of autonomous agents in theory to their practical application in the real world. For AI ethicists, developers, policymakers, and community leaders, the implications of "agent autonomy" can no longer be deferred.

This year's turning point reflects not one event but a convergence of breakthroughs and governance advances. Major technology platforms have released "agent interoperability protocols" that enable agents to coordinate supply chains, workflows, and collaborations. These protocols include Google's Agent2Agent and Agent Payments Protocol, Anthropic's Model Context Protocol, the OpenAI-Stripe Agentic Commerce Protocol, and the open-source Agent Network Protocol.

Agentic features in consumer tools also became default rather than optional. Google, Opera, Perplexity, and Microsoft integrated agent-like capabilities into browsers, which retrieve information and complete tasks such as making reservations, generating code, summarizing articles, and drafting emails.

Meanwhile, governance caught up**.** The EU AI Act began enforcing its strictest prohibitions and governance rules for general-purpose models. California's SB53 ("Transparency in Frontier Artificial Intelligence Act") required public disclosure of safety protocols, 15-day incident reporting, and whistleblower protections. Globally, the UN General Assembly's first Global Dialogue on AI Governance urged alignment of rules to "build safe, secure, and trustworthy AI systems."

The adoption of the ISO/IEC 42001 management standard accelerated as organizations embedded governance throughout the AI lifecycle. Salesforce, Microsoft, AWS, Google, Samsung SDS, and IBM have achieved or are pursuing certification, setting an industry precedent that is expected to drive broader adoption.

These shifts moved the debate from "Can we build autonomous agents?" to "Which autonomy do we allow, who oversees it, and how do communities share power?"

Trust Signals Across the Lifecycle: Why Interoperability Isn't Enough

In July 2025, a Capgemini study estimated that agentic AI could create $450 billion USD in value by 2028. However, trust in fully autonomous agents decreased from 43% to 27% due to concerns about privacy and ethics. People will not adopt systems they do not trust,

so visible trust signals must exist throughout the agentic lifecycle: from registration and selection, to runtime monitoring and outcome review.

Interoperability protocols are vital for cross-developer collaboration, but they are insufficient for ensuring safety. Each protocol handles different dimensions (such as tool access, message exchange, or permissions), thereby creating challenges in alignment, access control, and debugging. Without full-stack visibility and trustworthy signals, interoperability remains a technical feat that is vulnerable to failure or adversarial use, thereby undermining adoption.

Autonomy Choices: When and How We Let Agents Act

A central question is: when is full autonomy appropriate, and when must humans remain in the loop? An agent that restocks warehouse inventory carries lower stakes than one issuing microloans in marginalized communities, or routing emergency responders.

Effective deployment requires striking a balance between determinism and guided non-determinism. Salesforce recommends a layered framework:

- **Deterministic scaffolding = safety contract:** Actions that must never occur (e.g., PII exfiltration, regulatory violations) are encoded via validators, role-based access, and runtime enforcement. They're auditable and binary.
- **Guided autonomy = operational creativity:** Within those rails, agents plan and select optimal actions, producing efficiency, resilience, and higher value.
- **Fallback modes or "kill switches":** Enable human audit, breaks, and rollback if something goes wrong.

Designing for defence in depth means choosing the right balance of autonomy and control for each use case, capturing agentic value while maintaining legal, audit, and customer predictability.

From Model Governance to Agent Governance

Earlier AI ethics efforts focused on model-level audits for bias, explainability, and data governance. In an agentic world, governance must be real-time, scalable, and embedded.

ISO/IEC 42001 offers a lifecycle management scaffold, spanning design through retirement. Yet, many organizations lack the necessary staff, logging capacity, or observability tools for complex agent systems. Emerging approaches, such as BlockA2A, combine blockchain ledgering and smart-contract controls with agent communication channels to ensure immutability, cross-agent accountability, and revocation.

The shift ahead is from checkpoint to continuous governance, and from single-model auditing to inter-agent traceability.

Communities, Workforce, and Power

When automated systems produce harmful outcomes, real people, particularly marginalized communities, bear the brunt of the cost. In housing, policing, social services, or credit, an agent's bias or error causes bodily, financial, or civic harm. Community wisdom must inform governance, not follow it.

At the same time, autonomy is transforming the way work is done. Tasks within roles are changing more than roles are disappearing. The UN warns that 40% of global jobs could be affected by AI and automation. Preventing inequality will require apprenticeships and upskilling programs that are co-designed with the affected communities.

What to Watch in 2026 and Beyond

Two fault lines will shape the next phase:

1. **Growth vs. Trust.** The economic benefits of agentic AI are immense, but the lack of oversight and interpretability could block adoption. Organizations that prioritize transparency and reliable guardrails will lead.
2. **Global alignment vs. fragmentation.** The UN Dialogue, EU AI Act, and California SB 53 are promising starts, but without interoperable norms, the world risks a patchwork of regional regulations that will hinder innovation.

As AI systems begin to act independently, they cannot escape the social contract; they must extend it. The task ahead is to build agentic systems that are capable, accountable, and just, governed together rather than left to chance.

About the Author
Kathy is a Principal Architect / VP of Responsible AI and Tech at Salesforce. She is a member of Singapore's AI Verify Foundation and a Council Member on the WEF Global Futures Council on Data Frontiers. Previously, she served as a Visiting AI Fellow at NIST. She co-authored two editions of "Understanding Your Users: A Practical Guide to User Research Methodologies." She received her MS in Human Factors Engineering and BS in Applied Psychology from the Georgia Institute of Technology.

Cite this Article
Baxter, K. (2025). When AI Begins to Act on Its Own. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 122-124. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 14: Democratic AI – Community Control and Open Models

14.1 Learnings for Canada: Community-led AI in an Age of Democratic Decay

By Jonathan van Geuns, Independent Researcher

14.2 From Accessible Models to Democratic AI

By David Atkinson, Georgetown University

14.3 Open Science Practices for Democratic AI

By Ismael Kherroubi García, Kairoi, RAIN and MAIEI

DOI: 10.5281/zenodo.17328882

14.1 Learnings for Canada: Community-led AI in an Age of Democratic Decay

By Jonathan van Geuns **, Independent Researcher**

AI governance often performs as a façade. Behind the curtain, a crisis grows: due process erodes as automated systems replace notice, reasons and records, and appeal with opacity. The question is no longer about how to make participation meaningful, but how to protect the conditions that make participation possible at all. What happens when the backbone of democracy collapses, and what must communities build *now* to govern AI when institutions fail to?

In the US, procedural democracy is currently undergoing a stress test. The Heritage Foundation's Project 2025 blueprint and "Schedule F" style reclassification aim to strip civil-service protections from policy roles, consolidating political control over administrative decision-making, which is the very layer that translates rights into remedies. This is not an abstract separation-of-powers debate. It's a removal of people and processes that make contestation possible; it's political capture.

The lesson for Canada is not smug exceptionalism but urgency. If a similar swing arrives here, procedural guarantees must be hard-wired into how AI is procured, deployed and contested. Build for the rainy day, not the sunny press conference; a public stack assembled now so that governance doesn't evaporate when politics whiplash.

Ownership and Stewardship: Cooperatives and Libraries

Platform cooperativism offers a template for reclaiming digital infrastructure. Instead of terms of service drafted by corporate counsel, co-ops are governed by bylaws. Instead of one-share-one-vote, they operate on one-member-one-vote. Surplus returning as compute credits or community funds. Applied to models and datasets, co-ops let communities set access tiers, define acceptable uses, and recall consent. This is a form of counter-infrastructure, like credit union logic for models. An example is READ-COOP that develops and hosts its own automated text recognition solution.

If co-ops address who owns, libraries address who cares. Shannon Mattern's "Library as Infrastructure" reframed libraries as civic utilities. Extending that to AI, libraries can host small, audited models for public tasks; run clinics where staff help residents understand and contest automated decisions; maintain public use registers; and act as neutral intake for complaints tied to municipal deployments. This gives due process an address instead of a modal dialog box.

DOI: 10.5281/zenodo.17328882

Due-process-by-design

When automated systems decide who gets benefits, jobs, housing, or entry across a border, there should be at least four guarantees: notice, reasons, records, and human review. In practice, none of these rights are secure. People discover an algorithm's role only after harm occurs, if at all. Explanations, when provided, are generic or legally shielded. Records vanish into proprietary code, and human review means little when decisions are rubber-stamped by the same opaque logic.

Under renewed political pressure to weaken the administrative state, even these fragile norms risk being dismantled. Canada should not wait for the same erosion. If due process is not built into code, contracts, and regulation before systems scale, it will disappear exactly when it is most needed. The task now is to ensure that public algorithms remain subject to human accountability even when institutions falter.

