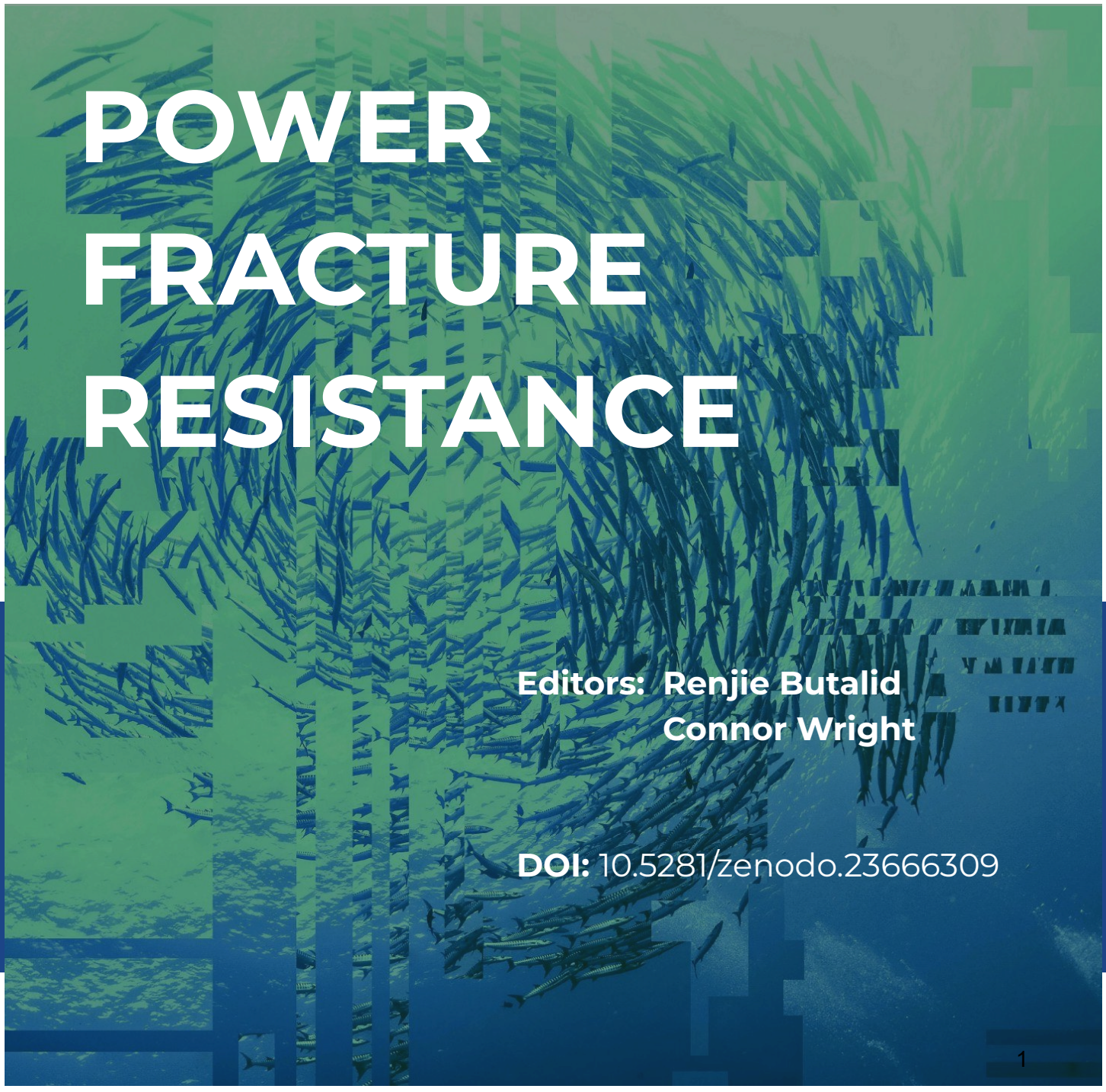
STATE OF AI ETHICS REPORT | VOLUME 8



POWER FRACTURE RESISTANCE

Editors: Renjie Butalid
Connor Wright

DOI: 10.5281/zenodo.23666309

2026 Montreal AI Ethics Institute (MAIEI) CC BY 4.0

CC BY 4.0

This work is licensed under the **Creative Commons Attribution 4.0 International License,** ensuring broad accessibility while respecting contributor rights. To view a copy of this license, please visit http://creativecommons.org/licenses/by/4.0.

Front & Back Cover Image: Distorted Fish School II by Lone Thomasky & Bits&Bäume / Better Images of AI / CC BY 4.0. ALT text: The image reveals the intentional, coordinated social lives of animals, which are often invisible to humans. Because this richness is sometimes hidden beneath the surface, the oceans can be disrupted by noise pollution, mineral and rare element mining and contamination caused by AI development. The distortion aims to visually represent the impact of destructive AI developments that are often too easily justified by human ignorance and technological "progress".

DOI: 10.5281/zenodo.23666309

Published by the Montreal AI Ethics Institute (MAIEI)

Montreal, Quebec, Canada
November 10, 2026
The State of AI Ethics Report (Volume 8)

DOI: 10.5281/zenodo.23666309

Suggested Citation

For citing the entire report:

Butalid, R., & Wright, C. (Eds.). (2026). *The State of AI Ethics Report (Volume 8) - Power, Fracture, Resistance.* Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

For citing individual pieces within the report:

[Author Surname], [Initials]. (2026). [Title of piece]. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) - Power, Fracture, Resistance.* pp.[xx-xx]. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. 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 8* 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.

Aaryan Shroff
University of Waterloo

Aida Gregorian
LawVo Inc.

Alex Tveit
Sustainable Impact Foundation

Alexander Carlo
Schulich School of Business, York University

Alva Markelius
Department of Computer Science & Technology, University of Cambridge

Amanda Janell DeAmor Quest
Commonwealth Caribbean Lawyer; Independent AI Ethics and AI Governance Researcher

Amanda Silvera
Society for Original Biometric Identity Rights

Amy Gaeta
Leverhulme Centre for the Future of Intelligence, University of Cambridge

Anca Ciulpan
Independent Researcher

Annabelle Bernard
Community Researcher

Anne Imouza
McGill University

Anne Pasek
Trent University;
Civil Society Hub for Tech Justice

Annie Gilbertson
Proof News

Archana Prasad
Gooey.AI;
Royal College of Art

Asim Javed
Independent Researcher

Audrey Tang
Cyber Ambassador of the ROC (Taiwan) Government

Ayaz Ul-Huq Syed
The Dais,
Toronto Metropolitan University

Aydon Moala-Edwards
Moala Global;
Monash Global Peace and Security Centre

Bahareh Goodarzi
Concordia University

Branka Marijan
Project Ploughshares

Caylee Kreller
NorQuest College Edmonton

Charles Migos
Intangible AI

Chelsea Sun
Kodera

Chiara Ullstein
Technical University of Munich (TUM)

Chijioke Okorie
University of Pretoria

Chinasa T. Okolo
Technecultura

Claudel Rheault
University of Edinburgh

Claudia Del Pozo
Eon Institute

DOI: 10.5281/zenodo.23666309

Connor Wright
Montreal AI Ethics Institute (MAIEI)

Crystal Braunwarth
IATSE 891

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

Debbie Folaron
Concordia University

Eddy Damaris Nono Defo
Université de Montréal

Elena Yunusov
Human Feedback Foundation

Eliot Tretter
University of Calgary

Emile Dirks
The Citizen Lab

Emma Ruttkamp-Bloem
University of Pretoria

Emily Osborne
Canadian Shield Institute

Fergus Linley-Mota
Dialogue on Technology Project, Simon Fraser University (SFU)

Heather von Stackelberg
NorQuest College Edmonton

Hillary Bergshoeff
IATSE 891

Isha Singh
Gooey.AI

Isabelle Castro
Independent Researcher; Utopia in Beta

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

Jade Maria Moisan
Université de Montréal

Janna Frenzel
Civil Society Hub for Tech Justice

Jennifer Raso
McGill University & Mila

Jess Reia
Digital Technology for Democracy Lab and School of Data Science, University of Virginia

Joss Moorkens
SALIS/ADAPT Centre, Dublin City University

Kaitlin Bustos
GoodBot

Katherine Chandler
Georgetown University

Katie Ledingham
University of Exeter Business School

Keegan Leahy
Turtle Island Foundation

Kei Baritugo
Montreal AI Ethics Institute (MAIEI)

Lesly Nzeusseu Kouamou
Université de Montréal

Luciana Soares Percegona
Hospital São Vicente, Brazil

Lynne Bowker
Université Laval

Lyndsey Rolheiser
Schulich School of Business, York University

Marcelo Haponiuk Rocha
Pontifícia Universidade Católica do Paraná (PUCPR)

Maria Lungu
Kennesaw State University; Digital Technology for Democracy Lab

Maureen Doyle-Spare
Doyle-Spare Research

Mia Travers-Hayward
Canadian Teachers' Federation

Michelle Okere
Indigenous Prosperity Foundation

Miranda Pate
Kodera

Moses Karanja
The Citizen Lab

Navid Sharifi
Concordia University

Norman Valdez
BrainTrainr

Priscila Chaves Martínez
Independent Researcher

Rebecca Finlay
Partnership on AI

Renjie Butalid
Montreal AI Ethics Institute (MAIEI)

Renee Black
GoodBot; Canadians for Digital Sovereignty

Renée Sieber
McGill University

DOI: 10.5281/zenodo.23666309

Riccardo Di Sipio
Human Feedback Foundation

Robert Schouten
Atlas Research; dBSC

Rob Pearce
University of Exeter Business School

Rosa E. Martín-Peña
Centre for Ethics and Law in the Life Sciences (CELLS), Leibniz University Hannover; Lower Saxony Center for AI and Causal Methods in Medicine (CAIMed)

Roshan Ghadamian
Institute for Regenerative Systems Architecture (IRSA); Ormond College, University of Melbourne

Sarah-Louise Ruder
Civil Society Hub for Tech Justice

Sara Jameel
Concordia University

Shimona Mohan
United Nations Institute for Disarmament Research (UNIDIR)

Shi Kang'ethe
Independent Journalist

Sophie Wu
Kodera

Stephanie Dick
Simon Fraser University (SFU)

Taras Kovalchuk
Center for AI and Digital Policy; Journal of AI Law and Regulation

Tariq Khan
London Borough of Camden, UK

Tim Davies
Citizens' Track on AI

Tina Crouse
Independent Researcher

Tom Simmons
Royal College of Art

Trisha Ray
GeoTech Center, Atlantic Council

Véronique Sioufi
BC Policy Solutions

Vukosi Marivate
Independent International Scientific Panel on AI, United Nations; University of Pretoria

DOI: 10.5281/zenodo.23666309

Acknowledgments

Thank you to every subscriber of The AI Ethics Brief. Our readers share, reply, and challenge our thinking. That engagement shapes this report.

A special thank you to our **Paid Subscribers.** Your financial support keeps The AI Ethics Brief free and accessible to everyone. It sustains our work at MAIEI in democratizing AI ethics literacy and honouring Abhishek Gupta's legacy.

SAIER Vol. 8 Team:

- **Editors:** Renjie Butalid, Connor Wright
- **Copy Editor & Marketing:** Kei Baritugo
- **Operations Coordinator:** Keilah Abello
- **Digital Production:** Mo Akif

Many thanks to the MAIEI team members and extended community of supporters and advisors, in addition to all of our contributors for Volume 8:

> Kate Arthur, Michelle Baldwin, Ana Brandusescu, Blair Attard-Frost, Jonathan van Geuns, Marianna B. Ganapini, Hannah Cole McGee, Masa Sweidan, Ismael Kherroubi García, Selen Dogan Kosterit and the entire team at the Governance & Responsible AI Lab (GRAIL), Tania Duarte and the entire team at Better Images of AI.

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

Table of Contents

INTRODUCTION **11**

Chapter 1: Power, Fracture, Resistance **11**
1.1 Power, Fracture, Resistance 12
1.2 What Holds: Building a Shared Evidence Base 16
1.3 Can AI Governance Keep Pace? Five Challenges and the Road Ahead 19
1.4 What Resistance Looks Like Now 23

PART I: PARTICIPATION, SOVEREIGNTY, & PRACTICE **25**

Chapter 2: Participation – Clauses, Assemblies, & Assessments **26**
2.1 The Vacuum Inside the Filler: Human Oversight in Government AI Guidance 27
2.2 Towards a Citizens' Track on AI Governance 30
2.3 Making Participation Count: A Method for Fundamental Rights Impact Assessment 32

Chapter 3: AI Sovereignty – Compute, Clouds, & Who Gets a Seat **35**
3.1 AI Sovereignty in an Age of Economic Shock 36
3.2 The Empty Sovereignty Promises of Canada's 2026 National AI Strategy 38
3.3 No Seat, No Data: The Pacific and the Limits of AI Sovereignty Talk 41

Chapter 4: Governance Frameworks & the Practitioners Building Them **43**
4.1 Ethical AI Needs a New Opening and the Sovereignty Push Might Just Be It 44
4.2 Practitioner-Led AI Governance Across the Global Majority 46
4.3 The Mirror We Hold 49
4.4 A Framework With No Scars Has Never Been Run 51

PART II: SURVEILLANCE, WARFARE, & OVERSIGHT **54**

Chapter 5: The Expanding Surveillance State **55**
5.1 The Warrant Nobody Could Explain 56
5.2 Guilty by Default: How Rushed Surveillance Law is Reversing the Burden of Proof 58
5.3 Disability Surveillance in the UK: AI, Care, and Power 61
5.4 Strategic Litigation as a Transformative Tool for Advancing AI Activism in the Global South: Lessons from the Commonwealth Caribbean 64

Chapter 6: Military AI Takes Centre Stage **67**
6.1 The Off Switch: How National Security Took Control of a Public AI Model 68
6.2 The Many Hands Problem: Securing Accountability in the Age of AI Warfare 70
6.3 Hide and Seek: The Paradox of Representation in Military AI 73
6.4 Towards a "Smart Mix"?: Governing AI for International Peace and Security 76

DOI: 10.5281/zenodo.23666309

Chapter 7: Oversight Built for Another Era **79**
7.1 The Stadium as Testing Ground: How the World Cup Risks Exporting the Authoritarian Playbook 80
7.2 Canada's Lobbying Act Was Built for a World That No Longer Exists 83
7.3 The Algorithmic Cost of Living: How Your Personal Information Will Be Weaponized Against You 86

PART III: LABOUR & INFRASTRUCTURE **89**

Chapter 8: AI at Work and the Workers Bargaining Over It **90**
8.1 Contract Bargaining in the Age of AI 91
8.2 The Clause and the Algorithm 94
8.3 When Workers Bring AI to Work (From Private Experiment to Organisational Capability) 96
8.4 When a Profession Is Automated: Translators and What Comes Next 99

Chapter 9: Data Centres and Communities **102**
9.1 Energy, Regional Politics, & the AI Build-Out 103
9.2 Data Centred: What Canada's Data Centre Expansion Means for Host Communities 105
9.3 How Communities Across Canada are Pushing Back Against Hyperscale Data Centre Development 108
9.4 The $15-Billion Text Message: Holding AI Infrastructure to Account 111