Participation is strong when baked into contracts. Municipal and provincial requests for proposals can require: dataset lineage and model cards; audited training/fine-tuning logs; the "rights pack" above; publish-and-update use registers; and decommission triggers if discrimination or due-process service-level agreements are missed. Look to New York City's "Automated Employment Decision Tools" regime. While not perfect, it is proof that audits and notices can be mandated and enforced as a floor, then go further on contestation and logs.

AI runs on water, power, and land. If communities can't decide when/where those resources are used, participation is cosmetic. Publicly funded AI should follow strict carbon and water budgets, train during low-impact hours, and publish energy ledgers for public scrutiny. When demand spikes, residents should have the right to throttle non-essential computation. Data-centre growth is an environmental and governance fact that demands local control.

What this Looks Like on Monday Morning

At the city library, a *Model Use Register* lists every algorithm used in public services. Residents can see what data each system draws on, and what to do if it gets something wrong. Librarians host clinics, helping people file appeals, draft review requests, and demand a human decision. On the city's procurement portal, vendors sign a due-process addendum before bidding: clear notice of automated use, reasons for outcomes, accessible logs, appeal timelines, and decommission clauses. The utility dashboard reflects if thresholds are breached, where residents can report through the same library desk. A visible due process.

DOI: 10.5281/zenodo.17328882

If the US spends the next years hollowing its administrative muscle, Canadian institutions will feel the gravitational pull. Regulatory arbitrage does not stop at borders. The counter is not grandstanding but infrastructure: co-ownership to decide ends, public stewardship to mind means, and justiciable due-process rights that survive any cabinet shuffle. Democratic AI will not be won in panels. It will be won in places where power changes hands or doesn't. Build them now, so that we the public hold the keys.

About the Author

Jonathan van Geuns is a practitioner, lawyer, researcher and writer working at the intersection of technology, governance, data and justice. His work examines how AI systems reconfigure power and participation, with a focus on community-led alternatives. He has worked with large international organizations and national governments, as well as super local initiatives. Jonathan's essays and projects explore democratic infrastructure, algorithmic governance, and the civic imagination needed for rights-based approaches to technology.

Cite this Article

Van Geuns, J. (2025) Learnings for Canada: Community-led AI in an age of democratic decay. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 126-128. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

14.2 From Accessible Models to Democratic AI

By David Atkinson, Georgetown University

There is a justified concern that closed AI systems, including generative AI systems such as ChatGPT, Claude, and Gemini, may not be democratic. In effect, we have outsourced decision-making power to a small, unelected council of tech giants.

Moreover, we have no practical way to weigh in on how the systems should be developed or deployed, nor do we have a voice in deciding which types of outputs are acceptable, who should have access, or how to ensure the benefits of the technology reach those who are most in need of AI's assistance.

An all too common refrain from several voices in the AI industry is that open-source AI models are the solution. Such models enable anyone with the necessary time, resources, and expertise to download, examine, modify, and utilize them. To many, "open" sounds synonymous with "free," "fair," and "for the people." Where closed models can be governed by a handful of companies by controlling access, implementing their preferred guardrails, and managing the fine-tuning process to guide models to perform in certain ways, open-source models offer a way to avoid these constraints. Open sourcing, many argue, *is* democratization.

However, these open-source-equals-democratization advocates, though often well-meaning, are mistaken.

The flaw in the argument is the underlying presumption that access equals democratization. In other words, by merely making the data, code, weights, evaluation data, and technical papers available to everyone, the knowledge and capabilities dormant in the technology are thereby democratized. But this makes little sense. If you give an anthill a laptop, have you democratized technology for all ants inside? Of course not. The same is true of AI components shared on sites like Hugging Face. How many landscapers now feel they can confidently participate in high-level discussions about AI thanks to their (probably unknown) ability to access Llama and Gemma models? I'd wager probably none.

Access alone is insufficient for a number of reasons. For one, it further separates the accessor from the company providing access. Google has little incentive to listen to virtually all of the people using their open-weight models. Meanwhile, Google can still do whatever it wants with Gemini, a model used hundreds of millions of times a month. The release of smaller open models serves as an effective distraction, allowing the core, most impactful systems to remain entirely opaque.

The shortcomings of access as a solution don't end there. The open models tend to be less capable than the best closed-source models. The number of people using any given open model is dwarfed by the number of people using any of the most popular closed models. Setting up an open model system requires a level of technical know-how limited to

a relatively tiny subset of the population. The cost to run the models for meaningful tasks or for many people is beyond what most individuals or small companies can afford, and even if someone has the money and know-how, they would need to find a sufficient amount of computing resources.

But perhaps the biggest knock against the idea of declaring democratization solved by tossing open-weight models to the masses is that it is not the type of democratization society needs or wants. Democratization isn't about being free to tinker with a toy; it's about having a voice in the systems that govern your life.

Meaningful democratization would entail a population able to sway the workings and uses of the models most likely to affect their lives. No amount of fiddling with a Llama model can compensate for how closed-source AI models make medical or financial decisions about people without providing them with a meaningful opportunity to weigh in. People should have a say in how power and money are consolidated in a handful of companies. Instead, those companies get to decide whether the environmental impact or a teen's suicide is a worthwhile tradeoff for a chatbot or image generator.

A democratic AI ecosystem would treat communities as co-governors, not passive users. It would include participatory oversight boards; the ability for affected communities to set guardrails and influence development before deployment; mandatory transparency reports about model behavior and data provenance; and the right for affected groups to contest harmful outputs.

Open source is an excellent idea in theory. It has many uses and upsides. But we should not fool ourselves into believing it is the same as democratization, and we should not settle for mere open source.

The challenge before us is not just openness but governance. We must demand systems of community control that subject the most powerful AI systems–the ones making potentially life-altering decisions–to public review, community-defined standards, and truly democratic oversight. We cannot allow what may be the most consequential political institution of the twenty-first century to remain the least democratic one of all.

About the Author
David is a postdoctoral fellow at Georgetown University researching how the law can be used to compel AI companies to act prosocially.

Cite this Article
Atkinson, D. (2025) From accessible models to democratic AI. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 129-130. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

14.3 Open Science Practices for Democratic AI

By Ismael Kherroubi Garcia , Kairoi, RAIN and MAIEI

AI applications and advancements are governed by Big Tech. Large, well-resourced private companies are who charter the course for a technology that is increasingly present in our everyday lives, whether we know it or not.

AI tools are used in different contexts, such as diagnosing patients, identifying criminals, allocating government benefits, and approving bank loans. Meanwhile, many generative AI applications are used by the general public and in the workplace to generate images, summarise research, brainstorm, and so on.

Before the proliferation of AI, many sections of the public have come to challenge the interests of Big Tech that the technologies promote. One way to challenge those interests is to "reclaim" AI; that is, "to distribute decision-making powers to different parties, from scientific communities to civil society" (Duarte *et al.*, 2025). In turn, an important toolkit for this effort is provided to us by the open science movement.

Open Science: A Very Brief Introduction

The open science movement has always been around in some shape or form; from ancient temples and libraries that stored knowledge and copied texts, to Enlightenment-era scientific institutions that allowed scientists to share and critique ideas and experiments. However, generally, we speak of the movement as a rather recent phenomenon. The advent of the internet effectively made much of open science possible. 1971 saw the birth of Project Gutenberg, which now serves as an online library of over 75,000 eBooks. In the scientific community, the desire for knowledge to be shared more widely was captured in the 2002 Budapest Open Access Initiative. As data and information sciences evolved, open access to knowledge became only one aspect of the movement. "Findability," "interoperability" and "reusability" would come to complement "accessibility" in the "FAIR Principles" (Wilkinson et al., 2016). Since 2016, the FAIR Principles have been adapted for many contexts, including for research software (Barker et al., 2022) and ML (Solanki et al., 2025).

Obfuscating Openness

Scientists and software engineers have led the way in making AI models more accessible and open to scrutiny. Hugging Face and OpenML are two such solutions: they allow for the sharing of datasets and code. These are key components of AI models, and to a degree, render them "open source." However, in recent years, the idea of an AI model's openness has become co-opted by Big Tech. And their motivations are clear: model openness is deemed essential for "sovereign AI." AI tools that are controlled by individuals or

DOI: 10.5281/zenodo.17328882

organisations, and free from external actors. “Openness” has become a marketing gimmick.

“Open source” is what enables sovereign AI. But the term has been misused by a number of tech firms. Meta is one company where “open source” has been used to describe their suite of LLMs, despite these models being unavailable to individuals and organisations based in the EU. A more common obfuscation of a model’s openness is in companies’ reference of their models’ “open weights.” “Weights” are a set of values fixed within an algorithm after training. However, having open weights does not help with a model’s reuse, or the sort of collaboration the open science movement encourages.