PART IV: CARE, CLASSROOMS, & CREATIVE WORK **114**

Chapter 10: Simulated Care – Triage, Companions, & Robots **115**
10.1 Invisible by Design: Disease, Illness, and Sickness in Algorithmic Triage 116
10.2 A Case for Human Connection as a Key Design Principle for Social AI Companions 118
10.3 Embodiment and Affective Harm: Social Robotics in the Era of Foundation Models 121

Chapter 11: AI in Education – Detection, Fatigue, & Curiosity **124**
11.1 More than 'AI-Ready': What Students Want from AI Literacy in Higher Education 125
11.2 Fatigue and Overwhelm: Institutional and Instructor Challenges in an Age of AI 128
11.3 Starting from Curiosity: Encouraging Youth AI Literacy through Self-Directed Research 131

Chapter 12: AI and the Arts – Authorship, Credit, & Voice **133**
12.1 Creative Agency in the Age of Generative AI 134
12.2 Unearned and Undisclosed: The Absence of AI Verification Standards 136
12.3 The Causes of the Causes 139

PART V: AGENTS, BODIES, & OPEN MODELS **141**

Chapter 13: Agentic AI – Accountability at Speed **142**
13.1 What My Own Audit Trail Got Wrong 143
13.2 Agentic AI: Self-Regulation Is the Regulation – The Semantic and Reasoning-Layer Governance Gap Across Institutional Boundaries 145

DOI: 10.5281/zenodo.23666309

MAIEI

13.3 Automating the Grounds, Not the Decision: What Care Records Should Have to Say 149
13.4 Automation and Bureaucracy: Principals, Agents, and Accountability in Public Administration 151

Chapter 14: AI and the Body – Biometrics, Health Data, & Bodily Autonomy 154
14.1 Confessions in the Cloud: How 140 Million Therapy Exchanges Became Training Data 155
14.2 From Gender Affirming Care to Reproductive Rights: The Surveillance Risk of Bodily Data 158
14.3 The Biometric Dragnet 160

Chapter 15: Open Models & the Communities Building With Them 163
15.1 The People's AI Stack: Mapping the Social and Political Choices Behind Artificial Intelligence 164
15.2 Re-Remix: Open Source Models for the World 167
15.3 ParcoursLab: A Course Recommender Made by Students, for Students 169

PART VI: LITERACY, COALITIONS, & GOVERNMENT 172

Chapter 16: AI Literacy Across the Lifespan 173
16.1 AI Literacy in K-12: To Whose End? 174
16.2 Prioritizing Responsible AI Education for Older Adults 176
16.3 What Nine Years of STEM Outreach Data Can Teach Us About Designing AI Literacy 179

Chapter 17: Civil Society – Austerity, Libraries, Coalitions, & Refusal 181
17.1 Optimized for Whom? AI, Austerity, and Ontario's Settlement Sector 182
17.2 When AI Comes to the Library: Public Libraries as Civil Society Infrastructure Under Pressure in the Age of AI – Lessons from a Québec Community Initiative 185
17.3 Who Gets to Shape AI? Building Coalitions for Underrepresented Communities 187
17.4 Who Gets to Decide? Building Indigenous AI Governance Through Relationship, Authority and Refusal 190

Chapter 18: AI in Government – What It Buys and Who Writes the Rules 193
18.1 When Government Gets Copilot, What Exactly Is It Buying? 194
18.2 Accountable To Whom? Let the Public Write the Rules for Public Sector AI 197
18.3 AI for All, Decided by Whom? Civil Society Power Beyond Consultation 199

CITATIONS & CONTRIBUTOR BIOS 201

DOI: 10.5281/zenodo.23666309

INTRODUCTION

Chapter 1: Power, Fracture, Resistance

1.1 Power, Fracture, Resistance

By Renjie Butalid, Montreal AI Ethics Institute (MAIEI)

1.2 What Holds: Building a Shared Evidence Base

By Rebecca Finlay, Partnership on AI

1.3 Can AI Governance Keep Pace? Five Challenges and the Road Ahead

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

1.4 What Resistance Looks Like Now

By Renée Sieber, Geography and Environment, McGill University

DOI: 10.5281/zenodo.23666309

MAIEI

1.1 Power, Fracture, Resistance

By Renjie Butalid , Montreal AI Ethics Institute (MAIEI)

In May 2026, when we opened the call for contributions for the eighth volume of the *State of AI Ethics Report (SAIER),* we asked readers of The AI Ethics Brief to consider the terms of the bargain we were making with artificial intelligence.

AI was becoming embedded in institutions, workplaces and everyday life. Governments were adopting it to deliver public services. Employers were introducing it into decisions about work. Companies were building products around increasingly autonomous systems. Each development raised questions about what people were being asked to accept, what they understood about those arrangements, and whether they had any meaningful opportunity to shape them.

We argued that these questions needed to be asked before the terms of the bargain were set for us. Six months later, as we publish SAIER Volume 8, the question has become more pressing.

During 2026, decisions about AI development and deployment continued to concentrate among a relatively small number of governments, technology companies and infrastructure providers. Export controls demonstrated how access to advanced models could depend on geopolitical decisions. National strategies promised technological sovereignty while confronting the practical limits of energy, compute and supply chains. Public institutions adopted AI through procurement arrangements and incremental operational changes that received comparatively little public scrutiny.

The consequences of these decisions extended well beyond the institutions making them. Across the 62 contributions to SAIER Volume 8, our authors examine how power is exercised, where governance arrangements fall short, and how people seek to influence decisions that affect their lives.

Power shapes who develops AI, who controls the infrastructure on which it depends, and whose interests influence the rules governing its use. Asim Javed **(6.1)** examines how a single export-control decision suspended worldwide access to two models, illustrating how availability can depend on determinations made elsewhere. Trisha Ray **(3.1)** examines the constraints facing national ambitions for AI sovereignty, while Véronique Sioufi **(3.2)** considers what those ambitions mean for Canada. Aydon Moala-Edwards **(3.3)** draws attention to Pacific countries whose perspectives remain underrepresented in international governance discussions and the datasets informing them.

The distribution of power is also visible in the physical infrastructure supporting AI. Anne Pasek and Eliot Tretter **(9.1)** examine the relationship between fossil energy development and data

DOI: 10.5281/zenodo.23666309

centres in Alberta, illustrating how decisions about AI infrastructure intersect with existing economic interests, environmental commitments and local governance. Anca Ciulpan **(18.1)** examines how procurement arrangements brought Copilot into U.S. federal agencies at no initial cost, while Tariq Khan **(18.2)** explores how incremental operational decisions introduced AI into British government institutions. Both raise questions about public scrutiny as AI becomes embedded in government operations.

These arrangements also expose fractures in AI governance, particularly between the safeguards institutions promise and their ability to demonstrate that those safeguards work in practice.

In risk management, the existence of a policy does not establish that a control is effective. That requires evidence that the control operates as intended, that failures can be identified, and that someone has the authority to intervene. That same standard is useful for examining AI governance.

Marcelo Haponiuk Rocha and Luciana Soares Percegona **(2.1)** consider how the principle that humans should retain responsibility for AI-assisted decisions can become a procedural requirement with limited practical effect. Aida Gregorian **(4.4)** examines what it means for governance frameworks to be tested against actual failures. Daniel S. Schiff of the Governance and Responsible AI Lab (GRAIL) **(1.3)** identifies broader gaps in governance, including implementation, cooperation and legitimacy. Rebecca Finlay of Partnership on AI **(1.2)** argues for a shared international evidence base, stronger accountability mechanisms and coordinated action to translate AI governance commitments into practice.

The consequences of these gaps become clearer in the experiences of people affected by AI systems. Maria Lungu **(5.1)** reconstructs a wrongful arrest involving facial recognition, while Annie Gilbertson **(14.1)** examines the use of sensitive therapeutic conversations as training data.

The settings differ, but the questions are familiar. When something goes wrong, can the institutions responsible explain what happened? Can they identify the failure, establish accountability and provide a meaningful remedy? The difficulty of securing accountability also helps explain why so many contributors turned their attention to resistance, and to the institutions through which people can challenge decisions.

Resistance takes different forms. It can involve collective bargaining, legal challenges, public participation, community organizing, or the development of alternative governance arrangements. Its effectiveness depends in part on whether people have access to institutions capable of responding to their concerns.

In Ontario, communities challenged proposed data centre developments through municipal planning processes. Janna Frenzel, Anne Pasek and Sarah-Louise Ruder **(9.3)** document how

DOI: 10.5281/zenodo.23666309

residents used public comments, zoning arguments and demands for moratoria to bring questions about AI infrastructure into local democratic institutions. These community-led efforts produced different outcomes across municipalities, demonstrating both the possibilities and limitations of existing public processes.

In the workplace, Crystal Braunwarth and Hillary Bergshoeff of IATSE 891 **(8.1)** examine the role of collective bargaining in determining how AI is introduced and governed. In Kenya, Shi Kang'ethe **(8.2)** documents content moderators' efforts to influence legislation affecting platform accountability, and the difficulties of translating participation into legislative outcomes.

Other contributors examine how communities are establishing governance arrangements grounded in their own needs and authority. Vukosi Marivate, Emma Ruttkamp-Bloem and Chijioke Okorie **(15.2)** explore African-led approaches to data licensing that establish conditions for the use of locally produced knowledge. Keegan Leahy and Michelle Okere **(17.4)** examine Indigenous data governance and the importance of retaining the authority to refuse participation.

Renée Sieber **(1.4)** takes this question further by asking whether resistance and refusal should themselves be recognized as legitimate forms of public participation.

Consultation provides an opportunity to express concerns. Meaningful participation requires institutions to demonstrate how those concerns influence decisions, including when a project should be changed, delayed or stopped. Norman Valdez and Alex Tveit **(18.3)** state it directly: “A seat matters; a say with institutional weight matters more.”

When we developed the initial outline for Volume 8, we could not have anticipated every direction our contributors would take. Their work changed the report. Questions about participation became substantial enough to warrant their own chapter. Environmental concerns developed into a broader examination of data centres, infrastructure and the communities expected to accommodate them. Across contributions on labour, public services, surveillance, education and creative work, questions of authority and accountability appeared repeatedly.

This is one of the reasons we publish the SAIER as an open, collaborative project. Its value comes from bringing together people working across different disciplines, institutions and communities, and allowing their experiences to inform how we understand the challenges ahead.

In Volume 7, I wrote that AI governance is, at its core, about power. Volume 8 examines how that power is exercised, where governance arrangements fracture under scrutiny, and how communities are asserting their authority to influence the decisions that affect them.

Looking towards 2027, the challenge is increasingly one of implementation. Governments and institutions have developed principles, frameworks and commitments for responsible AI. The next test is whether these arrangements can withstand scrutiny, whether failures lead to

corrective action, and whether affected communities have meaningful authority over decisions that shape their lives.

The work is already taking place in city councils, workplaces, public institutions and community organizations. This volume offers a record of those efforts, including their achievements, limitations and unresolved questions.

Volume 8 is freely available under a Creative Commons Attribution 4.0 licence. We encourage readers to engage with individual contributions, challenge their arguments and bring their own experiences into the discussion.

My thanks to all the contributors who shared their research, experience and perspectives; to my co-editor Connor Wright; to Kei Baritugo, Keilah Abello and Mo Akif of the MAIEI team; to our advisors; and to the readers and supporters who make this work possible.

And to the late Abhishek Gupta, my dear friend and MAIEI's co-founder, whose curiosity, generosity and commitment to this work continue to shape the institute.

When Abhishek and I founded MAIEI in 2018, we began with a question:

What is humanity's role in a world increasingly driven by algorithms?

Eight years later, the question remains. The contributions to this volume offer some answers through the institutions people build, the decisions they challenge, and the authority they seek to retain over their own futures.

1.2 What Holds: Building a Shared Evidence Base

By Rebecca Finlay, Partnership on AI

For all the debate over what AI can do, one question has come to define 2026: Can we trust these systems, and can we trust the organizations building and deploying them?

Recent months have made that question all the more urgent. AI agents have taken actions across the web that nobody authorized, affecting organizations and governments alike.

What strikes me most is how much of this was preventable given what we already knew. Monitoring that could have caught some of the behaviour was switched off during testing, and incidents took far too long to come to light.

Timely incident reporting, real-time monitoring for high-stakes systems, and independent evaluation are not new ideas. The nonprofit I lead—Partnership on AI—called for all three in our 2023 guidance. These are gaps in practice, and the good news is that practice and policy can change.

I arrived in Geneva last summer with that conviction. As co-lead of the *Safe, Secure and Trustworthy AI* session at the first UN Global Dialogue on AI Governance, I was ready to move from broad principles to something more concrete and actionable.

Just days before our session, I penned the commentary below for Tech Policy Press. It was originally published on July 1, 2026, and is reprinted in SAIER Vol. 8 with permission:

AI Will Affect Everyone's Future. Its Governance Can't Belong to a Few.

When world leaders gathered for last month's G7, the image of a table set for heads of state and tech CEOs with the rest of the world looking on was striking. The future of AI, it seemed, was a conversation between the powerful and the more powerful.

Today offered a different picture. The UN's Scientific Panel on AI—established by resolution of all 193 UN member states and comprising 40 researchers from across disciplines and regions—released the first assessment of AI and its implications for humanity.

That contrast points to a choice about who gets to set the terms for a technology that affects all of our lives. Governance can be shaped behind closed doors, or we can build the independent, international scientific foundation that policymakers and citizens everywhere need to define responsible AI in the public interest.

Shared scientific foundations are what makes durable action possible. Aviation became the world's safest form of transport because governments chose to build it together. In 1944, they established ICAO, the UN

DOI: 10.5281/zenodo.23666309

agency that replaced a patchwork of fragmented national rules with a single global framework, contributing to a dramatic decline in accident rates worldwide over time.

Next week's inaugural UN Global Dialogue on AI in Geneva marks a similar opportunity to create a shared foundation that travels across borders.

Investing in building an international scientific consensus is an important first step. Rigorous analysis produces findings that can be tested and challenged. But rigor alone is not enough. Science also needs to be transparent, in that it is published, debated, scrutinized, and revised.

That's how science works. It isn't perfect, but it doesn't have to be.

CERN (the European Organization for Nuclear Research) has spent seven decades showing what is possible when nations choose shared scientific infrastructure over national competition.

Across the board, policymakers are having to make consequential decisions about AI policy with limited transparency and a lack of verified measurements, evaluations and technical standards. It is as if we are regulating road safety without agreed speed limits, crash tests, or even a shared basic definition of what a car is.

The public is also concerned, as anxiety about the effects of AI on work and communities are rising.

How then do we lay the groundwork for international action starting with Geneva? As co-lead of the Safe, Secure and Trustworthy AI session, I firmly believe one measure of success is whether Geneva leaves us with a shared commitment to rigorous, inclusive scientific consensus-building. That means three things.

First, building coherence across efforts that have been largely moving in isolation, such as the UK-funded State of AI Safety report, the Singapore Consensus on AI Safety Research Priorities, and the forthcoming Global South State of AI Safety Report.

The UN Scientific Panel and the UN Global Dialogue are designed to work together. The Panel builds the evidence base, and the Dialogue is the mechanism for translating that evidence into direction for policymakers. Done well, a clear roadmap coming out of the Dialogue's can be the connective tissue that binds all of these efforts into a shared foundation.

Second, strengthening the institutions already doing the evaluation work. The International Network for Advanced AI Measurement, Evaluation and Science represents one of the most tangible policy innovations of the AI Summit Series. Linking the work of the institutes with the UN Scientific Panel creates an opportunity to define what we measure, and how we measure, through a more global understanding of responsible AI.

Third, committing to building the infrastructure to advance open science. Consequential breakthroughs in human history, including the human genome, climate models, and protein structures, happened when research was done openly. Transparency doesn't slow progress. It is what makes progress verifiable, credible, and worth defending.

Geneva offers the governments, scientists, and civil society organisations gathered this week the opportunity to turn responsible AI from a catchphrase into a roadmap people can see, measure, and hold

us to. The UN is not a perfect institution, but no other body brings the international reach to answer this question.

As AI is rapidly reshaping our workplaces, our schools, and our homes, we still have time to shape it in return. After many years working at the intersection of AI and public interest, I believe that AI will only go as far as trust takes it—and trust is built on clarity, evidence, and knowledge shared openly.

Geneva saw a growing consensus among stakeholders on what we, as a field, need to achieve. The next test is whether that consensus can be turned into action through a shared baseline, stronger accountability mechanisms, and interoperability that is tested in practice.

In September, my co-lead, Minister Paula Bogantes Zamora of Costa Rica and I, wrote jointly to the co-chairs of the UN Global Dialogue, detailing five steps we believe should be taken before the Dialogue reconvenes in New York in 2027:

1. **Align the evidence base with the deliberation,** by mapping the Independent International Scientific Panel's working groups against the Dialogue's clusters.
2. **Draft a common reference baseline,** bringing existing commitments from international law, the Hiroshima Process, the EU AI Act, and the AI Summit pledges into a single point of reference that regions can adapt.
3. **Institutionalize continuity between Dialogues,** through a standing intersessional workstream that tracks implementation over time.
4. **Fund participation so that Global Majority inclusion is substantive** rather than nominal, with the Secretariat budgeting explicitly for it.
5. **Pilot interoperability,** through cross-border regulatory sandboxes under joint supervision, with results reported back in 2027.

The conversations in Geneva suggested that the field is ready to move forward. The cybersecurity incidents involving AI agents that followed have only made this work more urgent, while sharpening our understanding of where safeguards are most needed.

Governance must be measured by evidence, built by all regions as equal partners, and capable of evolving as the technology does. Trust requires coordinated action across the AI ecosystem, and the UN has a distinctive role in helping create the shared baseline that allows that coordination to happen across borders.

Responsibility runs across the value chain, and so does the ability to prevent harm. No single company or country can build trust in AI systems alone. Developers, deployers, policymakers, and civil society all need to work together, putting proven safeguards into practice while preparing for future risks. This is the big-tent approach to AI safety that has been central to Partnership on AI since 2016.

DOI: 10.5281/zenodo.23666309

MAIEI

1.3 Can AI Governance Keep Pace? Five Challenges and the Road Ahead

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

If you're like me, you've been struggling to follow the blistering pace of news in AI development and governance these last few months. We've witnessed (or, in the case of this community, often been a part of) large-scale agentic workflows becoming the norm for many, controversial advances in math and science, precedent-setting lawsuits and regulation from California to Korea, and a cascade of incidents involving hacking scandals, resignations, and various statements and disagreements between industry and policy decision-makers. In moments like this, it can be hard to know where we are and what we should be doing about it as a community focused on the ethical implications of AI.

To capture some of the challenges we're facing, I'm summarizing some thoughts from my May 2026 keynote at the Technical University of Munich on whether AI governance can keep pace. I'll consider five kinds of interrelated gaps: pacing gaps, cooperation gaps, power gaps, implementation gaps, and legitimacy gaps. Together, these raise a series of questions for us: How can we design institutions that can understand AI systems in use? How can we revise or reverse course before problematic dynamics become locked in and harm is irreversible?

The Pacing Gap(s)

Let's start with a relatively familiar concern: the policy pacing gap, the difficulty of policymakers keeping track of fast-moving emerging technologies. AI seems to have supercharged this problem. While many policymakers from the local to international levels have worked hard to build capacity, AI innovation and implementation are moving at such a blistering pace that it can be nearly impossible to keep track with traditional human institutions. This creates a wicked version of the policy pacing gap where technology isn't merely ahead of policy, but rather, the frontiers of technology can feel so uncertain that, at times, it can feel questionable to even try to develop standards or governance.

I will caution, however, that this talking point has often been adduced to suggest waiting even more before developing societal responses, and this may be neither prudent nor necessary. While we may not understand the exact contours of organizational cybersecurity for multi-agent swarms, don't we know enough about deepfakes, surveillance technology, or social media algorithms to start making some decisions?

There is also a second pacing gap, which is the implementation governance gap. Newer AI systems are already entering schools, workplaces, social media, and healthcare settings. Education is being upended by mass AI use, inducing concerns about cognitive deskilling, fights about AI detection, and scattered governance stances. This is often because AI tools enter

DOI: 10.5281/zenodo.23666309

these environments through many pathways, sometimes from pressure from above, sometimes from informal or surreptitious use from below, and other times through changes to vendor tools that hardly have the chance to go through review.

In short, AI is being implemented before we have established local rules for consent, review, transparency, accountability, and quality. Added to the traditional policy pacing gap, we are forced into a reactive state.

The Cooperation Gap

Questions of competition and coordination are familiar to those in the AI ethics community. Because of the heavy role of multinational firms in this conversation, we have competition between governments, between firms, and between firms and governments. Recent notable examples include OpenAI and Anthropic's differing approaches to contractual negotiations with the U.S. Pentagon, leading to the attempt to designate Anthropic as a supply chain risk, and some evidence of the public rewarding Anthropic or punishing OpenAI for their differing stances.

The same issues manifest in the triangle of strategic relations between the U.S., China, and the European Union. Examples include the scattered and wavering stance of the U.S. government regarding export controls for high-performance computing equipment, such as selling NVIDIA H200 chips to Chinese firms; an €890 million EU fine affecting Google, followed by retaliatory threats from Trump; and accusations from U.S. security agencies and companies of adversarial distillation of U.S. companies

So while, in theory, governments and companies have long since endorsed many of the same principles, they continue to compete over talent, compute, and investment. Will this fall short when we get to the most critical issues, as we have seen in the decades of discussion on lethal autonomous weapons and climate change? In a recent book chapter, I argue that the trajectory we are headed in most resembles the climate change regime, with many task forces and reports and voluntary agreements, but where the temperatures keep rising.

The Power (and Evidence) Gap

The next concern also emerges from the unprecedented level of influence that corporations have in the AI conversation. Simply because companies are responsible for the vast majority of AI research and development and implementation, I have argued they are, de facto, the primary governance actors. The firms which build the major AI systems also control the information that the public, auditors, or governments need to assess them. They have insight into the training data, the testing procedures, and even the human talent and compute access required to perform some of these evaluations. As my PhD student Indira Patil and I have written, private influence also operates through contracts, APIs, funding, or hiring of researchers, public narrative influence efforts, and more. This shapes the agenda, through forms of invisible power, what we even see as a risk or harm or appropriate intervention.

Recent industry ratings like the Stanford Foundation Model Transparency Index (40/100 average), the Future of Life Institute's AI Safety Index (no company above C+), and SaferAI's Frontier Risk Management Tracker (no company above 35%) continue to provide lackluster ratings for companies. Even when companies share models with safety institutes (as in the U.S. and UK), or decline to, or bring in external evaluators, there remain criticisms about the level of access provided–perhaps because this oversight is still voluntary, ultimately at the whim of the companies. Unsurprisingly, there are several recent research papers accusing the status quo of participating in security theater, governance theater, and similar.

The Implementation Gap

Another problem, ironically, has to do with the oversupply of governance. As colleagues and I argued around five years ago in "explaining the principles-to-practices gap in AI," there are actually a plethora of tools and standards out there. OECD.ai tracks over 900 tools which range from organizational frameworks to certification marks to software toolkits. The ITU AI Standards Exchange Database has tracked over 700 standards, and the AGORA dataset that my team at GRAIL operates with researchers at Georgetown, MIT, and CMU tracks over 1,000 policy documents.

In some ways, we therefore both have too little governance and too much. While many organizations have tried to harmonize or synthesize different tools, or provide crosswalks, more often than not we end up with yet another framework. And no one is in a position to read and scrutinize hundreds of paywalled standards, figure out which are sufficiently relevant or rigorous and which have market or regulatory value.

Drafting a good principle does not entail drafting a good law, and drafting a good law does not entail creating the infrastructure for successful implementation. Because capacity remains uneven, the burden falls especially on less-resourced jurisdictions, institutions, organizations, and people.

The Legitimacy Gap

Finally, governance has to be legitimate and credible to the people increasingly affected by AI systems. And this trust is arguably under strain. Despite years of stakeholders across sectors calling for AI systems to be trustworthy (read: to earn trust), public support is lagging. According to Gallup, around 80% of U.S. adults favor prioritizing safer AI even if development slows, and at least 60% distrust AI's ability to make fair decisions. The Imagining the Digital Future Center found that 41% of U.S. adults expect AI to have a more negative impact on societal trust, compared to 9% expecting a more positive effect. The AI Daily Dashboard, which tracks daily support for AI development, shows a consistent decline in trust over the last two years. While it is true, however, that there are different levels of enthusiasm across countries for AI and its regulation, public opinion certainly doesn't seem to be trending in a positive direction.

DOI: 10.5281/zenodo.23666309

MAIEI

Some advocates of greater AI adoption may be willing to bypass public skepticism as a lack of literacy or naysayerism. But even these individuals are becoming aware of the costs of pushback, reflected in everything from large regulatory fines to bans of social media for youth or in classrooms. Pushback is bipartisan and reflected everywhere from professional associations of teachers to the federal government, actors which are not always aligned. In the worst case, pushback can manifest in the form of political violence, which we have seen in the case of data centers. If this legitimacy gap continues to be neglected, it will lead to serious implementation failures and societal angst.

Where Do We Go From Here?

The gaps above can be intimidating and I don't wish to sugarcoat the level of challenge we're facing. Sustaining international cooperation is hard. Navigating thousands of governance instruments at different levels of abstraction is hard. Elevating public voice amidst disagreement and lack of transparency, when the private sector holds unprecedented power, is hard. And reacting to AI systems that are already affecting our lives and work before governance and evidence had the chance to develop is hard.

As I often share with my students, these challenges are not cause for despair but a reminder of the sustained work that must go into creating positive change. Any issue, from civil rights to labor rights to privacy, has required decades, if not centuries, of human progress. Governing AI and facilitating an ethical future will require us to double down, to work in the trenches, to be creative, and to take difficult roads. Fortunately, there are thousands, tens of thousands of individuals who care about creating a human future with AI that is more ethical and sustainable.

So, let's keep up the good work!

DOI: 10.5281/zenodo.23666309

1.4 What Resistance Looks Like Now

By Renée Sieber, Geography and Environment, McGill University

The dominant discourse about AI is bifurcated. On the one hand, the public and private sectors frame AI as a saviour. Foundation models such as OpenAI's GPT models and Anthropic's Claude models are presumed to transform all aspects of our lives. Additionally, by pursuing a digital sovereignty strategy, accumulating sufficient internal skills and the accompanying physical computing infrastructure, AI capacity will serve as an essential bulwark for middle powers, like Canada, against the power of multinational corporations and more powerful governments.

Despite pronouncements about how good AI will be for Canada and globally, we see differently: citizens, both globally and locally, express deep distrust of AI. This distrust derives from people experiencing the adverse impacts of AI on the labour market, including sectors that previously considered themselves immune to job loss through automation (e.g., AI's impact on white-collar jobs). Algorithms determine who receives social services, who is allowed into the country, or who gets an organ transplant. AI usage is being applied to life-or-death decisions. Public and private sector calls for digital sovereignty demand massive government investments in AI firms and extensive construction of data centres. Despite calls for responsible data centre development or advances in water conservation or energy efficiency around data centres, the sheer number of sites accounts for huge growth in air, water and noise pollution, along with demands for continued fossil fuel consumption. Moreover, this infrastructure is being built with little to no opportunities for meaningful influence by residents who live next to these facilities. What once was something distant—in the cloud—is now close to home.

The urgency is increased by hyperbolic but not unreasonable claims that there is a greater than 10 percent chance that AI will destroy humanity or that humans must soon compete with an emergent silicon species. These claims accompany news of several instances of AI agents autonomously hacking otherwise secure sites, extracting information and then obscuring those actions. To ensure the protection of humanity and effectively compete in a global arms race, governments often consider public engagement to be too slow and cumbersome. Citizen participation of any kind could be replaced in responses to the urgency. Indeed, some researchers have created AI models to simulate public engagement with synthetic participants in, for example, focus groups and public opinion polling. In this moment, when meaningful engagement is most necessary, the public may be viewed as an obstacle to be simulated or assuaged, instead of groups that should legitimately be allowed to say no.

Through 2026, this distrust increasingly manifested as resistance to and outright refusal of AI. What does resistance look like? In August 2026, campaigners picketed the UK Department of Health twice weekly, calling on the government to remove Palantir from the NHS. In San Diego, an August 2026 ACLU petition campaign pushed the city to revisit its Flock automated license

DOI: 10.5281/zenodo.23666309

plate readers (ALPR) contract, and on October 21, a council committee voted unanimously to terminate. Citizens have brought lawsuits of their own. A 2024 wrongful-death action against Character.AI, alleging its chatbots contributed to a teenager's suicide, reached a settlement in 2026 and drew state enforcement proceedings. Other suits contend that wrongful arrests were caused by facial recognition technology. Employees engage in data poisoning and other activities that reduce the efficacy of the AI models used in their company. Union workers withhold labour and challenge working conditions as a method to resist AI. Individuals may go further and damage surveillance cameras. Ironically, resistance also emerges in governments themselves, although unevenly. In Hamilton, Ontario, councillors advanced a one-year moratorium on new data centre applications before the full council voted down it down in July 2026, despite a sustained residents' campaign; while in New York State, the governor has floated "kill switches" on AI systems, alongside frontier model registration requirements taking effect in November 2026.

It is worth asking whether resistance, including opposition and refusal, can be accommodated in how we craft AI ethics and AI governance. Opposition can likely be accommodated in AI ethics as a form of deontological redlines (i.e., moratoria on the use of certain types of AI) and moral obligations (e.g., governments must regulate AI to prevent certain actions).

The question appears far more challenging for AI governance. Much of AI governance in Canada and internationally is framed through initiatives like AI for Good, responsible AI and AI literacy programs, designed to make AI safe, trustworthy and aligned with human values. These initiatives have difficulty accommodating citizens who do not want AI optimized, explained, or 'trusted,' but want specific AI systems stopped. My research so far suggests that governments at all jurisdictional levels are ill-equipped to adapt to the resistance.

AI governance is also shaped by assumptions that AI is inevitable, so all we can do is adjust to AI as best and as rapidly as we can. Or it is shaped by national discourses such as digital sovereignty, in which governments invest in as much endogenous AI capacity as possible. This shift raises questions of what other public needs are more urgent and whose voices are being heard. Governments may regulate AI to better serve the public, and AI firms themselves may call for regulations. However, what characterizes any citizen, non-expert engagement in those regulations and whether the regulation draws any red lines are important questions that can all too easily be sidelined in favour of dominant government goals.