The distinction between “open weights” and “open source” causes some confusion, which has been exploited by the likes of Meta, OpenAI and DeepSeek. In a world where “openness” can be attached to any AI technology for marketing purposes, it is no surprise that data stewards and open science advocates have had to fight back and reiterate the spirit of the open science movement.

Reclaiming Open Science

Following years of collaboration with multiple stakeholders, and following a co-design process, the Open Source Initiative (OSI) defined “Open Source AI” in late 2024. The definition re-emphasises the value of open source AI tools: their reuse, study, modification and sharing. The definition also establishes what needs to be shared to be open source: information about data, code and parameters (such as weights). However, how much data can be shared without infringing on people’s freedoms, such as privacy and intellectual property? The definition provides some clarity: what is shared is information about datasets underpinning AI models, and how to obtain those datasets that are publicly available.

But this remains a key challenge, and the OSI followed up their definition with a report on data sharing and governance in February 2025 (Tarkowski, 2025). They suggest following different governance models for each type of data, which may be open (accessible and shareable without restrictions), public (accessible without authentication), obtainable (may be acquired through subscription or some other mechanism), or un-shareable and non-public (legally protected).

The spectrum of data types hints at openness itself being a matter of degree. What’s more, data constitute only one element of AI models, and the open science movement invites us to consider the many practices surrounding data. The OSI’s mention of “information about data” may relate with “metadata,” which can follow different standards according to a model’s domain, and be captured within “documentation,” which itself can follow diverse standards according to their target audience (Kherroubi García et al., 2025). These additional layers of complexity motivate the need to continue building and promoting frameworks that help scrutinize claims about AI models being “open.”

About the Author

Ismael is the Founder and CEO of Kairoi, the AI governance consultancy. He is also the Founder of the Responsible Artificial Intelligence Network (RAIN), and Participant Panel and Ethics Advisory Committee member at Genomics England. Ismael holds a master's in Philosophy of the Social Sciences, where he sought to uncover enabling conditions for multidisciplinary collaboration in scientific projects. As a result, a key tenet of his work is to advocate for an epistemic humility.

Cite this Article

Kherroubi García, I. (2025) Open Science Practices for Democratic AI. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 131-133. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

PART V: COLLECTIVE ACTION

DOI: 10.5281/zenodo.17328882

Chapter 15: AI Literacy – Building Civic Competence for Democratic AI

15.1 AI Literacy: A Right, Not a Luxury

By Kate Arthur, Independent, Author

15.2 AI Literacy: Building Civic Competence for Democratic AI

By Tania Duarte, We and AI

15.3 From Co-Creation to Co-Production: How Communities Are Building AI Literacy Beyond Schools

By Jae-Seong Lee, Electronics and Telecommunications Research Institute (ETRI)

DOI: 10.5281/zenodo.17328882

15.1 AI Literacy: A Right, Not a Luxury

By Kate Arthur, Independent, Author

In September 2025, UNESCO's International Literacy Day focused on "promoting literacy in the digital era." I was invited to deliver the keynote, an opportunity that reflected a growing recognition that literacy itself is being redefined and driven by the rapid advancements of AI.

I can't recall a time when I couldn't read or write; a privilege tied to access to education. Born in the UK, raised and educated in Nigeria, Argentina, and Saudi Arabia, and later settling in Canada, I learned through shifting languages and cultures that literacy is more than reading and writing, with its meaning evolving across time and place.

Literacy encompasses having the knowledge, skills, values, and behaviours that let us connect, share, and participate in society. With access to material tools and communication networks, we are literate when we can turn data into knowledge and respond meaningfully. Through public awareness, we understand the importance of literacy and, through education, we practise and improve the skills to become engaged and active citizens.

During the First and Second Industrial Revolutions, reading, writing, and numeracy enabled people to adapt to new economies and civic life. The Third Industrial Revolution added computing, requiring us to learn how to communicate, create, and build with technology. A digital world was being formed; one that mirrored both the good and the bad of our physical one.

Today, the Fourth Industrial Revolution is again transforming how humans engage. Advances in AI and other technologies are changing the way information is accessed, processed, and shared. For the first time, humans are not the only ones who can access data, transform it into knowledge, and respond meaningfully; so too can machines. And, unlike past revolutions that unfolded over decades, today's advances are accelerating at exponential speeds.

To be literate today means building on traditional and computing literacies, and now also AI literacy. This entails knowing how AI works, being able to question its results, and using it responsibly. It includes understanding the ethical and environmental impacts of the technology, as innovations take a toll on people and the planet. It also means having access to the tools and networks that enable participation, just as paper and pencils were once necessary to learn to read and write. With nearly three billion people still lacking internet access, and 800 million people remaining without basic literacy, technology risks reinforcing existing inequalities and leaving many voices unheard.

Around the world, 2025 saw new policies and programmes put AI literacy at the core of education and workforce skills planning. The US introduced an Executive Order on Advancing Artificial Intelligence Education for Youth. China made AI education mandatory,

DOI: 10.5281/zenodo.17328882

investing in teacher training and labs to reduce the urban-rural divide. The EU incorporated AI literacy into the AI Act, calling for both technical and ethical training for employees and users.

Other countries are advancing AI literacy agendas. Singapore and Finland led early public campaigns, while India introduced AI in schools and youth programs. South Korea uses AI to personalize learning and create digital textbooks. Australia developed a national ethics framework, and Estonia partners with tech firms to support teachers and students. In the Middle East, the UAE announced age-appropriate AI training for all grades.

These initiatives reflect growing international awareness of AI literacy as a foundation for social and economic participation. Yet progress is uneven, particularly in emerging markets and low-income countries, where limited infrastructure and resources make widespread access more challenging.

Governments, industry, and civil society all have a role to play in developing coordinated strategies for AI literacy that align with global standards while remaining locally relevant. Immediate steps include the integration of AI literacy into national AI strategies, AI public awareness campaigns, and financial investments into communication networks (internet), material tools (devices), and literacy skills training for all.

AI literacy must be built into lifelong learning, from early education to adult training and continuing education, including seniors, and across formal and informal learning settings. Schools need to lay the AI literacy foundations across disciplines, building year to year as both technical and ethical abilities strengthen. This includes investments in teacher training and accelerating curriculum revisions. Community centres, such as libraries, can ensure that internet access and computers are available, along with AI workshops offered free to the public. Reskilling and upskilling in the workforce allows for an inclusive, values-driven approach to AI technologies, as well as responsible use.

A Fifth Industrial Revolution is on the horizon. Education and workforce training must strengthen human-centric skills, such as creativity and critical thinking, to ensure strong human agency in AI design, development, deployment, and interaction.

AI literacy is not a luxury. It is a right. It ensures social mobility, protects democracy, and drives ethical innovation. To be able to interpret data into knowledge, to share it, and to record it, is to be literate. And to have the right to belong.

About the Author
Kate Arthur is a writer and lecturer specializing in AI literacy. Her work has reached over one million young people and thousands of leaders worldwide. She advises governments and NGOs, including UNESCO and UNICEF, on AI literacy and youth digital education policies. Kate is the author of Am I Literate? Redefining Literacy in the Age of AI, and is working on Algorithm to Adulthood, a new book helping young people build the skills to navigate life in a digital world.

Cite this Article
Arthur, K. (2025) AI Literacy: A right, not a luxury. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 136-137. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

15.2 AI Literacy: Building Civic Competence for Democratic AI

By Tania Duarte , We and AI

There has been a recent proliferation of frameworks and AI literacy programmes, at a national level, and in local and individual community initiatives. Some are driven by a desire to catch up schools and publics with the new skills required to navigate the impact of AI in the workplace, whilst others aim to safeguard against potential AI-enabled hazards. Less attention has been given to the provision and definition of the knowledge and competencies required for meaningful public participation in AI governance. As such, AI literacy initiatives have been falling short of delivering the key components needed to facilitate the necessary public scrutiny to hold the systems, design and use of technologies, and those who wield them, to account.

This is unsurprising given that AI literacy is often built on the model of digital literacies which have, through commercial capture, promoted technical training in a way that encouraged learners to become reliant on the technology, thus creating locked-in customers and profit for the tech industry. AI literacy is similarly influenced by technodeterministic agendas. Even in the curricula or content that acknowledge sociotechnical relationships, “AI development is [often] framed as *a priori* commercially valuable—and inevitable” (Duarte *et al.*, forthcoming, in AI and Ethics Handbook (Larry Medsker, Ed.) *Springer Nature*).

Following a roundtable intended to inform education policy advice in January 2025, The Royal Society, the independent scientific academy of the UK, identified that “AI literacy lies at the intersection of scientific, information, and data literacy. It is not only a matter of fostering technical skills but also of empowering individuals to critically engage with AI’s ethical, social, and economic dimensions“.