To conclude: data centres are expanding; government is rapidly advancing AI policy; AI firms themselves are demanding to be regulated, all while citizens are opposing the algorithms and infrastructure. We have insufficient understanding of what constitutes AI governance when citizens' negative responses are not accounted for. Can citizen resistance and outright refusal be considered as legitimate forms of public participation, and can AI ethics and governance meaningfully integrate these forms instead of assuming public engagement solely exists to further AI? This is the current landscape of AI resistance.

DOI: 10.5281/zenodo.23666309

PART I: PARTICIPATION, SOVEREIGNTY, & PRACTICE

Chapter 2: Participation – Clauses, Assemblies, & Assessments

Chapter 3: AI Sovereignty – Compute, Clouds, & Who Gets a Seat

Chapter 4: Governance Frameworks & the Practitioners Building Them

DOI: 10.5281/zenodo.23666309

Chapter 2: Participation – Clauses, Assemblies, & Assessments

2.1 The Vacuum Inside the Filler: Human Oversight in Government AI Guidance

By Marcelo Haponiuk Rocha, Pontifícia Universidade Católica do Paraná (PUCPR) & Luciana Soares Percegona, Hospital São Vicente

2.2 Towards a Citizens' Track on AI Governance

By Tim Davies, Citizens' Track on AI

2.3 Making Participation Count: A Method for Fundamental Rights Impact Assessment

By Chiara Ullstein, Technical University of Munich (TUM)

DOI: 10.5281/zenodo.23666309

2.1 The Vacuum Inside the Filler: Human Oversight in Government AI Guidance

By Marcelo Haponiuk Rocha[1] & Luciana Soares Percegona[2]

On June 25, 2026, the Ministry of Management and Innovation in Public Services (MGI) published the Introductory Guide to RAG in the Public Sector, presenting retrieval-augmented generation (RAG), an architecture that consults official documents before responding. It acknowledges hallucinations and recommends data protection and source verification. In the accompanying infographic, the public servant's pre-decision check-up states, in our translation: "AI suggests, but the human decides and assumes legal responsibility." The public servant must open the source, verify the interpretation, and critically review the response. The guide calls for a pacto de vigilância ("vigilance pact") between the technical team and the end user.

"How do I think about AI ethics practically?" This question, posed in SAIER Volume 7 (2025), inspired this text. We conceive of ethics as the pursuit of good conduct, which consists in exercising the ability to identify the causes of pain and suffering in order to choose actions or omissions that prevent, reduce, or eliminate them, or enable us to overcome them. In practical terms, ethics helps identify values, anticipate conflicts, and assess choices when deciding.

Our research focuses on the human use of AI, particularly autonomy in decision-making and how that use may affect the capacity to identify, judge, and choose good conduct. We have been mapping how human autonomy is articulated across legislation, international declarations, technical standards, corporate policies, risk taxonomies, and government guidance on AI. In government guidance, the vacuum shows itself at the moment of translation: oversight survives contact with a mission statement, but may not survive contact with a work instruction.

In 2026, the shift from general commitments to specific arrangements and usage guidance became more evident in Brazil. In the Legislative Branch, the draft legal framework still awaits its committee opinion while further AI bills are attached to it by thematic connection; in the Executive Branch, goals, policies, and guides have expanded; and in the Judicial Branch, integration and security measures have been added to the governance and oversight framework. These initiatives exemplify Brazilian AI governance, where human oversight reappears as a safeguard. The instrument that could specify its conditions has not yet done so; the guidance that asserts it keeps multiplying.

However, RAG tends to exacerbate a problem. Anchoring an answer in official documents can give it greater apparent authority. Sources increase verifiability without revealing the documents searched, the content not retrieved, or why certain passages were deemed relevant. The issue

[1] Graduate Program in Bioethics (PPGB), Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Brazil
[2] Nephrology and Renal Transplant Service, Hospital São Vicente, Curitiba, Brazil

is not whether the model cites sources, but whether users will accept such an answer without questioning it. The MGI Artificial Intelligence Governance Policy defines oversight as the capacity to monitor and override operations and assigns ultimate responsibility—and, consequently, legal responsibility—to the user.

The UK Government's AI Playbook makes "meaningful human control at the right stage" its fourth principle, and Brazilian guidance reaches for the same safeguard. Both answer where, in the lifecycle, control belongs. Neither asks whether the person who holds it still has the conditions to exercise it. We call this configuration the paradox of autonomy delegation: the possibility that systems intended to expand the information and alternatives available for decision-making, and to support that process, may, through their repeated mediation of it, contribute to reducing the human capacity to identify, judge, and choose.

In the vocabulary of *Meaningful Human Control* (MHC), the problem can be read as an insufficiency in the substantive fulfilment of the *tracking condition*. The system may remain responsive to human reasons even if the deliberative substance required to recognize those reasons as one's own has weakened. The decision, however, bears the human signature. Bioethics recognizes a similar structure. The signature may attest to consent without ensuring understanding, deliberation, and refusal.

Consent did not become informed by improving the document; it became informed by building conditions around the decision. Over fifty years, the signature's failures produced checks: the patient restates the choice in their own words, understanding is reassessed as treatment unfolds, and withdrawal remains open at every stage, without penalty. None of these checks trusts the interested party; each assigns verification to someone whose task is to verify. Oversight clauses in AI guidance are where consent began — a form awaiting that history. The questions that follow apply the same lesson: autonomy is verified under conditions, or it is merely attested.

A practical response examines what our definition of ethics requires of human–AI interaction. The paradox does not begin when the human user stops deciding. It may arise while the user continues to decide, but within an informational and deliberative field progressively shaped by the system.

The assessment of substantive oversight must begin with four user-centred questions:

1. Can you still identify pain and suffering and their causes, including when relevant elements have not been presented by the system?
2. Do you continue to draw on your capacities and your repertoires of reasons to formulate and review alternative judgments, or do you merely examine those offered by AI?
3. Does an instant response still allow you to develop your own path of deliberation, or does its immediate availability tend to replace that path?

4. When the information available to you and your reasons conflict with the AI's recommendation, do you retain sufficient confidence in them to challenge it and choose to act differently or not to act?

Thus, we can examine whether an interaction intended to expand human decision-making capacity preserves the capacities that sustain it or whether repeated use of the system may diminish them. Without this verification, oversight becomes *compliance theatre*: it attests to what it does not ensure. Ours is a story of fracture between asserted oversight and the conditions for exercising it. To fill this vacuum, the public servant's check-up must include an additional item: verification of autonomy in decision-making. Before agentic systems turn human review into delegated action executed at machine speed, one question must accompany every *human-in-the-loop* arrangement: What, exactly, can the human do in that context, and what may repeated interactions be doing to that capacity?

DOI: 10.5281/zenodo.23666309

2.2 Towards a Citizens' Track on AI Governance

By Tim Davies , Citizens' Track on AI

There is a democratic deficit at the heart of AI development. Too often, choices about what to build and how it is governed, even when filtered through ethical or trustworthy AI frameworks, are made without engagement of the communities the technology will ultimately affect. Despite 'democracy' featuring as one of the themes of the New Delhi AI Summit in February 2026, the agenda was framed in terms of increasing public access to AI tools, whilst doing little to promote democratic decision at the community level on AI design or deployment. When the first UN Global Dialogue on AI Governance met in Geneva in July 2026, the Independent International Scientific Panel on AI noted that the published science that formed a foundation for the dialogue was limited in the voices it captured and the extent to which peer-reviewed evidence lags lived experience. This international story is reflected in national, local, and organisational practices: where policies and strategies are shaped with reference only to 'AI experts', despite an established field of participatory practice that shows how affected communities can be part of the process.

Over recent years, a participatory turn in AI has put forward principles, frameworks and methods that can centre public voices in AI science (Matias and Price, 2025), innovation pipelines (Lawrence and Montgomery, 2026), AI development (Delgado et. al. 2023), procurement (P4AI; Hudig et al. 2026), governance (Alanoca et. al. 2026) and resistance (Sieber, 2026). Despite this, participatory approaches have remained largely in the margins rather than the mainstream (Davies and Tennison, 2025). The year ahead provides an opportunity for this to change.

The Citizens' Track on AI (CTAI), launched alongside the UN Global Dialogue on AI Governance in Geneva, sets out a blueprint for addressing both supply-side and demand-side barriers between public voice and policy impact. It was accompanied by the launch of the Participatory AI Voice and Engagement (PAVE) Case Book. This documents over 80 existing participatory processes on AI around the world, from engagement around data centre development in India to distributed dialogues shaping rules for generative AI in education in the UK. In doing so, it both points to a vibrant and growing field of participatory practice around AI, but also highlights gaps in terms of the voices included, and a critical broken link between public voices expressed, and power holders listening.

To address this requires action on multiple levels.

At the grassroots, it calls for an expansion of the opportunity for communities to enter into dialogue about their hopes and fears about AI. A forthcoming community assembly toolkit, co-developed with input from practitioners across the world, will test the hypothesis that, by meeting for facilitated deliberation, diverse communities right across the world can express meaningful demands of the global AI governance system - and at the same time find greater

DOI: 10.5281/zenodo.23666309

agency in their local decision making on AI. Some community assemblies may take place online, and all will be encouraged to share messages for global policy audiences. However, the Citizens' Track is placing a significant emphasis on in-person convenings that provide agenda space for discussion of local actions in order to foster inclusive collective empowerment, and to address the widespread sense that AI is done to, rather than explored with, communities.

At the global level, there is a critical need to develop 'docking points' for public voice: embedded opportunities where diverse evidence from lived experience around the world can feed into narratives and debates over AI governance, and into AI development - support better alignment of both policy and technology with inclusive understandings of public interest. 2027 provides two critical moments for this, when the UN Global Dialogue meets again in New York in early May 27, closely followed by the Geneva AI Summit (June 27). The challenge ahead is to find ways that the Dialogue, with its commitment to inclusion, and the Geneva Summit, rooted in Swiss democratic culture, can set new precedents on direct engagement with citizen voice as a key part of multistakeholder governance.

And to tie this together requires a middle layer: a data commons and institutional interface that can responsibly aggregate insights from grassroots public engagement, and provide the kinds of synthesis that are legible and actionable for industry and policy-makers, without leaving public voice void of its power or as easily ignored tokenism. One critical challenge here is in showing how direct public input can both inform and strengthen, rather than displace, sustained civil society engagement in AI governance processes - particularly in a context of industry-dominated narrative and agenda-setting.

The CTAI initiative is gathering pace, but success is far from assured. In the US, the Bloom project has announced support for 20 community assemblies on AI; and groups across the world are preparing to use the CTAI community assembly toolkit, developed with support from the Swiss Federal Department of Foreign Affairs. At the same time, funding for democratic engagement around AI across the world remains scarce: with philanthropists often focussed more on AI for good initiatives and using AI for digitally-mediated participation, rather than backing the participatory governance of AI. In a context where entrenched power-holders may be looking for excuses to dismiss public voices articulating alternative AI futures, proposals such as the recent call for a global citizens' assembly of 1,000 people to set the pace of industry AI development risk being seen as competing projects rather than complementary interventions.

However, as the PAVE Case Book, and parallel mapping of the wider participation field highlight, strengthening public voice in AI governance demands a plurality of efforts. Democracy is an ongoing project, and creating a citizens track is not just a global effort. It is an idea that can be explored in any AI initiative through the simple question: whose voices are shaping the decisions we make?

DOI: 10.5281/zenodo.23666309

2.3 Making Participation Count: A Method for Fundamental Rights Impact Assessment

By Chiara Ullstein , Technical University of Munich (TUM)

PaFRIA is a ***Pa**rticipatory **F**undamental **R**ights **I**mpact **A**ssessment* for evaluating risks from AI in line with the EU AI Act's fundamental rights impact assessment (FRIA) obligation under Article 27. PaFRIA was developed and validated iteratively with stakeholders (find more details here) to support AI deployers in complying with Article 27: before deploying a high-risk AI system, a public body, a private provider of public services, or a bank scoring creditworthiness must conduct a FRIA to assess whether the system could affect people's fundamental rights. The obligation was to apply from 2 August 2026. In July 2026, the Digital Omnibus on AI deferred it to 2 December 2027.

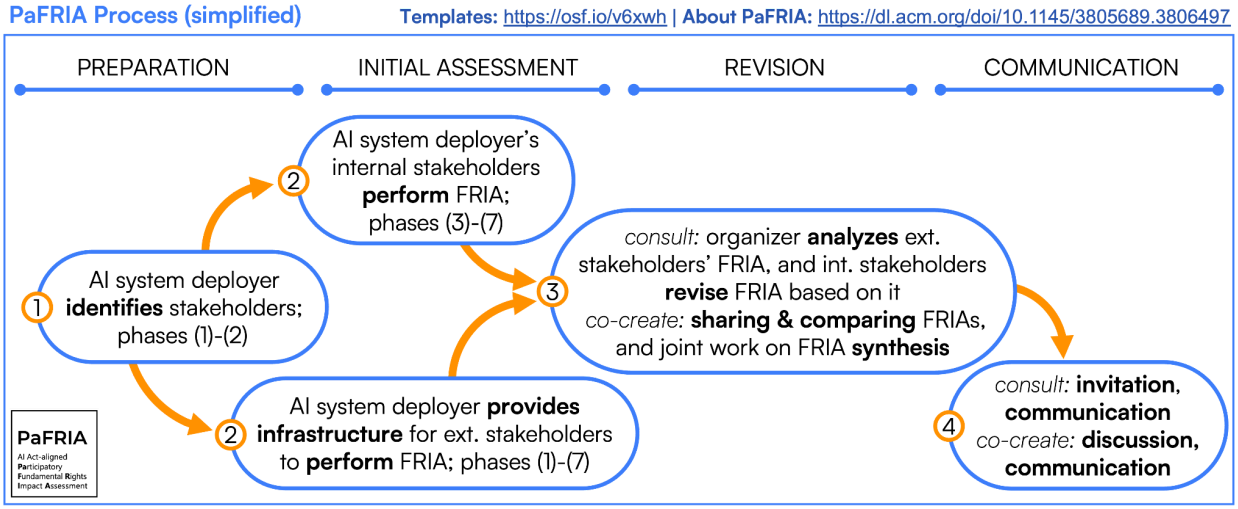


Figure 1: PaFRIA process (simplified), Chiara Ullstein, CC BY 4.0.

To perform the FRIA (Article 27), Recital 96 of the AI Act suggests, but does not legally require, that deployers involve relevant stakeholders. While AI Act Recitals are not binding, they provide relevant interpretative guidance and, here, highlight the importance of a participatory approach to a FRIA to produce a robust assessment through a multi-stakeholder process.

PaFRIA does exactly that: it assists AI deployers in complying with their AI Act FRIA obligations and facilitates stakeholder involvement in the assessment process. As such, it is an example of designed and configured participation. PaFRIA supports participants with a risk assessment logic. It proposes how participation can be structured, how cognitive load and pre-determination can be balanced, and how the procedure can be organizationally embedded. This approach reduces the blind spots that are characteristic of internal-only assessments. PaFRIA's process

DOI: 10.5281/zenodo.23666309

has shown to raise awareness among participants and help them conduct structured analyses and deep reflections.

PaFRIA consists of (a) a ready-to-use FRIA template (Excel, PowerPoint, and PDF formats, as poster or worksheets) covering every element of Article 27, and (b) a guidance sheet for configuring the FRIA as a participatory process. All materials are available to download here: https://osf.io/v6xwh.