For public power to be truly possible, the confidence that is needed is not the confidence to use AI tools, or even to take responsibility for protecting oneself against harms such as misinformation, AI psychosis, automation, data exploitation. The confidence needed is instead to determine *if* and *when* AI use is even appropriate; whether the adoption will yield the promised benefits and if so, for whom, under what circumstances, and with what controls.

In response to the roundtable outcomes, the Royal Society commissioned a review of AI literacy frameworks, which from a longlist of 68,000 articles, identified three dimensions: technological (how AI works), practical (responsible use of AI) and human (e.g., societal implications). The analysis illustrated “a clear skew towards the technological dimension, with the practical dimension secondary, and the human dimensions least developed... Ethics and rights do often get some mention, but this is mostly superficial” (Hillman and Holmes, *A Rapid Review of AI Literacy Frameworks*, Forthcoming).

DOI: 10.5281/zenodo.17328882

An exception was the Council of Europe's (CoE) recommendation on AI literacy, which includes eighteen principles related to the human dimension. Among the principles are AI as a socio-technical assemblage; ML as different from human learning; the role of human responsibility in the democratic control of AI systems; the importance of understanding the historical evolution of AI and its transformative potential; the importance of going beyond "useful AI skills for the workplace" in order to focus on broader issues around the impact of AI on all workplace stakeholders; choosing when not to use AI as well as how to use AI responsibly; a consideration of the ethics of the full AI lifecycle; the impact on human rights, democracy, rule of law, the environment, peace and international security policing and justice, the geopolitics of AI, interculturality and plurilingualism, healthcare, and the role of educator and learner.

However, CoE aside, the lack of literacy provision aimed at enabling democratic or public power in favour of functional, skills-based literacy, has resulted in an emergence of Critical AI Literacy. Critical AI literacy, "unlike the term AI literacy often entails a decontextualized, instrumentalist approach to teaching and learning about AI, critical AI literacy emphasizes the ways in which AI technologies are situated in larger systems of power." Critical AI literacy is elsewhere described as "ability to critically analyse and engage with AI systems by understanding their technical foundations, societal implications, and embedded power structures, while recognising their limitations, potential biases, and broader social, environmental, and economic impacts."

Critical AI literacy learning objectives are resulting in creative delivery methods, in both formal and informal education. At the community level, We and AI have been conducting workshops encouraging engagement with the artefacts, people and materiality of the AI supply chain, as there are various initiatives in formal education based on exploring metaphors.

This movement has also resulted in some useful courses and curricula for universities which are available online such as Teaching Critical AI Literacies, Civics of Technology Curriculum, and OSU's AI Literacy Center.

Ironically, the use of generative AI in particular can also be seen as potentially degenerative to literacy, as "using AI is not about communicating. It's about avoiding communicating. It's about not reading, not writing, not drawing, not seeing. It's about ceding our powers of expression and comprehension to digital apps that will cushion us from fully participating in our own lives."

About the Author

Tania is the Founder of We and AI, a UK volunteer non-profit focusing on critical AI literacy advocacy and advisory, and related research, workshops and resources. With the We and AI community, she runs the global Better Images of AI library which provides images and inspiration for more transparent and inclusive communication about AI. Tania is currently working on enabling the decoding of narratives related to AI hype, AI inevitability, and through the use of AI metaphors.

Cite this Article

Duarte, T.. (2025) AI Literacy: Building civic competence for democratic AI. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 138-139. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

15.3 From Co-Creation to Co-Production: How Communities are Building AI Literacy Beyond Schools

By Jae-Seong Lee , Electronics and Telecommunications Research Institute (ETRI), South Korea

AI is no longer the domain of experts; it has become a public language through which societies negotiate power, fairness, and responsibility. Increasingly, meaningful discussions about AI happen beyond schools: in communities where people learn, debate, and design together. In these spaces, citizens are not passive users but active social learners and practitioners. Learning AI is not just about mastering tools but about relearning how to live together in a shared digital world. This article examines how community initiatives foster AI literacy as civic practice, showing how citizens evolve from subjects of technology to its co-creators and co-producers.

From Understanding to Co-Creation and Co-Production

Traditional AI literacy has focused on understanding how algorithms work and how to use them. This knowledge has been essential for co-creation, where citizens and governments build shared agendas around ethical technology. Yet, as AI shapes public decisions and social values, understanding alone is no longer enough. Communities must now pursue co-production, collaborating with institutions to redesign AI itself. This shift demands critical AI literacy: the capacity to question systems and shape their governance.

In co-creation, participants learn by doing: reviewing data, proposing policy alternatives, and translating technical issues into public dialogue. Many participants describe this experience as “making technology part of my own life.” Such learning nurtures confidence, solidarity, and political awareness.

In co-production, citizens go further, not as advisors but as partners in design and policy. They define public values, set priorities, and decide what to pursue or forego under limited resources. Through hackathons and pilot projects, they co-design, test, and refine solutions, feeding insights back into institutions. In doing so, they develop ethical reasoning, collaboration, and collective decision-making, skills that turn technical literacy into democratic capacity.

The Co-Creation Model: The vTaiwan Case

Taiwan’s vTaiwan initiative (born from the civic-tech collective, g0v, pronounced "gov-zero") is a leading experiment in co-creation, a model that turns the process of deliberation itself into a classroom for democracy. Since 2014, vTaiwan has combined

online deliberation tools such as Pol.is with offline workshops that bring together citizens, experts, and legislators.

Participants describe the process as “learning democracy in real time.” On “Pol.is,” clusters of public opinion are visualized, allowing people to see how divergent perspectives can converge into shared principles. Facilitators translate complex technical and ethical issues into accessible public language.

The model gained global attention in 2023 when vTaiwan joined the OpenAI “Democratic Inputs to AI” project with Chatham House and the AI Objectives Institute. ML analyzed large volumes of feedback, but human facilitators integrated results through value-based synthesis rather than computation alone. Participants began to see AI as a democratic tool rather than a technocratic threat. Still, sustaining broad and inclusive participation remains an ongoing challenge for such deliberative experiments.

The Co-Production Model: The Amsterdam Algorithm Register Case

If vTaiwan represents co-creation, the Amsterdam Algorithm Register illustrates co-production: a model in which citizens share real authority over the design, monitoring, and adjustment of AI systems. Managed by the city government, the Register publicly discloses every algorithm used in municipal services, outlining their purposes, logic, data sources, bias-mitigation strategies, and oversight mechanisms.

More importantly, the Register grants citizens the right to examine, question, and influence these systems. “It was the first time I could actually see how city algorithms make decisions about me,” one participant shared. “It made me want to ask better questions.” Through formal feedback loops, residents can request audits or suggest changes, and in several cases, citizen proposals have led to algorithmic revisions or deactivations.

Amsterdam’s experience shows that real learning and change occur when citizens gain authority as well as transparency. Rather than viewing algorithms as distant systems, participants act as co-governors, deciding which functions to maintain, retire, or reform. This empowerment turns oversight into shared governance, embedding AI literacy in everyday democratic life.

Conclusion

Co-creation through traditional AI literacy and co-production through critical AI literacy represent distinct but connected stages of civic learning. Co-creation centres on shared understanding and participation, while co-production centres on shared power and responsibility. In co-creation, citizens help imagine new policies. In co-production, they exercise real authority, setting priorities, managing trade-offs, and institutionalizing feedback. This shift from symbolic participation to collective governance marks the deepening of AI literacy as a democratic practice.

AI literacy fosters social change only when it expands from understanding to co-productive action. The most meaningful activities are unfolding beyond classrooms: in civic spaces where citizens see, question, and reshape the algorithms that shape them. Yet these experiments remain uneven. Care workers, older adults, and migrant communities still face barriers of time, access, and language that restrict their participation. Bridging these divides requires that citizens take an active role in shaping science and technology, and that their contributions be fairly recognized and rewarded. This is not a mere supplement but a fundamental condition of democracy.

About the Author

ICT policy expert with a PhD in Science and Technology Policy based in South Korea. Currently a Senior Researcher at ETRI's ICT Strategy Research Laboratory. Former Postdoctoral Researcher at KISTI, focused on data-driven policy. Actively engaged in global AI governance initiatives, panelist for Leiden University's Digital Roundtable on AI Governance, policy researcher with GoodBot, and committee to AI ethics framework research at AIEI. Served on the Program Committee for the 2025 AAAI/ACM Conference on AI, Ethics, and Society.