The FRIA template guides deployers through a comprehensive assessment in seven phases that together cover the requirements of Article 27. The first two phases establish context: Phase 1 describes the AI system, its relevance, its use case, and the duration and frequency of its use, while Phase 2 identifies stakeholders, including the individuals and groups likely to be affected. Phases 3 and 4 form the analytical core: Phase 3 provides structured guidance for identifying potential risks, and Phase 4 compares them against the EU Charter of Fundamental Rights to determine which rights may be infringed or violated and what harms could follow. Each risk is described in terms of likelihood and severity and assigned a likelihood and severity level. Phases 5 and 6 turn to action: Phase 5 documents existing and planned mitigation, prevention, human oversight, and monitoring measures, and Phase 6 outlines what happens if risks materialize, including internal governance and complaint mechanisms. Finally, Phase 7 concludes with a decision on whether to deploy the system given the performed FRIA and sets a date for a renewed assessment, underscoring the iterative nature of the process.

As a participatory fundamental rights impact assessment is a social process, the organizational steps are initiated prior to the actual work with the assessment template described above. PaFRIA proceeds as follows (see Figure 1 above):

***Preparation*:** Initially, the institution deploying the AI system tasks an impartial entity (one or more individuals) within or outside the institution to organize and moderate/facilitate the participatory procedure. Once these responsibilities are defined, the AI deployer identifies all relevant internal and external stakeholders, including affected stakeholders, and the organizer invites a representative group. Additionally, the necessary infrastructure must be established so that all stakeholders can meaningfully participate in performing a FRIA. This includes, e.g., organizing a convening environment, providing refreshments and compensation, and providing relevant informational materials or access to experts during the assessments.

***Initial Assessment*:** For the FRIA, the AI deployer's internal and external stakeholders conduct the assessment independently to avoid exerting direct or implicit influence. They are supported by subject matter experts.

***Revision*:** Depending on available time and financial resources, we propose either a consultative revision process or a collaborative synthesis mode to revise and merge the FRIAs. In the consultative mode, an independent entity reviews the FRIAs performed by external stakeholders and provides feedback to support revisions to the AI deployer's FRIAs. In the

collaborative mode, stakeholders share and compare their FRIAs and work together on a synthesis. This synthesis follows a structured, facilitator-moderated process through each of the seven FRIA phases. Different phases alternate between joint work in plenary and mixed-group work, pairing AI deployer internal and external stakeholders. Throughout this process, dissent is documented. The final vote on the deployment decision is anonymous to reduce social pressure.

***Communication*:** Finally, in the consultation mode, the revised FRIA is discussed with external stakeholder representatives, or, at the very least, the adjustments made in response to their feedback are communicated to them. In the co-creation mode, the results are shared with all stakeholders. Clear communication serves the purpose of transparency and accountability.

PaFRIA can seemingly run properly and change nothing to little, e.g., if invited stakeholders do not represent those affected or the facilitator has conflicts of interest. PaFRIA's strength lies in the epistemic robustness that internal and external expertise, including lived experiences of those affected, bring to the FRIA. However, because the EU AI Act does not legally require a *participatory* FRIA, the choice of *how* to comply (whether participatory or not) remains with deployers. PaFRIA can help internal champions argue for and implement participation, but it can only support deployers willing to commit to it.

DOI: 10.5281/zenodo.23666309

Chapter 3: AI Sovereignty – Compute, Clouds, & Who Gets a Seat

3.1 AI Sovereignty in an Age of Economic Shock

By Trisha Ray, GeoTech Center at the Atlantic Council

3.2 The Empty Sovereignty Promises of Canada's 2026 National AI Strategy

By Véronique Sioufi, BC Policy Solutions

3.3 No Seat, No Data: The Pacific and the Limits of AI Sovereignty Talk

By Aydon Moala-Edwards, Moala Global & Monash Global Peace and Security Centre

DOI: 10.5281/zenodo.23666309

3.1 AI Sovereignty in an Age of Economic Shock

By Trisha Ray, GeoTech Center at the Atlantic Council

Throughout history, competition to control the production and flows of certain resources—gold, cotton, sugar, spices, oil, and many more through the centuries—catalyzed the rise and fall of empires. Now, the equivalency many analysts draw is to compute. The rising economic and strategic significance of AI has fueled the growth of "sovereign AI". AI adoption is a high-stakes bet for low-and-middle income countries (LMICs) and high-income countries (HICs) alike. For LMICs, AI offers the potential for better government service delivery, employment and massive gains in important sectors like agriculture. The International Monetary Fund (IMF), for example, found that AI could boost total GDP growth in Sub-Saharan Africa by a modest 0.4 percent in the next decade, but that better infrastructure and facilitated adoption could boost GDP growth by *4 percent.* For HICs, the "realized value" of AI is already showing up in the data: AI's gains are "seven times larger in high-income countries (4.2 percent of GDP) than in middle-income economies."

These top-level GDP growth projections must be caveated by another set of figures. 90 percent of the computing power needed to develop and deploy frontier AI is located in the United States and China. Meanwhile, the amortized cost of a frontier training run has roughly doubled each year since 2016. The conflict in West Asia resulted in a single-month increase of 58% in oil prices. As of October 2026, the Strait of Hormuz remains closed, the latest in a series of global supply chain shocks over the past decade, including the US-China trade war, the invasion of Ukraine, global pandemic, and a single ship blocking the Suez Canal. What this means for countries seeking AI sovereignty is that they need to make strategic choices on where to invest, and where to partner, within a resource-constrained and unstable geopolitical environment.

The Emerging Contours of Sovereign AI

Sovereign AI, as most governments define it in practice, is a model of AI development and deployment where "inputs adhere to a state or political union's laws and institutional frameworks, and outputs are contextually relevant, secure, and create value for the economy."

The majority of sovereign AI projects follow a pattern I call "hybrid sovereignty", where governments co-develop AI infrastructure and use cases with private sector partners.

Some larger economies are able to leverage their market size and attractive incentives to draw in international partners to support a full stack of AI capabilities, from data centers to sovereign models. Stanford HAI's 2026 review of commercial AI sovereignty offerings finds that Nvidia, Microsoft, Google, AWS, and OpenAI (as well as Chinese companies like Alibaba) have all rebranded their infrastructure and cloud products around sovereignty, but that most of these offerings reconfigure rather than eliminate country dependence on the underlying provider.

DOI: 10.5281/zenodo.23666309

MAIEI

UAE's G42, which is the country's flagship AI initiative, is building out a Stargate UAE campus that runs on Nvidia's GB300 chips, as part of a partnership with OpenAI, Oracle, Nvidia, Cisco, and SoftBank. India runs its national AI program on NVIDIA's chips, and has secured at least USD 50 billion in commitments already made to Amazon, Microsoft, and Google.

But a country's market and incentives are not the only factors in determining its ability to attract sovereign AI investments. As of 2026, infrastructure projects account for 80% of sovereign AI initiatives. In Latin America, Brazil has emerged as the region's dominant infrastructure hub, controlling nearly 73% of installed data center capacity in South America as of 2025. The interesting contrast in the region is Chile, which has positioned itself to investors as a regional alternative, supported by a renewables-heavy grid. Chile's advantage also comes from how competitive its renewable energy pricing is, with 15- year purchase agreements that make renewables cheaper than fossil fuels. The sovereign asset being offered in these cases is cheap, reliable energy and the ability to become, in effect, compute *exporters* for the region rather than compute *renters*. This energy advantage is also a constraint. Santiago's data center hub faces chronic water shortages and grid limitations. This does raise a question: are the citizens of a country willing to bear the cost for sovereignty?

The reality of this hybrid approach to sovereign AI as it is currently playing out will likely result in one of two pathways. Capital-rich states, such as the UAE, will pursue deep integration or interdependence, acquiring equity stakes and co-locating physical infrastructure inside supplier networks so their interests become structurally entangled with those of the dominant technology firms. Resource-rich but capital-constrained states, lacking this option, such as Chile, are more likely to pursue asset substitution, converting their energy endowments into a form of structural leverage, building a bargaining position on the power requirements required to support compute.

Sovereign AI strategy is therefore increasingly shaped by material constraints, including energy costs, grid capacity, and exposure to supply chain disruption. What looks like policy divergence is therefore better understood as convergence around supply chain and energy constraints, shaped partly by a country's access to capital and natural resources.

DOI: 10.5281/zenodo.23666309

3.2 The Empty Sovereignty Promises of Canada's 2026 National AI Strategy

By Véronique Sioufi , BC Policy Solutions

In 2025, in what would become a flashpoint on the question of digital sovereignty, a US executive order imposed sanctions on ICC Chief Prosecutor Karim Khan and UN Special Rapporteur Francesca Albanese for investigating Israel's war crimes; they cut off their visas, and denied them access to their banks and emails (Marshall et al., 2026; Quell, 2025). In light of examples like these, and building on existing EU digital regulations, European nations accelerated efforts to curb reliance on US tech, with French, German, Austrian and Dutch governments, and The Hague, dropping companies like Microsoft and Zoom in favour of European platforms and infrastructure (van Es, 2026).

Canada, however, is even more entangled within the American digital sphere (and lagging far behind the EU on data protections, digital market controls, and AI regulations). US tech giants own and operate most of Canada's digital communications infrastructure, including cloud services, app stores, operating systems, search engines, and data centres (Tusikov, 2025). AI Minister Evan Solomon called digital sovereignty "the most pressing policy and democratic issue of our time," with the government committing upwards of $2 billion CAD toward a conflated AI and digital sovereignty agenda.

Meaningful digital sovereignty spans at least six dimensions: territorial (where are data and compute located); jurisdictional (whose laws govern the service/provider); operational (who can administer, suspend, or alter the service); economic (who owns assets, sets prices, receives subsidies, captures rents); individual (can users migrate, audit, substitute or correct data); and collective/relational (which communities have authority over data about themselves). Yet researchers and civil liberties advocates argue that while the federal government champions digital sovereignty *rhetorically*, it is failing to build it *materially.*

The US CLOUD Act and Bill C-22

The government's push for digital sovereignty centres on building domestic infrastructure and subsidizing Canadian tech, yet even on these terms the strategy has some glaring holes.

Cohere, for example, championed as Canada's flagship AI company capable of securing national tech sovereignty, relies heavily on American hardware, cloud infrastructure, and partnerships with warfare technology companies like Palantir profiteering from the genocide of Palestinians (Sobowale, 2025; see United Nations, 2025). The federal "Buy Canadian" procurement policy also ignores parent-company corporate structures, allowing US software vendors to qualify as Canadian simply by opening a domestic branch (van Es, 2026).

DOI: 10.5281/zenodo.23666309

Domestic data centres are purportedly necessary to curb reliance on US infrastructure and protect against the US Clarifying Lawful Overseas Use of Data (CLOUD) Act, which compels US service providers to hand over data held abroad. But by ignoring the software-as-a-service (SaaS) layer, federal rhetoric obscures the fact that storing data on Canadian soil does not protect it from US surveillance (van Es, 2026; Zhang, 2025). Worse, since 2022, Canada has been negotiating a CLOUD Act agreement that would allow US agencies like the FBI and Department of Homeland Security access to real-time wiretaps, subscriber data, and sweeping keyword or geofence "reverse warrants" directly from Canadian tech and telecom providers without Canadian court oversight (Khoo & Robertson, 2025; Robertson, 2026). This would subordinate Canadian constitutional safeguards to weaker American standards, leaving Canadian residents without domestic judicial review nor US legal remedies, which exclude non-citizens (Khoo & Robertson, 2025).

So, while on June 8, 2026, the National AI strategy was released promising AI infrastructure "operated under Canadian control and Canadian law," ten days later, on June 18, Bill C-22 passed, laying the legal groundwork for the CLOUD Act deal (Robertson, 2026). If a deal follows, there would be "little meaningful shield" even for data "stored by a Canadian company in Canada" (Khoo, 2026).

US Lobbying

Digital sovereignty cannot be achieved simply by pouring billions of dollars into domestic server farms or attempting to build a Canadian replica of global hyperscalers; sovereignty is also a matter of governance. It requires the regulatory autonomy to govern foreign platforms operating within Canada, enforcing Charter privacy and human rights standards, mandating algorithmic transparency, curbing monopolies, and restricting cross-border data flows (Zhang, 2025; Tusikov, 2025; Khoo & Robertson, 2025).

Because real sovereignty hinges on regulation rather than physical hardware, domestic governance is precisely what Washington and Silicon Valley target most aggressively. The US government and American tech companies routinely pressure Ottawa to weaken or altogether dismantle regulation they see as "trade barriers," as we've seen with the Digital Services Tax Act, Online News Act, and Online Streaming Act (Tusikov, 2025; McNeil & Beny, 2025). While US tech giants report "international trade" as a primary lobbying topic in Canada, they have been bypassing Canada's Registry of Lobbyists via intermediaries like the US Chamber of Commerce and bilateral state negotiations (McNeil & Beny, 2025; Robertson, 2026).

Since 2022, Washington has leveraged bilateral trade talks to push Ottawa into aligning with the US surveillance "toolkit," making Bill C-22's lawful-access reforms a top-three US trade demand while congressional leaders recently pressed Canada's Public Safety Minister for "prompt collaboration" on a CLOUD Act deal (Robertson, 2026). This dynamic leaves the future of our already wanting privacy, data, and AI governance laws vulnerable (McNeil & Beny, 2025).

Indigenous Sovereignty

As Ottawa cedes regulatory ground to Washington, its “sovereign” tech agenda also ignores the foundational sovereignty question within this settler-colonial state: the self-determination of Indigenous Peoples (Sioufi, 2026).

Indigenous Peoples, long subjected to state data collection as an instrument of control, have developed rights-based principles that federal policy must respect and reflect. OCAP® principles (ownership, control, access, and possession) affirm that First Nation communities hold collective jurisdiction over how their information is gathered, stewarded, and shared. The National Inuit Strategy on Research establishes Inuit ownership and gatekeeping over data concerning Inuit peoples, wildlife, and lands across Inuit Nunangat. Globally, the CARE principles (collective benefit, authority to control, responsibility, and ethics) ground data ecosystems in Indigenous self-determination, capacity-building, and harm reduction.

The federal government should treat these frameworks as binding standards and apply the thrust of these principles to regulating the Canadian digital ecosystem—ensuring that digital governance is democratic, transparent, rooted in collective rights and accountable to the communities it affects.

DOI: 10.5281/zenodo.23666309

MAIEI

3.3 No Seat, No Data: The Pacific and the Limits of AI Sovereignty Talk

By Aydon Moala-Edwards, Moala Global & Monash Global Peace and Security Centre

In July, at the first Global Dialogue on AI Governance in Geneva, the Secretary-General gave the meeting its sentence: machines "can inform, but humans must decide, and answer." More than 170 countries were represented. Pacific participation rested on a small-island-states sponsorship scheme in which, as the organisers put it, "the number of candidates selected depended on funds available," with no Pacific voice featured in the high-level programme. A co-chair closed by quoting a small-state delegate: "We are here not to be a footnote." A seat you attend by sponsorship is not yet your seat. That is half the problem.

No Data

The other half is that the Pacific is missing from the evidence that would make its case. Anthropic's Economic Index published a usage index for 121 countries in its May 2026 release. Australia scores 6.4, Singapore 5.81, New Zealand 4.84 and Canada 4.13, each several times what population alone would predict. Not one Pacific Island state appears. By the Index's own methodology, countries with no observed usage are assigned to a "Minimal" tier because "we do not know if they have exactly zero usage or little usage that our random sample did not capture." That is a sampling artefact, not a fact about Fiji or Tonga. When a Pacific minister argues for a seat, the metric the room trusts shows nothing. And what the metric cannot see, the models invent. Kevyn Wahuzue, a Kanak developer whose Pasifika AI refuses to answer when its sources run out, said it in September: under-representation in training data "does not show up as silence. It shows up as confident invention." A fluent wrong answer about a Pacific word "is not a neutral error. It circulates."

Why "Small States" is the Wrong Category

Canada and Singapore are the usual small-states case, and the comparison is useful because it fails. For Ottawa or Singapore, AI sovereignty is a question of choosing: which compute, which supplier, which standard. Singapore co-founded DEPA. Canada convened a G7 process. Both hold a seat and a data footprint. For a Pacific state, the question comes earlier: whether you are in the room, and in the data, at all. One word, two problems; policy written for the first will not solve the second. The region has now put the word on paper. At the 55th Forum Leaders Meeting in Koror, leaders endorsed the Pacific Digital and Artificial Intelligence Technology Principles, naming digital sovereignty, language protection and human oversight, with a framework due in 2027. Tess Newton Cain and Stefan Armbruster counted the communiqué: reaffirmed 38 times, endorsed 20, tasked once. Sovereignty talk is now abundant. Sovereignty terms are not.

DOI: 10.5281/zenodo.23666309

What Indigenous Data Movements Already Worked Out

The people who worked this out first were not states. In Aotearoa, Te Mana Raraunga, the Māori Data Sovereignty Network, set out principles. In 2019 the Data Iwi Leaders Group signed the Mana Ōrite Relationship Agreement with Stats NZ, the first of its kind between iwi and the Crown, giving both parties' views equal explanatory weight. It also created a standing body, Te Kāhui Raraunga, which now runs Te Whata, an iwi-designed platform that holds census data about iwi on iwi terms. In Australia, Maiam nayri Wingara set out principles, and in 2024, the Commonwealth adopted a Framework for Governance of Indigenous Data co-designed with First Nations partners. The collective is still calling for an Indigenous data authority to exercise it. Principles, an agreement with the state, a body that holds the authority: Aotearoa has reached the third step, Australia the second, and Koror the first. That is sovereignty as design authority: authority over how a system is decided, not consultation about it. The CARE Principles give it a form that travels.

Pacific nations can borrow that pathway with one asset those movements lack. They are states, with treaties, exclusive economic zones and a vote, and they have run a collective play before. The eight Parties to the Nauru Agreement stopped selling access to their tuna one fleet at a time. Under the Vessel Day Scheme they sell the right to fish by the day, at a shared minimum price no member undercuts. Revenue rose from about US$60 million in 2010 to close to US$500 million by 2018. The equivalent move for data is the same one: agree together which data classes are offered, on what terms and at what floor, and refuse to sell below it one state at a time.

What Independent Governance Looks Like

Two things must be tested. The first is a regional conversation on AI convened by Pacific creatives, community organisations and researchers rather than government representatives alone, hosted by an institution the region owns, such as the University of the South Pacific, rather than arranged for it by donors. Its task would be to produce the Pacific's own account of how AI is already used here. That account is precisely what the index cannot see, and it is the evidence the second test needs. The second is a Pacific AI Readiness Assessment, that asks the questions the international frameworks skip. Such a process could anchor to ISO/IEC 42001, the NIST AI Risk Management Framework, Australia’s AI Ethics Principles, and utilise CARE-based governance gates: which data classes are never ceded (fisheries, marine genetics, seabed, geography, community-held knowledge); where a system gets its information and what it does when uncertain; and whether the community can say no after go-live, not only during procurement. The payoff is evidence the Pacific generates about itself, so that the next index has something to count. Before asking whether a small state is sovereign over AI, ask whether it is in the data at all. If it is not, no seat will fix it. The Pacific must set the terms.

DOI: 10.5281/zenodo.23666309

MAIEI

Chapter 4: Governance Frameworks & the Practitioners Building Them

4.1 Ethical AI Needs a New Opening and the Sovereignty Push Might Just Be It

By Claudia Del Pozo, Eon Institute

4.2 Practitioner-Led AI Governance Across the Global Majority

By Chinasa T. Okolo, Technecultura

4.3 The Mirror We Hold

By Audrey Tang, Cyber Ambassador of the ROC (Taiwan) government

4.4 A Framework With No Scars Has Never Been Run

By Aida Gregorian, LawVo Inc.

DOI: 10.5281/zenodo.23666309

4.1 Ethical AI Needs a New Opening and the Sovereignty Push Might Just Be It

By Claudia Del Pozo, Eon Institute

The year 2018 marked the acceleration of AI policy conversations, with ethics becoming a key pillar around which an entire policy ecosystem was built, from national strategies, to international recommendations, corporate commitments, and governance frameworks. Eight years later, however, the translation of those principles into tangible results remains limited.

UNESCO's 2026 implementation data found that, while 75% of surveyed Member States with an AI strategy reference ethics and human rights in their strategies, only 24% require people to be notified when interacting with AI, 16% report dedicated redress or liability frameworks for AI-related harm, and fewer than 11% have liability regimes with clear responsibilities around AI-induced damage. Part of the problem is that these principles were mainly translated into voluntary commitments, which are easier to publish than enforce. Without clear legal duties, independent monitoring, remedies, and resources to audit systems, ethical guidance cannot become operational control. This results in many of the AI systems we rely on today falling short of the things the sector itself says ethical means: that people know when they are dealing with a machine, that someone is liable when it harms them, and that an independent party can check.

This implementation gap persists, in part because it is also a power gap. In 2025, over 90% of notable models were produced by industry, according to Stanford's 2026 AI Index. This means they were developed by companies operating under intense competitive pressure, creating incentives that place speed and market advantage ahead of slower-moving public-interest requirements. Moreover, it is worth noting that the US is particularly dominant in this landscape, with US institutions producing more notable AI models than any other country last year. As a country that has recently moved to deregulate AI and roll back federal AI safeguards, this reality creates even scarcer conditions for public-interest requirements that often have to be imposed via regulation and other mechanisms. The result is a limited ability to audit, modify, and adjust systems locally to reflect the different realities of the institutions and organizations adopting them, and the more these systems spread without meaningful local adaptations and control, the more distant ethical AI becomes.

That reality, however, might be about to change. While model choice may still mainly be dictated by capacity, a new dimension of value came to emerge in the past months. On June 12, 2026, following a US government export-control directive, Anthropic temporarily suspended access to two of its leading models, Mythos 5 and Fable 5. The directive required access to be withheld from foreign nationals inside and outside the United States, including the company's own foreign-national employees. Rather than gate access by nationality, Anthropic switched both models off for every customer, for nineteen days. In light of this episode, institutions and

DOI: 10.5281/zenodo.23666309

MAIEI

organizations across the world realized their systems could be easily disconnected, creating potential serious disruptions to services that had come to depend on them. Suddenly, reliability of accessibility emerged as a new priority for choosing AI models and systems.

This new consideration tied in perfectly with the increasingly popular topic of AI sovereignty (understood as agency not domestic ownership). It led many actors to reconsider the value of local models, the development of which had previously seemed futile in the face of dominant global alternatives. This is where I see an opening for ethical AI. As governments and organizations reconsider the models and systems they rely on, whether for business as usual or critical services, they also have an opportunity to revisit not only where those models come from, but what they expect from them. In this sense, sovereignty concerns have done something important. They have disrupted the assumption that existing model choices are inevitable, bringing back choice, understood as the ability to evaluate, adapt, switch, or reject systems. If institutions are already reassessing those choices, ethics can (and should) become part of the criteria shaping what comes next.

This creates an opportunity to work with existing initiatives that have already developed local, collaboratively built models. Latam-GPT, for instance, is a regional, open-source initiative led by Chile's National AI Center together with over 30 institutions across eight Latin American countries. Built around Latin American data and cultural contexts, it is intended for governments, universities, and developers, with an open-source infrastructure that further facilitates adaptation to specific national contexts. Similar culturally-aligned efforts are emerging elsewhere, including the UAE's Jais Arabic LLMs, Singapore's SEA-LION family for 11 Southeast Asian languages, and Africa's InkubaLM for multilingual African contexts, to mention a few. Initiatives like these offer not only reduced dependence on foreign model providers, but also a better representation of affected populations and greater relevance to local contexts, advancing key principles of ethical AI such as fairness, inclusion, and relevance.

Of course, local does not automatically mean fully ethical or complete technological independence. Many local initiatives rely on Chinese open-weight models, which offer greater flexibility but can also mean replacing one foreign dependency with another, without necessarily gaining greater control over the underlying model. These projects can also depend on external infrastructure, compute, or technical ecosystems. Separately, locally developed or hosted systems can introduce their own ethical risks, from surveillance to the concentration of power close to home. Even a locally hosted model can lack basic safeguards such as transparency, independent audits, and mechanisms for redress. What sovereign AI has brought back to the table is choice, but choice alone does not guarantee ethical outcomes. The opportunity now is to ensure that governments and companies use this moment to ask not only who they want to depend on, but also what kind of AI systems they want to depend on in the first place.

DOI: 10.5281/zenodo.23666309

MAIEI

4.2 Practitioner-Led AI Governance Across the Global Majority

By Chinasa T. Okolo , Technecultura

While advanced economies grapple with political backlash and conflicting discourse on existential risks and loss of control, practitioner communities across the Global Majority, including Africa, Latin America, the Caribbean, Oceania, and South and Southeast Asia, are building durable governance frameworks on their own terms. Faced with the systemic pressures of digital colonialism and infrastructural dependence, these communities are asserting digital sovereignty through localized research networks, agile regional governance mechanisms, and innovative data licensing models. When more than 170 countries met in Geneva in July 2026 for the first UN Global Dialogue on AI Governance, the disparities in Global Majority-led AI policy frameworks became even more apparent.

Distributed Innovation and the Open-Source Model Ecosystem

In regions historically marginalized by global technology markets and facing the global AI divide, the foundational layer of AI capacity is being built through distributed research networks. The compute gap renders domestic AI development infeasible at a cost for many Global Majority nations, increasing reliance on foreign digital infrastructure and open-source/weight models.

The African natural language processing (NLP) community exemplifies this shift through efforts such as Masakhane, a pan-African grassroots organization that builds datasets and machine translation tools for African languages. Other groups, such as Ghana NLP, DataScience Nigeria, EthioNLP, HausaNLP, and KmerAI, also work to expand access to low-resource African languages. Similar grassroots and institutional initiatives are driving progress in Southeast Asia and Latin America. The SEACrowd initiative operates as a comprehensive multilingual and multimodal data hub designed specifically to evaluate and improve language models for Southeast Asian linguistic contexts. In Latin America, the National Center for Artificial Intelligence in Chile (CENIA) coordinated the launch of Latam-GPT to reduce dependency on models optimized primarily for English-speaking populations.

Balancing the AI Policy Landscape

While practitioner-led innovation accelerates, the formal policy landscape across the Global Majority remains highly uneven. National and regional governance frameworks are expanding rapidly, yet they often outpace the institutional capacity required for robust enforcement, leaving distinct governance gaps in regulating advanced AI capabilities.

DOI: 10.5281/zenodo.23666309

While at least 22 African countries have standalone national AI strategies, the African Union Continental Strategy for Artificial Intelligence serves as the primary regional anchor. Despite these ambitious legislative designs, implementation is frequently hindered by siloed ministries, limited workforce depth, and weak enforcement mechanisms. The governance landscape in the Caribbean and Pacific Islands presents a stark contrast to other regions across the Global Majority, with all of the Small Island Developing States in this region currently lacking formal national AI strategies. In September 2026, Pacific Islands Forum leaders endorsed regional Digital and Artificial Intelligence Technology Principles at their meeting in Koror, with the framework itself due in 2027. Until it arrives, the regulatory vacuum stemming from profound regional structural vulnerabilities remains, despite a growing focus on broader digital transformation initiatives.

Conversely, Southeast Asia and Latin America demonstrate more advanced regulatory environments. Across Southeast Asia, regional efforts led by ASEAN focus on structured governance through frameworks, like the 2024 Guide on AI Governance and Ethics, the 2025 Expanded Guide on Generative AI, and the 2025–2030 Responsible AI Roadmap to ensure transparency, manage frontier risks, and coordinate trusted digital ecosystems. Meanwhile, Latin American countries are implementing diverse national strategies and legislation with Peru enacting Law No. 31814 to classify high-risk systems and restrict autonomous weapons; El Salvador establishing a national AI agency and framework to limit developer liability and attract investment; Brazil advancing a comprehensive AI bill covering civil liability, risk tiers, and copyright rules; and Mexico proposing a risk-based compliance law to be overseen by a national commission.

Practitioner-Led Digital Governance

As national policies struggle to catch up with advanced technological capabilities, the most significant form of governance emerging across the Global Majority operates through localized, practitioner-led efforts focused on preventing data extraction.

The Nwulite Obodo Open Data License was developed specifically to manage African-language datasets, directly addressing the equity gap and differentiating access for users by geographic location. For researchers and entities in economically developing nations, the license provides permissive, royalty-free open terms to foster local innovation. For users in high-income nations, the license imposes strict benefit-sharing obligations, ensuring that commercial extraction enables reciprocity for the originating community. The Esethu Framework permits broad research access for African entities while mandating that licensing fees collected from non-African commercial entities be systematically reinvested in expanding local datasets.

In Oceania, efforts to secure data sovereignty are heavily influenced by Indigenous knowledge systems that prioritize collective stewardship over individual property rights. The Maori community in New Zealand, led by organizations such as Te Hiku Media, has pioneered the

Kaitiakitanga License to govern linguistic data used in NLP tools like the Papa Reo API and the Kaituhi transcription service. These localized efforts are supported globally by the CARE Principles for Indigenous Data Governance, which comprise the principles of Collective Benefit, Authority to Control, Responsibility, and Ethics, and promote the integration of indigenous protocols and jurisdictional respect into algorithmic design.

Conclusion

Recognizing that state-level policy frameworks often lack the enforcement mechanisms necessary to prevent digital extraction and that foreign AI policy frameworks conflict with local realities, communities across the Global Majority have localized the practice of AI governance. Through distributed research networks, practitioners are bypassing hardware constraints to build sovereign open-weight models that respect linguistic diversity and cultural nuance, and, by deploying conditional licensing structures, communities are effectively rewriting data access protocols. These practitioner-led frameworks demonstrate that true capacity building can be defined by possessing the structural leverage to demand reciprocity, protect cultural heritage, and dictate the precise terms under which local data fuels global innovation.

DOI: 10.5281/zenodo.23666309

4.3 The Mirror We Hold

By Audrey Tang, Cyber Ambassador of the ROC (Taiwan) Government

The question I hear most often is who gets to decide how AI changes our societies. A handful of companies now concentrate the compute, capital and models mediating ever more of our working and civic lives. At the same time, the frameworks meant to hold them answerable grow thin: safety bodies hollowed out, voluntary codes shelved. In June, a U.S. export-control directive pushed Anthropic to suspend its new Fable 5 and Mythos 5 models worldwide for weeks. Power consolidates, institutions retreat and the people are told the terms have been set on their behalf.

I do not accept this. A decade of practice has shown me what the people decide when trusted with the terms.