Cite this Article

Lee, J.S. (2025) From Co-Creation to Co-Production: How communities are building AI literacy beyond schools. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 140-142. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 16: Civil Society and AI – Nonprofits, Philanthropy, and Movement Building

16.1 From Proximity to Practice: Civil Society's Role in Shaping AI Together

By Michelle Baldwin, Equity Cubed & Alex Tveit, Sustainable Impact Foundation

16.2 Indigenous Approaches to AI Governance: Data Sovereignty, Seven-Generation Thinking, and Long-Term Stewardship

By Denise Williams, First Nations Technology Council (former CEO)

16.3 How Nonprofits Are Using AI: What's Working, What's Not, and What They Need to Succeed

ByJenni Warren & Bryan Lozano, Tech:NYC Foundation

DOI: 10.5281/zenodo.17328882

16.1 From Proximity to Practice: Civil Society's Role in Shaping AI Together

By Michelle Baldwin, Equity Cubed; and Alex Tveit, Sustainable Impact Foundation

Civil society organizations translate systemic failures into human stories, and human needs into systemic change. In 2025, as AI reshapes the landscape of social impact, that work is entering a new phase. Civil society is no longer waiting to be positioned as a beneficiary of innovation. We are stepping forward as its co-architects.

Civil society understands something AI still struggles to learn: context shapes outcomes as much as code. We've seen predictive models reproduce inequities; and we've seen trust, our sector's most vital infrastructure, treated as an afterthought rather than a foundation. Yet civil society also holds another kind of infrastructure: expertise built from years of proximity to communities. These deep reservoirs of insight remain underutilized in AI research and policy design; knowledge born from walking alongside communities in relationships of trust. When this knowledge is ignored, so too are the people it represents.

Relationships as Foundation

The most meaningful advances in 2025 were not technological; they were relational. Across Canada, organizations are proving that proximity generates wisdom no dataset can replicate. When Indigenous-led networks created data sovereignty frameworks, they were not just managing information, they were practicing reciprocity. When Black-led coalitions developed accountability metrics for hiring algorithms, they translated lived experience into technical specification. These are not consultation exercises, they are models of collaborative governance.

Civil society brings what algorithms cannot: accountability rooted in trust, proximity, and care. But that trust is accompanied by domain expertise: contextual knowledge, data, and relational intelligence that should inform how AI systems are designed, trained, and evaluated. The institutions funding and regulating AI must recognize this capacity not as anecdotal input, but as critical data infrastructure. The future we need is one where civil society's participation in AI governance is not granted as permission but recognized as essential for legitimacy, reliability, and resilience.

Coordination as Civic Infrastructure

Civil society has always worked differently. We collaborate, share strategies, pool resources, and learn in public because our legitimacy depends on collective outcomes, not market share. Shared procurement models are emerging. Community-governed data trusts are moving from theory to practice. Pooled funding mechanisms are being tested to let smaller organizations access AI expertise without compromising their independence or values.

DOI: 10.5281/zenodo.17328882

Canada already has the ingredients for such coordination: multidisciplinary research networks, philanthropic capital, Indigenous and community leadership, and social impact organizations embedded in every region. What remains missing is connective tissue, such as intermediaries, funding architectures, and knowledge commons, that turn isolated pilots into shared public infrastructure. The call of this moment is not for more pilots, it is for integration: systems that link community-held data, public research, and philanthropic insight into a shared ecosystem for ethical innovation. This has been the year we recognized this gap; 2026 must be the year we fill it.

Resourcing and Responsibility

The infrastructure gap is not only organizational; it's financial. Public innovation funding still flows primarily to universities and corporations, treating nonprofits as implementers rather than innovators. When governments design AI programs, community-based and philanthropic actors must be eligible as principal investigators and co-creators. Philanthropy, too, must shift from project cycles to long-term investment, funding not just tools but governance capacity: community advisory boards, algorithmic audits, participatory design, and the staff time to sustain them.

Civil society's contribution to AI is not just moral oversight; it is technical and practical. Our sector manages data ecosystems that reflect lived complexity; exactly the kind of nuance AI needs to function responsibly. Supporting this means funding data stewardship, shared learning platforms, knowledge commons and digital infrastructure that amplify collective intelligence rather than extract it.

Democracy and Decolonial Practice

As democratic institutions worldwide strain under polarization and authoritarian resurgence, Canada has an opportunity not to proclaim leadership but to embody pluralism. Our strength lies in convening across worldviews: Indigenous governance models rooted in reciprocity; diasporic networks linking local action to global insight; researchers and nonprofits co-creating ethical infrastructure. Decolonial AI is structural, not symbolic. It redefines who holds authority, how knowledge is valued, and how benefits are distributed. Related inquiries into democratic governance and decolonial AI practice extend these questions into institutional contexts.

If AI governance is to serve democracy, it must be co-created through trust and relationship. Civil society's work shows that accountability can be participatory, that transparency can be communal, and that technology guided by care can act as quiet resistance against concentration of power.

DOI: 10.5281/zenodo.17328882

Learning as Collective Practice

Realizing that scaling AI responsibly means scaling learning, not only deployment. We need documentation commons where nonprofits publish what works and what fails; peer networks where practitioners troubleshoot in real time; and funding models that reward openness and shared progress. Our capital isn't only financial, it's social and moral.

A Practice of Care

Responsible AI will not emerge from technical specifications and expertise alone. It will grow from staying in relationship with communities, with each other. Civil society's greatest insight is that care itself is intelligence. The work ahead is not about scaling technology but about scaling trust, turning coordination into coherence, and collaboration into shared civic power.

The question is no longer whether civil society can lead. It's whether governments, funders, and technologists are ready to lead *with* us. To build a future where legitimacy is measured not by control, but by connection.

That future is within reach, if we choose to fund it, govern it, and build it together.

About the Authors

Michelle is reimagining how capital, governance and tech can advance community-led futures. As Co-Executive Director of Impact United Academy and Senior Associate at Equity Cubed, she supports impact investors exploring equitable finance and tech. Formerly Senior Advisor at Community Foundations of Canada, she now collaborates with SuperBenefit DAO, All In for Sport DAO, Women in AI, Tech Stewardship and MAIEI, exploring how AI, Web3, and blockchain can redistribute power and reimagine philanthropy. She also teaches at Huron University College.

Alex Tveit is a systems-focused leader advancing community empowerment and inclusive innovation. He is Co-Founder of Sustainable Impact Foundation, Emerging Technology Fellow at Community Foundations of Canada, and serves on boards spanning climate, democratic deliberation, healthcare, and other impact areas. His work integrates systems thinking with community-centred approaches, addressing complex challenges through collaboration and knowledge mobilization. With a focus on ensuring technology contributes positively to society, Alex champions equity-driven innovation and partnerships that foster resilience and systemic change.

Cite this Article

Baldwin, M. & Tveit, A. (2025). From Proximity to Practice: Civil Society's Role in Shaping AI Together. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 144-146. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

16.2 Indigenous Approaches to AI Governance: Data Sovereignty, Seven-Generation Thinking, and Long-Term Stewardship

By Denise Williams, First Nations Technology Council (former CEO)

The Relational Nature of AI Governance

> *Artificial intelligence is an extraordinary tool. It is also a revealing reflection of who we are and who we are becoming.*

Across the world, Indigenous peoples are demonstrating what it is to be in relationship with technology as a future ancestor. As AI rapidly reshapes our relationships with each other and our systems, we are called to consider where this new ancestor is learning from. Who does it listen to? Who is advising it on the teachings that will ensure that the power of its intelligence brings to bear the best of human potential and not the alternative? When these intelligences are designed in conjunction with thousands of years of Indigenous intelligence, we significantly enhance our contribution to future generations. Our near-term work in AI integration aims to expand and advance our understanding of governance, sovereignty, and leadership in this context.

The work to advance this effort is why Indigenous leaders are actively reframing AI governance in accordance with Indigenous ways of knowing, being, and seeing. The *Assembly of First Nations (AFN)* and the *First Nations Information Governance Centre (FNIGC)* continue to be visionary in this work. Through the OCAP® principles (Ownership, Control, Access, and Possession)**,** they have built a framework for what responsible data governance looks like when it is, in fact, grounded in sovereignty. Data, in this worldview, isn't something you own. It's something you're responsible for and accountable to. Meaning data is not a resource; it is a relative, and that significantly changes the dynamic of the relationship.

Why Seven-Generation Thinking Matters Now

Seven-generation thinking asks us to honour the generations who came before and consider carefully those yet to come. It's a way of reimagining governance and innovation that expands accountability across time and dimensions.

Before launching an AI model, we might ask:

- How will this system affect our languages, lands, and grandchildren?
- What stories will it amplify, and which might it erase or change?
- What will future generations inherit from our design choices today?