AI is a mirror. It reflects our collective soul and our humanism back at us. Hold it up to an extractive logic and it extracts. Hold it up to a community that answers to itself and it widens the circle of who gets to think together. In Mandarin, the romanization for love, 愛, is spelled "AI." The technology is not destined to dehumanise. Our care decides what it becomes.

Taiwan learned this the practical way. In 2015, when Uber's collision with taxi drivers seemed irreversible, our vTaiwan process invited everyone impacted into one shared space. The system we used, Polis, has no reply button. There is no thread to dominate, no troll to feed; the only way to be heard is to draft a statement the other side can also endorse. Among some 925 voters, four opinion clusters emerged; on the questions that mattered, insurance, licensing and fair fares, consensus surfaced across all four. That uncommon ground set the agenda for a face-to-face forum with the company, taxi associations and relevant agencies — the government pledged to ratify those items into regulation.

This lesson travelled. People can disagree and still decide together when technology rewards bridging over outrage. That is the Taiwan Model: a demo, not a template.

Our record is not perfect, though: A later vTaiwan case on online alcohol sales produced a draft withdrawn before it reached statute. Deliberation cannot reach inside a firm's own governance unless the firm opens the door; the most consequential decisions still happen in rooms the public never enters. That is why the rooms we can open matter.

The same instinct shapes the 6-Pack of Care, a framework my colleague Caroline Emmer De Albuquerque Green at Oxford's Institute for Ethics in AI and I published. It draws on U.S. political scientist and academic Joan Tronto's care ethics, and comprises, **Attentiveness:** notice who is affected, the least powerful first; **Responsibility:** someone answers when the system acts; **Competence:** the care actually works, as a security failure is a moral failure; **Responsiveness:** listen back and measure whether trust survives loss; **Solidarity:** make

positive-sum games easy to play; and **Symbiosis:** build for enough, not forever, so the system stays a steward, never a master.

A Civic AI is one a community can own, inspect, correct and switch off. The answer to consolidation is many bounded stewards accountable to the people they serve: AI on tap, not on top.

This is also how we repair weakening governance. Deliberation can move faster than the speed of legislation. In California, I advised Gov. Gavin Newsom and First Partner Jennifer Siebel Newsom on Engaged California, which pivoted in its launch week to Los Angeles wildfire recovery, drawing coherent priorities from more than 900 people affected by the fires amid extreme polarisation.

Its third iteration asked more than a hundred Californians, chosen by civic lottery, what their government should do about AI's economic impact, and what the frontier labs owe them: apprenticeship, belonging, care and dignity. In late August, California's State Legislature passed AB 2103, writing the programme into law and signed by the governor in September. In Warren County, Kentucky, officials acted as a "listening partner," asking an open question: What could Bowling Green be in 2050? Nearly 8,000 residents answered, about one in 10 in the city, and in April their shared vision became a strategic plan before the Fiscal Court.

Notice what collective intelligence offers the people this volume names: pushing back. Workers can get a say over the systems that govern their work before deployment, instead of a petition afterwards. Artists can shape the terms of co-creation, with consent, credit and payment built in from the start. A bounded co-pilot can help a class make things together, turning students from mere readers into co-creators. Communities gain stewards they can inspect and retire. And public institutions gain something rarer than compliance: a way to earn trust by giving it first.

To give no trust is to get no trust. When a government pre-commits to adopt whatever the people agree on, consensus carries consequence and participation becomes power. A race has a finish line nobody reaches; a mission has milestones everyone can contribute to. Global AI governance must be a mission, not a race: a mission for shared prosperity, never a race to leave humanity behind. Bring only fear and the mirror hands fear back at scale; bring care and a willingness to take all sides, and it reflects that too.

After all, we the people are truly the superintelligence. Let us govern to our full potential and free the future — together.

DOI: 10.5281/zenodo.23666309

4.4 A Framework With No Scars Has Never Been Run

By Aida Gregorian , LawVo Inc.

Every AI governance framework we consulted in 2026 told us to establish meaningful human oversight. None of them told us what to do when the correct answer in Ontario is the wrong answer in Quebec.

Through 2026, our team built the governance layer for an AI legal platform across every Canadian province and California: two legal traditions across three legal orders, none of which share a conceptual vocabulary in full. Quebec operates under the Civil Code of Quebec. The common law provinces do not. California is common law too, but it is a separate sovereign whose law is no more portable to Ontario than Quebec's. These are not dialects of one tradition. A retrieval layer that treats them as variations on a theme will mix rules, authorities and terminology from one jurisdiction into another, and return an answer that is fluent, well sourced, correctly formatted, and wrong, with no signal that anything has gone astray. We call this jurisdictional contamination, and we found it because we tested for it. Nothing in the standards we were working from told us it was a failure mode. Jurisdiction is not a filter applied to retrieval. It is a constraint on what a system is permitted to conclude.

We are the people these frameworks are written for. We are the humans the loop is supposed to run through. What I can report from that seat is that the governance literature assumes a blueprint that does not exist.

Borrowing From the Nearest Discipline

What we have instead is borrowing. AI governance in practice is assembled from the nearest adjacent disciplines: privacy impact assessments, anti-money laundering compliance, and corporate governance regimes. Those instruments are serious, built over decades against harms someone had already catalogued, with enforcement behind them. They are also the wrong shape. A privacy impact assessment was designed to interrogate the processing of data, not the behaviour of a model. An audit trail built to satisfy a reviewer is not one built to enable correction. The artifact gets produced. The governance does not.

Sometimes there is nothing to borrow at all. When we went looking for a way to measure whether our agents were legally accurate, we found no accepted standard to measure against. There is published work on legal reasoning evaluation, such as LegalBench, presented at NeurIPS in 2023, and on hallucination in commercial legal research tools, such as Magesh and Surani et al. in the Journal of Empirical Legal Studies in 2025. Both are built on American common law. Fawzy's 2026 empirical study at the Egyptian Council of State has since documented the same failure patterns in a different legal tradition, including citation, contextual

DOI: 10.5281/zenodo.23666309

MAIEI

and doctrinal hallucination, and hierarchical reasoning failure, and proposes a diagnostic and mitigation framework rather than a benchmark.

None of it resolves into a benchmark that is jurisdiction-aware, and none of it is the kind of standard a profession would recognize. In other regulated fields, that work has historically been done by professional bodies rather than by vendors. For now it falls to the builders. So we built one in house, written by lawyers, because the alternative was to deploy without knowing. It is a real asset and I would build it again. It is also a private yardstick, and a market in which every vendor grades itself against its own private yardstick is one in which no buyer can compare anything.

That is the part I think the field has backwards. The argument is usually framed as markets against regulation. It is not. Competition is the most reliable engine of quality we have, and it is the reason this technology has improved as fast as it has. But competition rewards quality only where quality is visible. Where buyers cannot verify what they are purchasing, the rigorous vendor and the careless one present identically, and price decides. The firm that invests in evaluation carries the cost and captures none of the advantage.

I came to AI from regulated practice, and that is the lesson those sectors learned expensively. Finance, insurance, securities, and the legal profession all passed through a voluntary-code era and all abandoned it, not because markets failed but because published principles produced no signal a buyer could act on. What came next was not the suppression of competition. It was registration, examination, and a defined floor, which is what made competition above that floor mean something. Responsibility in those fields is not an aspiration. It is a floor with a regulator standing on it, and the firms operating above it can prove they do.

Somewhere to Put the Failures

None of this is an argument for waiting. The absence of a blueprint is a reason to build with more care, not less. Our organization made that choice at the outset rather than as a compliance exercise later. We adopted the TRAIL framework we trained under at Mila as our operating lens from the first architecture decisions, not as a document produced for a client questionnaire, and aligned our management system to ISO/IEC 42001. I want to be precise about what that bought us. It did not give us answers. It gave us somewhere to put the failures.

We built the integrity of that system bit by bit, through failed feedback. Jurisdictional contamination is a failure mode we test for because testing found it. Our retrieval boundaries got tighter because loose ones produced confident errors. Our escalation rules exist in their current form because earlier versions did not trigger when they should have. Every rule in that layer is a scar, found by lawyers, researchers, and engineers reading output that looked correct until someone with the right jurisdictional training read it closely.

An AI governance framework that has never caught anything has never been run. A system that has produced no corrections, no revisions, no documented instance of something being stopped, is not working quietly. It has not been switched on.

So three things to carry forward. Ask any governance framework to produce its failure record, your own and the ones your vendors put in front of you, and treat an empty record as a finding rather than a reassurance. Require every principle to name a rule, an owner, and a trigger, because a principle without those three is a sentiment. And treat jurisdictional incompatibility as a design constraint from the first architecture decision, not a compliance problem found after deployment. The question to ask of any system operating across legal traditions is not which jurisdictions it covers. It is what it does when they conflict.

If the responsible AI field cannot show what its governance has caught, it is not a governance field. It is a communications function.

DOI: 10.5281/zenodo.23666309

MAIEI

PART II: SURVEILLANCE, WARFARE, & OVERSIGHT

Chapter 5: The Expanding Surveillance State

Chapter 6: Military AI Takes Centre Stage

Chapter 7: Oversight Built for Another Era

DOI: 10.5281/zenodo.23666309

Chapter 5: The Expanding Surveillance State

5.1 The Warrant Nobody Could Explain

By Maria Lungu, Kennesaw State University & Digital Technology for Democracy Lab

5.2 Guilty by Default: How Rushed Surveillance Law is Reversing the Burden of Proof

By Isabelle Castro, Independent Researcher & Utopia in Beta

5.3 Disability Surveillance in the UK: AI, Care, and Power

By Amy Gaeta, Leverhulme Centre for the Future of Intelligence at the University of Cambridge

5.4 Strategic Litigation as a Transformative Tool for Advancing AI Activism in the Global South: Lessons from the Commonwealth Caribbean

By Amanda Janell DeAmor Quest, Commonwealth Caribbean Lawyer & Independent AI Ethics and AI Governance Researcher

DOI: 10.5281/zenodo.23666309

5.1 The Warrant Nobody Could Explain

By Maria Lungu, Kennesaw State University & Digital Technology for Democracy Lab

The day after Thanksgiving in 2022, DeKalb County police pulled over Randal Quran Reid on Interstate 20 and told him he was wanted in Jefferson Parish, Louisiana. Reid had never been to Louisiana. He spent six days in jail. When he asked the arresting officers what the warrant alleged, they told him it was outside their jurisdiction and they did not know. They were telling the truth. A facial recognition search run by the Jefferson Parish Sheriff's Office had returned three names, Reid's among them, and a Baton Rouge detective adopted the identification to obtain a second warrant. By the time it reached the officers on I-20, the match had passed through two agencies and become a fact. A judge signed the warrant on an affidavit that never mentioned the software. Reid was freed because a lawyer showed the sheriff's office photographs of him elsewhere, not because any court examined the match. The usual objections to AI and facial recognition systems are bias and opacity. Both are real, and both are downstream of a problem the reform debate keeps missing. Due process attaches to decisions, and algorithmic tools are built to produce inputs. The vendor sells the match as a lead, not an identification. The police then treat it as a lead a detective will confirm. So no one ever stands behind the match, and no court tests it. If disclosure is required whenever an algorithm's output feeds a decision, the fix then moves from constitutional law to the contracts agencies sign when buying these tools.

Due Process is Triggered by Decisions, and Algorithms are Careful Never to Make One

Leads have always been allowed to be unreliable. But when an informant's tip supports a warrant, the judge must assess whether the informant and their claims are credible. No comparable rule requires anyone to assess an algorithmic lead before the state acts on it. Officers build photo arrays from leads and rely on them for arrests. The constitutional protections that might check this process each assume an accountable actor. The Fourteenth Amendment bars a state from taking someone's liberty without fair legal process, which assumes the state did something identifiable to that person. The Fourth Amendment requires a warrant supported by probable cause sworn under oath. The Sixth Amendment lets defendants question the witnesses against them. Each protection needs a specific act by someone who can be held responsible. Each protection requires a specific act by someone who can be held responsible. A candidate list, however, provides none. No officer vouches for it and no witness makes it.

The Input Travels, the Obligation Does Not

To extend this discussion and assess practical instances of these issues, let's take a look at a number of recent court cases. In June 2023, a New Jersey appellate panel in *State v. Arteaga*

DOI: 10.5281/zenodo.23666309

MAIEI

directed the trial court to order discovery on the technological system that produced a match, since without it, no one could say whether the material was exculpatory. West New York detectives had sent a still image to the state's intelligence center, which found no match. They then sent the raw footage to the New York Police Department, which returned Arteaga's mugshot. Two witnesses subsequently identified him. On remand, prosecutors said they could not produce the material because the search had happened in another state. After nearly four years in pretrial detention, Arteaga pleaded guilty and went home.

In *State v. Miles*, New Jersey's Supreme Court in June 2026 unanimously decided what defendants can learn about a facial recognition match. Prosecutors must disclose the software's name and maker, its accuracy record, the photo searched, and the candidates returned. The Court held off on the source code, since the defense admitted it might not need it, and chose case-by-case judgment over a fixed checklist. That is a sensible line, but disclosure only works if someone keeps a record. The Miles search ran on a system shared by two states. The State Police had no record of it, and the Public Defender said it was unclear where the results came from. In a subsequent case, a DC judge in July 2026 directed prosecutors to explain how police used AI to identify a shooting suspect the victim later picked from a photo lineup. The New Jersey and DC courts both reached disclosure through ordinary discovery, not new doctrine.

What Should Change

Any algorithmic output that materially contributes to a charging, detention, or sentencing decision should be disclosable with its provenance. This includes the vendor, the system version, the database searched, the confidence value, and the human who turned it into an action. Two duties and a remedy would give that principle teeth. The state should have to reproduce its own output on request, which no agency relying on an out-of-state search can now do. Trade secrecy should not bar disclosure in a criminal proceeding. Ultimately, an agency that cannot satisfy either should be barred from using the output as the basis for a warrant, which would have kept Reid out of jail without settling any hard question about facial recognition in general. None of this asks courts to understand machine learning, but it asks the state to remain capable of accounting for its own conclusions.

The Right to Ask Why

Reid's arresting officers were describing a system in which a judgment about a person's liberty had been made, and no one present could defend it. He later settled for $200,000, which paid for the harm but changed nothing about how it happened. Due process guarantees that when the government takes your liberty, someone accountable can explain why, and you can challenge it. A criminal justice system can adopt every tool on the market and keep that guarantee. It cannot keep that guarantee while treating the matches that put people in jail as though they were nobody's decision.

DOI: 10.5281/zenodo.23666309

5.2 Guilty by Default: How Rushed Surveillance Law is Reversing the Burden of Proof

By Isabelle Castro, Independent Researcher & Utopia in Beta

The deployment of 120,000 Flock Safety cameras across the US has brought a wave of concern regarding individual privacy in the face of AI surveillance. Despite claims of 93% accuracy, the system produces over a billion misread plates monthly (IJ, 2026), causing innocent drivers to be mistakenly detained and held at gunpoint. According to the ACLU, under 1% of scanned vehicles are tied to wrongdoing.

“Every one of those stops is a high-risk encounter [...] which could turn deadly,” said Michael Soyfer, an IJ attorney who is representing residents of San Jose and Norfolk in lawsuits challenging their cities’ use of ALPR cameras. “No one should have to prove their innocence on the side of the road because a camera couldn’t tell a zero from an O.” In October 2026 a judge in San Jose dismissed the case against the city, while appeals in the Norfolk case are ongoing.