DOI: 10.5281/zenodo.17328882

For Indigenous peoples, these questions are not theoretical, they represent governance in action. The kind of action required to societally shift our perspective from the enjoyment of deploying a short-term solution, to a responsibility we hold and are now committed to guiding and nurturing in this long-term relationship.

Where It's Working: Community-Led Innovation

Real-world examples are already showing what ethical, community-led AI looks like:

- **First Nations Information Governance Centre (FNIGC)** has established regional data hubs, where communities govern their own information. These hubs ensure that AI applications, from healthcare to housing, begin with Indigenous consent and community-defined ethics.
- **Animikii Indigenous Technology**, based on Lekwungen territory (Victoria, BC), builds platforms that give Indigenous nations full control over their digital data. Their software reflects values of respect and reciprocity, demonstrating that innovation and sovereignty coexist.
- **Inuit Tapiriit Kanatami (ITK)** is advancing Inuit data sovereignty by applying AI tools to Arctic climate and health research. Inuit organizations determine how data is used and shared, ensuring that research prioritizes Inuit wellbeing over external interests.
- **Cowessess First Nation** in Saskatchewan uses AI to manage solar energy systems, blending traditional stewardship with advanced analytics. Their approach demonstrates that clean technology can align with cultural values and self-governance.

These communities are not waiting to be invited into the digital future, they are building it, guided by their own protocols and laws.

The Work Ahead: From Policy to Practice

To move from intention to action, Canada must understand and center Indigenous digital governance as an essential component of its national AI policy.

This means recognizing Indigenous digital jurisdiction as an expression of inherent rights and treaty responsibilities. Just as Indigenous peoples express stewardship and rights recognition in relationship to land, water, and resources, we also extend that expression of stewardship and jurisdiction to digital society, because that knowledge also originates from our territories and peoples.

It means embedding OCAP® and CARE principles (Collective Benefit, Authority to Control, Responsibility, and Ethics) directly into federal and provincial AI and data laws, ensuring that Indigenous frameworks guide how data is collected, stored, and used.

It also means funding Indigenous-owned digital infrastructure, including community data centres, connectivity projects, and Indigenous-led AI research hubs; from design to delivery. These investments are not transactional, they are also part of the relationships we are building together and a demonstration of a commitment to justice, economic wellbeing and freedom, and reconciliation in this chapter of digital societies' evolution.

And finally, it means sharing power. Indigenous Nations must hold decision-making authority in Canada's AI ethics councils, standards bodies, and policy design processes. True co-governance means designing systems together, from the beginning.

These actions, along with the many offered by Indigenous peoples across the country, could provide a roadmap for practitioners and policymakers today. This could start with a simple shift in AI governance from risk management to relationality: a model that centers on consent, trust, and long-term accountability.

Honesty and Hope

The path forward requires bold and precise action. Building governance models that uphold this work and these principles will require governments to share authority, companies to shift business models, and institutions to learn and unlearn, in the pursuit of two-eyed seeing. The good news is, this work is already happening in communities, classrooms, and institutions everywhere.

The important thing to remember is that Indigenous peoples have always been technologists. Our ancestors engineered governance systems, trade networks, and ecological knowledge systems throughout this land's most complex and evolutionary times of change. The same principles that designed and sustained those systems can also sustain this next evolution of technological systems change. This is truly a journey of shared liberation, if we choose to see it that way and invest in it accordingly.

The digital world we are building will remember what we teach it. Let's ensure it remembers respect, humility, and our commitment to life in all its forms. If we do that, AI can become not just intelligent, but wise.

> *When we design and govern for seven generations, our technologies become good ancestors*

About the Author

Denise Williams is a proud member of the Cowichan Tribes and former CEO of the First Nations Technology Council. A Dialogue Fellow at SFU's Morris J. Wosk Centre for Dialogue, she leads national conversations on Indigenous digital sovereignty, AI governance, and economic well-being. Denise holds an MBA from SFU's Beedie School of Business and works across universities, philanthropy, and government to build systems grounded in reciprocity, innovation, and Indigenous leadership that shape a more equitable future.

Cite this Article

Williams, D. (2025) Indigenous approaches to AI governance: data sovereignty, seven-generation thinking, and long-term stewardship. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 147-149. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

16.3 How Nonprofits Are Using AI: What's Working, What's Not, and What They Need to Succeed

By Jenni Warren & Bryan Lozano, Tech:NYC Foundation

When Dave Spencer at New York Sun Works trained an AI model to analyze program reports, he expected a small efficiency gain. Instead, the task that once consumed an entire workday now takes ten minutes. Across the social-impact field, similar experiments are transforming daily work. From museums generating personalized storybooks to libraries offering multilingual search, nonprofits are discovering that AI's value lies in the quiet redesign of everyday tasks.

Decoded Futures, a program of the Tech:NYC Foundation, has trained over a thousand nonprofit leaders in hands-on, practical AI applications. Its training approach combines technical learning with human-centered design, helping participants translate new tools into meaningful practice.

Three key components make this learning stick:

1. **In-person, exploration-based learning:** Participants experiment with AI tools, connecting concepts directly to their own organizational challenges.
2. **Tech expert support:** Each cohort includes access to practitioners who can troubleshoot, recommend tools, and translate technical ideas into nonprofit contexts.
3. **Workflow and problem mapping:** Participants identify specific pain points first, then design clear workflows where AI can add measurable value.

Alongside these practices, every project is grounded in three ethical guidelines that ensure responsible and inclusive use:

1. **Personally Identifiable Information:** Free AI tools aren't private; never share sensitive or personal data unless it's through a secure platform.
2. **Keep a Human in the Loop:** Always review AI outputs before relying on them or sharing externally.
3. **Customize for Your Community:** Out-of-the-box models rarely reflect every culture or experience; nonprofits should adapt AI to serve their specific audiences.

Follow-up interviews reveal a clear pattern: AI has become a practical collaborator across fundraising, education, workforce development, and public service. 95% of alumni continued using AI after the program, building more than 200 new workflows. Their stories reveal what's working, where friction remains, and what kind of support will determine whether this next wave of adoption advances equity or deepens divides.

DOI: 10.5281/zenodo.17328882

What's Working

Everyday efficiency: AI is saving nonprofits time and mental bandwidth. At CUNY K16, Rachel Hutchins uses AI for charts and regression analyses that used to take hours. These changes shift effort away from repetitive work and toward mission-critical programs.

Mission-aligned innovation: Organizations are using AI to enhance creativity and access. The Bronx Children's Museum built a story-generation tool featuring Bronx kids, expanding literacy access. The Brooklyn Public Library launched a multilingual recommendation system across 60 branches. Both examples show AI applied with cultural sensitivity.

From pilots to systems: Some nonprofits are embedding AI into core operations. At The Leadership Academy, Alexander Negron integrated AI into Salesforce and Monday.com to automate workflows and support staff learning. United Neighborhood Houses is building a dashboard linking 45 settlement houses, creating real-time visibility across 800,000 New Yorkers served.

Internal capacity building: Adoption spreads fastest when someone inside champions it. At Berkeley College, Randy Gomez built a cross-department GPT to identify partnerships and trained colleagues. At CUNY K16, Rebecca Beeman trained 18 senior leaders and surveyed adoption rates to guide decisions. Such initiatives turn curiosity into institutional readiness.

What's Not Working

- **Institutional hesitation:** At larger organizations, some alumni use AI daily but face a system-wide moratorium. Similar caution across large institutions slows innovation until policy catches up, forcing some staff to innovate quietly.
- **Resource constraints:** Smaller nonprofits often lack the engineering support necessary to maintain their tools. Innovators easily build prototypes but struggle to sustain progress without technical mentorship or funding.
- **Weak measurement:** Few organizations formally track AI outcomes. Most rely on anecdotes, which makes it harder to evaluate the results.
- **Fragmented support:** Leaders describe needing continuing education, troubleshooting spaces, and policy guidance. They want peers to learn with, not just one-off training sessions to attend.

What Nonprofits Need to Succeed

The next phase of AI adoption depends on ecosystem design, which includes shared infrastructure, peer learning, and steady investment.

- **Community of practice:** Networks or spaces to exchange examples. Peer learning transforms scattered pilots into a collective force for progress.
- **Hands-on guidance:** Continued office hours or access to experts who can help implement ideas in real time.

- **Advanced training:** Deep workshops on automation, workflow design, and ethics, especially those scaling from prototypes to systems.
- **Organizational support:** Writing AI policies, integrating tools, or connecting with funders. *What nonprofits need most is patient capital for responsible scaling.*

Funders can meet these needs by investing in communities of practice, applied learning grants, and shared ethical infrastructure.

The Path Forward

AI in civil society is evolving from pilot projects into everyday infrastructure. The leaders driving this change are proving that innovation rooted in mission can thrive without massive budgets. Even as funding is pulled back across the US, nonprofits are learning to scale impact and reshaping the social sector with AI. What they need now is sustained funding, technical mentorship, and communities that foster continuous learning.

That is how AI transforms the social sector, becoming a durable tool for equity and change.

About the Authors

Jenni Warren is the Program Director of Decoded Futures at Tech:NYC, leading efforts to equip nonprofits with the transformative power of AI. With over 15 years of experience in learning and development both internationally and domestically, she has built innovative programs at the intersection of education, technology, and equity. Jenni holds a masters degree in Early Childhood Education from The Ohio State University and is passionate about using tech for social impact.

Bryan is the Director of the Tech:NYC Foundation. Bryan is leading the foundation to scale the organization's K-12 CS education and workforce development initiatives and deepen the tech sector's support of other economic development and social impact issues. Under Bryan's leadership, the Foundation has launched Decoded Futures, an initiative empowering NYC nonprofits with AI tools and expertise to enhance their social impact. A graduate of Stony Brook University and the NY Coro Fellows Program.

Cite this Article

Warren, J. & Lozano, B. (2025). How Nonprofits Are Using AI: What's Working, What's Not, and What They Need to Succeed. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 150-152. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Chapter 17: AI in Government – Public Sector Leadership and Implementation

17.1 Unions, Lawsuits and Whistleblowers: Public Sector Leadership from Below

By Ana Brandusescu, McGill University

17.2 AI in Government: Accessibility, Trust, and Sovereignty

By Tariq Khan, London Borough of Camden County Council, United Kingdom

17.3 The Hard Work of AI in Government

By Jennifer Laplante, Government of Nova Scotia, Canada

DOI: 10.5281/zenodo.17328882

17.1 Unions, Lawsuits and Whistleblowers: Public Sector Leadership from Below

By Ana Brandusescu , McGill University

We live in dangerously unprecedented times. Authoritarian rule is expanding across the globe. In March 2025, V-Dem's *Democracy Report* highlighted that 72% of the world population is under autocratic rule;, the highest since 1978. The rise of corporate and state violence has been exacerbated by technologies like AI.

In 2025, generative AI changed the game. The best-known LLMs have seen unprecedented adoption and push into our daily lives, from intimate conversations to office communications. Generative AI tools are being deployed in darker contexts like war and genocide; contributing to new forms of post-traumatic stress disorders; causing psychosis and suicide; endangering drought-stricken communities; and massive theft of all types of art. This is a side of AI that many politicians and tycoons do not discuss. Instead, they celebrate the potential for economic growth, efficiency, and cost-saving. We have heard this refrain countless times, despite continuous public skepticism and low trust in AI.

Deregulation also reshaped the landscape. We have entered an era defined by the loosening of oversight, seamlessly aligned with the political climate of authoritarianism and enormous investments in compute and technology firms. The end of investments is nowhere in sight. In fact, they are increasing, and this is justified under the banner of national security and digital sovereignty.

Public sector leadership means political will. It also means bureaucratic power where public servants on the other side of the political equation play a significant role in AI development, deployment and implementation. True accountability demands moments of discomfort, such as acts of individual and collective dissent. Such actions are important in a moment where governments are increasingly pressured to not only adopt AI systems but also to adopt governance by AI. The silver lining of 2026 may rest with the public servants to shape a new kind of *public sector leadership from below*.

Amid growing job instability, 2026 is poised to be a year of increasing unionization. A pivotal time to take collective action and safeguard livelihoods. Public sector officials should join forces with fellow union organizers to mobilize against the ongoing injustices embedded in AI development and deployment. The Tech Workers Coalition aims to strengthen solidarity and trust between tech developers and social justice organizers that centre lived experiences of communities and groups.

Public sector officials also should take note from workers in Kenya and the US. In Kenya, lawsuits have been filed against Meta, OpenAI, and TikTok. As data workers unionize for vital protections such as better pay and mental health support. Due to increasing global backlash over the effects of generative AI, the Global Trade Union Alliance of Content

Moderators was established in April 2025 to protect workers from algorithmic wage and labour exploitation. In the US, workers organized to strike against the film industry, critiquing the negative impacts of AI. They were successful in preventing production companies from deciding when they could use and not use AI. At a time of austerity, the ability for workers to engage in class action lawsuits is critical.

2026 may mark a time of government contraction and closure. Regardless of how open or closed states become, robust whistleblower protection laws will remain essential to keep state power in check and hold governments and corporations accountable. Whistleblowing is paramount for detecting fraud, triggering government investigations. Canada ranks among the worst countries in the world for protecting whistleblowers, even with substantial parliamentary committee reviews of the Public Servants Disclosure Protection Act.

Whistleblowers face racism and retaliation, and given their precarious positions, they desperately need robust protections and incentives. Before it was weaponized by the current US administration, the *False Claims Act* represented such a robust framework. Three points made the FCA particularly strong: it prohibited retaliation against government employers or contractors; it guaranteed access to competent counsel; and the whistleblower would obtain a financial reward if their work led to an enforcement action. The Silenced No More Act in California passed in May 2021 to ensure that workers who experienced workplace discrimination or harassment get heard and supported. In a time of fascism, taking a stand is the most courageous thing one can do.

2026 will mark a moment of full force AI adoption across government. All the more reason, then, to collectively slow down AI adoption and implementation. This is a time to draw hard lines around governments implementing some AI but not all AI. 2026 is a year of making clear distinctions between AI and generative AI, and an honest reckoning of what technologies are truly needed for government to function properly. It is a chance to revisit all the unglamorous AI already embedded in government: ML, natural language processing, AI without the generative part. To ask, is more data really needed? Can we do with what we have? 2026 is an opportunity to resist and build small when corporate players are asking to do otherwise.

About the Author

Ana Brandusescu is a researcher, policy analyst and social scientist who works on a more accountable use of AI. Under her PhD capacity, she researches privatization and political power in the scale of AI governance. As a Balsillie Scholar, she examined privatization in governance by AI. As the 2019-2021 McConnell Professor of Practice at the Centre for Interdisciplinary Research in Montreal, she researched AI policy and public investments and served on Canada's Multi-Stakeholder Forum on Open Government.

Cite this Article

Brandusescu, A. (2025) Unions, Lawsuits and Whistleblowers: Public sector leadership from below. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 154-155. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

17.2 AI in Government: Accessibility, Trust, and Sovereignty

By Tariq Khan, London Borough of Camden County Council, United Kingdom

Across the world, governments are under pressure. Slowing growth, rising public debt, ageing populations, and inflationary drag are straining economies. In the UK, the Institute for Fiscal Studies warns that, unless the government can deliver unusually high productivity gains, its fiscal plans could become unsustainable. Similar tensions are playing out across Europe, North America, and Asia. The challenge is shared: deliver more with less, while restoring public trust in institutions. The question has become now *whether* to adopt AI, but *how* to do so in ways that enhance accessibility, preserve trust, and maintain digital sovereignty.

AIs' Short-Term Benefit: Productivity or Accessibility?

The private sector's experience with AI offers a cautionary tale. MIT's *State of AI in Business 2025* found that, while 95% of companies have experimented with generative AI, only 5% have achieved measurable gains. Most pilot projects stall because tools fail to integrate into workflows or adapt to complex environments. Gartner's data echoes this: fewer than 4% of IT leaders believe Microsoft Copilot currently delivers significant business value. Governments risk repeating these mistakes.

The UK's Department for Business and Trade provides a different perspective. Its 2025 evaluation of Microsoft 365 Copilot showed modest *productivity* gains but profound *accessibility* benefits. Neurodiverse staff reported that Copilot helped them organise thoughts and manage workloads; non-native English speakers found it boosted confidence and communication quality. It made collaboration easier, meetings more inclusive, and writing tasks less daunting.

Governments exist to remove obstacles that limit full participation in society. Accessibility is not a side effect of AI adoption; it is one of its clearest public-interest victories. Yet accessibility doesn't make headlines and remains under-celebrated. Governments that are serious about legitimate AI adoption should see accessibility as a foundational pillar, not a fringe benefit.

Trust, Digital ID, and Who Builds It

Trust remains the true currency of digital government, and nothing tests it more than identity. The UK's recently announced digital ID programme, designed to manage everything from right-to-work checks to public service access, faces early scrutiny. Cybersecurity experts warn that centralising sensitive identity data creates a tempting target for attackers. Civil society groups fear exclusion, surveillance, and opaque governance.

However, the real question is *who* builds and controls such systems. In 2025, the UK deepened its ties with Palantir through a £1.5 billion strategic partnership for defence and data capabilities, one of several major contracts the firm now holds across government, including National Health Service (NHS) federated data platforms, border and migration analytics, and national security systems. Palantir's history of supplying intelligence tools to the US Immigration and Customs Enforcement (ICE) deportation programme and the IDF has raised ethical concerns internationally. Its growing footprint inside the UK public sector risks importing those same controversies, particularly where opaque data-sharing agreements and predictive analytics intersect with citizen rights. These dynamics blur the line between public infrastructure and private strategy, eroding the very trust that digital government depends on.

Dependency, the Wild West, and the Erosion of Sovereignty

Public-sector pay scales make it difficult to recruit and retain technical talent. As a result, governments turn to consultants and vendors to fill capability gaps, outsourcing not just code, but strategy and control.

This has created what many inside the public sector now see as a *"wild west"* of AI procurement. The public sector is being inundated by technology suppliers, from global giants to fledgling startups, making audacious claims about the efficacy of their systems. Many of these claims cannot be independently validated. There are no shared benchmarks, no national testbeds, and limited technical capacity within government to verify impact. The result is a marketplace of hype and opacity, where vendor marketing often outruns measurable outcomes.

The Local Government Association's *State of the Sector: AI 2025* found that, while 95% of councils are exploring AI, two-thirds rely on external suppliers. This over-reliance risks a deeper loss; that of digital sovereignty. When identity systems, health data platforms, or AI infrastructure are built and maintained externally, the state's ability to govern its own technology weakens. In the name of modernisation, governments risk surrendering the very control that makes modernisation possible.

Rebuilding for Trust and Resilience

If governments are to lead responsibly in the AI era, they must rebuild legitimacy not through rhetoric but through participatory design, involving citizens directly in shaping how technology is governed.

Some of the most promising experiments in this space are happening at the local level. London Borough of Camden's *Data Charter* sets a new benchmark for democratic data governance: residents helped define the principles for how their data should be collected, shared, and used, creating a living social contract between the council and its community.

Equally significant is WAVES, developed by Demos in partnership with Google.org and London Borough of Camden. WAVES is an AI-enabled deliberation platform that gives

residents a structured way to debate complex policy questions, such as housing, climate, or urban safety. By combining large-scale language models with civic facilitation, WAVES helps policymakers surface collective intelligence rather than polarised opinion.

Together, initiatives like the Camden Data Charter and WAVES show what a more deliberative democracy around AI can look like, one where trust is not manufactured but earned through openness and shared governance. They demonstrate that the state can use AI not merely to deliver services, but to deepen citizenship itself. If governments lose control over identity, data, and trust, they lose something deeper than operational sovereignty, they lose legitimacy.

About the Author

Tariq Khan is an award-winning technology leader and CIO with a career spanning both private and public sectors. Having spent most of his career in the private sector, he now leads large-scale digital and AI transformation in local government. A Cambridge AI and Ethics graduate, he is passionate about using technology to drive social impact and innovation.

Cite this Article

Khan, T. (2025). AI in Government: Accessibility, trust, and sovereignty. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 156-158. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

17.3 The Hard Work of AI in Government

By Jennifer Laplante, Government of Nova Scotia

Public Sector Leadership and Implementation

Leading AI adoption in government is both a privilege and a balancing act. It takes patience, coordination, and resolve. It is easy to talk about innovation but much harder to achieve it inside an organization built on continuity, accountability, and public trust. Government is not designed to move fast and break things; it is designed to move carefully and keep things running. We exist to serve, not to disrupt. That means we must modernize safely and ensure that the systems supporting our citizens continue to operate while we improve them.

The Reality of Off-the-shelf AI

Within large enterprise systems, adopting an AI tool alone will not deliver transformation. Meaningful change requires a deep examination of how data moves through systems, how decisions are made, and where friction and inefficiency occur. This work demands time and alignment. True transformation depends on re-engineering processes, cleaning and structuring data, and integrating systems in a way that enables sustainable improvement. It is not about adding a tool; it is about strengthening the foundation beneath it.

AI performs best when applied narrowly and purposefully to a defined, repeatable process that benefits from improved speed, accuracy, or insight. Broad, general-purpose tools may assist with certain tasks, but genuine transformation, the kind that changes how people work and how services are delivered, comes from focusing deeply on specific areas and redesigning the processes that matter most.

This approach brings its own challenges. It often results in the adoption of multiple specialized solutions, what might be described as one-offs. Many originate from startups, which brings welcome innovation but also questions of long-term viability.

Meanwhile, we must also maintain the large enterprise systems that keep everything else running: services, payroll, health records, and land and motor vehicle data. Those systems are our backbone; they cannot fail. They require constant updates, patching, and attention. Supporting them is a major part of the work. We are always managing competing priorities, maintaining what is essential while testing new things that could make a difference. That is the daily reality of digital government.

The Weight of Legacy

Every large organization faces legacy systems; government simply has more of them. Over the years, departments built what they needed, often on their own. The result is a web of systems that must somehow connect, supported by people who understand how to

keep them operating. Moving to shared services and common platforms is the right direction, but it is complex. Citizens rely on these systems every day for essential services.

Few people get excited about investing in back-end infrastructure. It is like spending money on new wiring or plumbing: no one sees it until it breaks. Yet everything depends on it. Without a strong foundation of clean data, stable systems, and secure networks, AI will not succeed. Departments build isolated tools or pilots that solve small problems but do not connect to the bigger picture.

The Reluctance of Risk

Pilot projects only have real value when there is a path to sustain what works. Too often, projects are funded as short-term experiments without certainty of future support. When that happens, even successful efforts can stall and enthusiasm fades. It is understandable; budgets are limited and priorities shift. But if there is no realistic chance of scaling, sometimes it is better not to start. People's time, energy, and optimism are valuable. We should focus on fewer initiatives with the potential to last, where lessons learned can turn into action and the work leads to something enduring.

People, Process, and Priorities

Technology transformation is ultimately about people. Public servants take pride in the processes that ensure fairness and accountability, but those same processes can make change difficult. They provide rhythm and reliability, yet also reinforce comfort zones.

Introducing AI means questioning how work is done. It requires retraining, new roles, and new thinking. For some, that is exciting; for others, unsettling. Leaders must show that AI is a support, not a threat, and that it allows people to focus on higher-value problem solving. The experience and judgment of public servants are what make AI implementations practical and ethical.

Governments also face thousands of competing priorities. Choosing where to apply AI is not simple. Some opportunities directly serve citizens, while others improve internal efficiency. Each AI initiative must connect to a clear, measurable, and meaningful outcome. The hardest part of leadership is saying no to ideas that are interesting but not strategic.

A Path Forward

Despite the challenges of legacy systems, risk aversion, and competing priorities, the opportunity is immense. Governments hold rich data, deep expertise, and the ability to create impact at scale. Success depends on doing the unglamorous work first: strengthening foundations, aligning systems, training people, and building governance that supports safe innovation.

AI in government is not about replacing people or automating judgment. It is about amplifying capability, improving efficiency, and building smarter, more resilient systems that serve citizens better. Progress may be gradual, but it is real. Each pilot, each modernized system, and each data improvement builds confidence and capacity. In time, that steady progress will transform how government works, not through one breakthrough but through many deliberate steps forward.

We cannot lose sight of what comes next. Advancing our digital, data, and AI capabilities is not a separate ambition; it is the next step toward greater efficiency, deeper insight, and better service for citizens. It is how we build a government that is modern, resilient, and ready for the future.

About the Author

Jennifer Laplante joined the Government of Nova Scotia after working more than two decades in digital transformation, data governance and AI. She holds an MBA and an MSc in Computing and Data Analytics. Prior to joining government, she served as the Chief Growth and Investment Officer for Canada's Ocean Supercluster, where she helped companies innovate and make forward-thinking investments in their technology strategies.

Cite this Article

Laplante, J (2025). The Hard Work of AI in Government. In R. Butalid, C. Wright, & I. Kherroubi García, (Eds.), *The State of AI Ethics Report (Volume 7) - AI at the Crossroads: A Practitioner's Guide To Community-Centered Solutions*. pp. 159-161. Montreal AI Ethics Institute. DOI: 10.5281/zenodo.17328882. Available at: https://montrealethics.ai/state.

Special Tribute: Abhishek Gupta Remembered

In The AI Ethics Brief #166, we shared a moving illustrated tribute to MAIEI co-founder Abhishek Gupta by **Luca Baraldi** and **Laura Zambarda** of symboolic.ai.

Originally published in Italian, we are honoured to republish the tribute here in both English and Italian as part of *SAIER Volume 7*, which is dedicated to Abhishek's memory.

For those who wish to share memories of Abhishek or leave a digital memorial, please visit: https://www.forevermissed.com/abhishekgupta.

English version:

Original *Italian* version:

DOI: 10.5281/zenodo.17328882