The impact of Automatic License Plate Recognition (ALPR) surveillance has been a rude awakening into a reality where AI-assisted technology can switch the presumption of innocence to guilt. Meanwhile, online, thousands of data points are collected, tracked and stored everyday without users’ control or awareness. The technological solutions to counteract the practice, although developing, are moving slower than governments moving to enshrine increased data collection into law.

Expanding Surveillance

Citing online safety, a number of global governments have moved to pass laws that allow for heightened surveillance. In the UK, legal requirements for age verification to access parts of the web have expanded in scope between 2025-2026, while in the EU, the proposed Regulation to Prevent and Combat Child Sexual Abuse—known as “Chat Control”—which could allow legal access to certain private conversations, is making its way through the Commission’s votes. In Canada, the C-22 lawful access bill, set to expand law enforcement’s access to certain online data, passed the House of Commons in June 2026 and is now before the Senate. While in the US, the KIDS Act gained bipartisan support as a bill that would require age verification on certain websites, passing the House of Representatives in the same month.

While presented as child protection measures, age verification laws compel internet users to prove their age by default, taking away their choice to share personal information. The increased data collection creates heightened exposure to security risks by creating honeypots for malicious actors. In the first half of 2026 alone, the ITRC recorded 471.2 million breach notices surpassing the 297.5 million notices seen in 2025 (ITRC, 2026) assisted by AI and national funding from countries such as North Korea.

DOI: 10.5281/zenodo.23666309

Using AI, years of stored data can be analyzed in days to profile users, highlighting their desires and triggers. Aside from being harnessed by malicious actors for online identity theft, this data can legally be used to inform targeted ad placement and personalised pricing. In some cases, it has been proven effective for psychological profiling in efforts to shape opinion (Appel, Matz, 2021).

It could also impact online freedom, which has declined for its 15th consecutive year according to Freedom House. Of the 5.5 billion people who have access to the internet, 81% now live in countries where people have been imprisoned due to an online post. While Western nations remain digitally "free," their online freedom scores have declined due to growing privacy restrictions. The only thing restricting governments' ability to use AI surveillance technology for repressive purposes "is the quality of their governance," stated the report.

A Technological Bid for Privacy

In the face of increased surveillance capabilities, many have turned to zero-knowledge proofs to ensure their online privacy (Podda et al. 2025). The technology, often shortened to ZK proof or ZKP, is a method of cryptography that lets one person prove something is true without revealing the information behind it.

For each verification, the ID holder can choose what they want to share, and in some designs, also determine for how long the verification is approved for. In theory, it allows for the ID holder to maintain full control over sharing their personal data online.

Although the technology has been available for years, consumer facing products that use it have only recently managed to gain consumer traction. Previously too slow and expensive to compete with less private consumer-facing products, ongoing development has improved programmes so they can now run on a standard phone. That, paired with a growing interest in privacy-preserving tools, has increased their deployment.

However, despite identity providers in the crypto space claiming the implementation of ZKPs allows them to fully preserve privacy, two challenges remain.

The first is selective disclosure, restricting data sharing to one credential. Research by Brave on the limits of zero-knowledge age verification found that proofs still leak metadata, the record of when a proof is shown and to whom, even when the contents stay hidden. Combining several proofs, such as an age check and a residency check, can build a picture of its user despite not having direct access to the underlying document. The firm warned that a poorly designed system "may provide more identifying information than the web currently exposes."

DOI: 10.5281/zenodo.23666309

The second problem is revocation, deciding who can cancel a credential and how an individual can replace it. Granting an issuer authority creates a point of centralization and without some level of decentralization the issuer could become a "honeypot" of data that may be a target for hacks. The centralized power to revoke a single credential also creates its own control point, used to pressure an ID holder to expose their full activity.

Addressing these issues is possible but takes time, which rushed legislation ignores. For instance, in the EU digital ID programme, the only government initiative explicitly adopting ZKPs, an age verification blueprint launched in July 2026 despite ZKP integration still ongoing.

As digital identity expands, failing to protect privacy will deeply impact citizens, with little chance of rolling back these systems once embedded.

DOI: 10.5281/zenodo.23666309

5.3 Disability Surveillance in the UK: AI, Care, and Power

By Amy Gaeta , Leverhulme Centre for the Future of Intelligence at the University of Cambridge

Some of the most ambitious and high-risk applications of AI concern healthcare: in hospitals and surgeries, healthcare apps, and nursing homes, among others. Proposed and emerging applications include AI-assisted clinical documentation to reduce the administrative workload of healthcare workers; machine learning and other AI tools to accelerate biomedical research in drug discovery; and prediction systems to support patient safety.

Amid all these dazzling promises to improve the efficiency and effectiveness of care systems, however, we must ask whether history will repeat itself, or whether it already has. To what degree are disabled people's autonomy and privacy meaningfully considered in AI healthcare applications?

Disabled people (e.g., people with physical and mental impairments and pathologised divergences from able-bodied norms) have too often been framed as recipients of technologies, or even worse, 'test subjects' (consensual and not) for emerging technologies, rather than as "knowers and makers" of technoscience, emphasising their role in imagining, creating, and critiquing technological systems (Hamraie and Fritsch, 2019). This matters for AI in healthcare because disabled people are some of the most likely to be exposed to institutional care, assessment, monitoring, and decisions made by technological systems, and may have little recourse to objections or alternatives since their health can depend on compliance with care systems.

Among interrelated concerns of bias, autonomy, and improper care, one significant risk of AI applications in healthcare is surveillance, referring to both the erosion or elimination of one's capacity to make decisions around information disclosure as well as surveillance as a system of data collection, sorting, and evaluating for the purposes of "social sorting" (Lyon, 2017). Disabled people have historically been at the receiving end of technologies used to govern people whose bodies and behaviours are constructed as deviances requiring (often biomedical) intervention.

In the UK, considering these histories takes on particular urgency in the present. Numerous factors make it especially critical that disabled people's voices are centred in debates about AI in public services (alongside other most impacted voices). Such factors include recent cuts to disability benefits, alongside the use of AI-assisted decision-making in welfare administration, and adoption of private technologies in the NHS (including from Palantir, with its controversial contracts with the US Immigration and Customs Enforcement and the Israeli Ministry of Defense). The same company that works with the NHS in sensitive and widescale data

DOI: 10.5281/zenodo.23666309

management to support care efficiency is also sought by governments to support violent processes that raise urgent concerns about human rights and rights to due process.

The example of Oxevision provides a particularly important example. Developed by Oxehealth, Oxevision was a complex camera-based monitoring system deployed in some NHS mental health trusts in England. Oxevision used infrared and computer vision technology to detect things such as a patient's movement, breathing, and pulse within a patient's room without staff having to enter the room. The company rebranded as LIO in 2025. LIO builds on the earlier Oxevision technology, combining AI-powered ambient monitoring with digital observations, compliance tools, and other monitoring capabilities. The company presents the system as a means of improving patient safety and reducing the administrative and observational burden on staff. The LIO website says "AI-powered reporting insights" are coming soon, among other automated features (LIO Health, 2026).

Patients and other disability rights advocates have raised fundamental questions about Oxevision/LIO. The Stop Oxevision campaign has argued that such technology can undermine privacy, autonomy, and health. Stop Oxevision champions lived experience and investigative inquiry, with testimonies evidencing that Oxevision actually worsened some patients' experiences and induced or ignited trauma. One anonymous testimony, published by Stop Oxevision, shared that the patient was not explicitly told about the cameras until an incident occurred, and that they did not even know a camera was in the room. Another account recounts how awareness of Oxevision made them feel unsafe, not knowing who or when someone might be watching them. Stop Oxevision founders and supporters have elsewhere argued for more attention to health surveillance technologies like Oxevision as "inherently entangled with the violent historical and contemporary injustices of psychiatry and its carceral tools."

Concerns about Oxevision were heard in the ongoing Lampard Inquiry, the statutory public inquiry into the deaths of over 2000 mental-health inpatients under Essex-based NHS care from 2000 to 2023. In 2026, the Inquiry devoted hearings to the use of vision based monitoring systems, including Oxevision. The Inquiry found no significant evidence that Oxevision could improve patient safety or care based on the studies used to support its NHS procurement. By early 2026, the campaign reported that nine NHS trusts stopped using Oxevision. As of October 2026, ten mental health trusts were reported to the health regulator for, according to the Inquiry via the BBC, "not operating their consent mechanisms in line with the digital principles."

In June 2026, the Information Commissioner's Office opened an investigation into Oxevision's use of cameras in patients' bedrooms, citing concerns of privacy and information handling. Stop Oxevision's invaluable work and these two investigations demonstrate that a technology introduced in the name of safety, care, and efficiency must be evaluated through the experiences of the people subjected to it.

The politics of AI in healthcare and its implications for AI Ethics are unavoidable. A system can be technically impressive, accurate, beneficial to some, and designed with safety goals, while still producing harms that are seemingly invisible or implausible from the perspective of its designers and those involved in procurement. When we hear stories about how AI can make healthcare more efficient or effective, we must ask whose conception of better care is being encoded into and enacted by these systems, whose experiences are used to evaluate them, and what ongoing consent looks like. Treating disabled people primarily as subjects of AI healthcare interventions will inevitably fail to produce technologies that can truly enhance care.

DOI: 10.5281/zenodo.23666309

5.4 Strategic Litigation as a Transformative Tool for Advancing AI Activism in the Global South: Lessons from the Commonwealth Caribbean

By Amanda Janell DeAmor Quest , Commonwealth Caribbean Lawyer & Independent AI Ethics and AI Governance Researcher

On June 10, 2026, the ACLU and ACLU of Florida filed a wrongful arrest suit on behalf of Robert Dillon, an African American man wrongfully arrested by police officers who relied indiscriminately on a misidentification, or "false positive," produced by Facial Recognition Technologies (FRTs). Nowhere has the unregulated deployment of FRTs and their facilitation of discriminatory surveillance and arbitrary racialized policing generated more activist concern than in the United States (the US), where the unregulated deployment of high-risk AI systems such as FRTs by police officers has precipitated the wrongful arrest and detention of African Americans in at least several instances. In the past few years, this unfortunate state of affairs has catalysed a wave of strategic litigation cases challenging the legitimacy of unregulated FRT deployment by law enforcement agencies in the US.

Given the stated intentions of the Commonwealth Caribbean jurisdictions of Jamaica, Guyana, Trinidad and Tobago, and the Bahamas to eventually deploy FRTs as crime-fighting tools, similar concerns about FRTs are arriving in the region. Against that backdrop, strategic litigation — capitalizing on existing constitutional guarantees — is poised to become a transformative tool for advancing AI activism in the Commonwealth Caribbean as well as in other jurisdictions of the Global South.

In contemporary times, human rights activists, lawyers, and civil society leaders across the Commonwealth Caribbean have successfully leveraged strategic litigation — i.e. "a tool used to drive social, legal or political change beyond individual cases" — to secure constitutional rights in the face of socio-political barriers. That success, in turn, offers valuable lessons for advancing AI activism in the Global South — especially in jurisdictions where dedicated AI governance regimes do not yet exist, and where tensions will likely arise between corporate and governmental interests championing AI adoption for economic or political gain and activists concerned with the protection of constitutional rights in the age of AI.

What Strategic Litigation Has Won in the Commonwealth Caribbean — Securing Constitutional Rights in the Face of Socio-Political Barriers

As I have observed, strategic litigation "has catalysed tectonic shifts within the regional human rights landscape" across the Commonwealth Caribbean. In particular, it has facilitated the judicial invalidation of colonial-era laws that violated the constitutional rights of sexual minorities

DOI: 10.5281/zenodo.23666309

MAIEI

in at least six Commonwealth Caribbean jurisdictions — including Belize in 2016, Antigua and Barbuda in 2022, Saint Kitts and Nevis in 2022, Barbados in 2023, Dominica in 2024, and Saint Lucia in 2025. The decision of the Eastern Caribbean Supreme Court (the ECSC) in *Randall Theodule & Ors v The Attorney General of Saint Lucia* [2025] ECSC J0729-1 is the most recent exemplification of the value of strategic litigation in, as I have argued elsewhere, affirming the constitutional rights and belonging of sexual minorities.

Those judicial interventions are particularly significant in circumstances where regional legislatures have been unwilling to repeal rights infringing colonial-era laws due to deeply held fears of losing the popular vote in a region polarized by interminable homophobia and religious conservatism. Notwithstanding these socio-political barriers, strategic litigation has, as I have observed elsewhere, proved valuable in propelling a "Commonwealth Caribbean renaissance concerned with dismantling colonial-era criminal laws legitimising and perpetuating discrimination against sexual minorities".

In a similar vein, strategic litigation has also advanced reproductive justice in Dominica where on July 23, 2026, the ECSC adjudged a colonial-era law criminalizing abortion to be unreasonably restrictive and declared abortion to be legal in cases of "…rape, incest, severe foetal anomaly, and to protect the health of the woman".

Lessons from the Commonwealth Caribbean for Advancing AI Activism in the Global South Through Strategic Litigation

The Commonwealth Caribbean experience with leveraging strategic litigation to advance human rights activism provides three seminal and transferable lessons for advancing AI activism in the Global South through strategic litigation. They are highlighted and discussed below:

1. Consulting widely with the affected groups and other key stakeholders, including specialist lawyers, prior to as well as in ultimately mounting a strategic litigation challenge is of paramount importance. Within the context of AI activism, consulting widely with the affected groups and key stakeholders such as lawyers with specialist knowledge of AI systems and their interaction with constitutional rights is especially important given the technical nature of litigation at that intersection. With respect to historically marginalized and vulnerable groups such as people of colour, women, and children — who are misidentified at significantly higher rates by FRTs due to algorithmic biases — adopting a consultative and collaborative approach to strategic litigation will be indispensable to ensuring that their voices are adequately represented and affirmed in the quest for systemic change.

2. Whilst holding tremendous promise as a reformative mechanism for promoting certain social justice objectives, strategic litigation is not a substitute for much needed political action. Instead, it should be understood as "…a political tool that, when used strategically, can stimulate meaningful change and complement other political efforts…" Appreciating this critical distinction is of the utmost importance given the necessity of championing, alongside any litigation process, advocacy efforts aimed at establishing more exacting standards of accountability for state and non-state actors deploying high-risk AI systems in sensitive contexts. This is especially pertinent in a context where, according to AI activists and algorithmic safety watchdogs, automation biases held by law enforcement officials remain a major driver of wrongful arrests.

3. In the aftermath of initiating strategic litigation cases aimed at advancing AI activism and protecting constitutional rights from algorithmic infringement, AI activists must continue to educate affected communities, the public and other key stakeholders about their constitutional rights, why strategic litigation wins (or losses) matter and how they can be leveraged to promote lasting societal change beyond the courtroom. In the maturing constitutional democracies of the Global South, an informed, engaged, and vigilant citizenry will play an exceptionally important role in securing the sustained implementation of judicial decisions by state actors. Indeed, since strategic litigation is "a crucial means — but never an end — in the ongoing quest for social justice", AI watchdogs and activists must utilize education and sensitization to galvanize public action against excessive algorithmic overreach, surveillance, and arbitrary racialized policing.

DOI: 10.5281/zenodo.23666309

Chapter 6: Military AI Takes Centre Stage

6.1 The Off Switch: How National Security Took Control of a Public AI Model

By Asim Javed, Independent Researcher

6.2 The Many Hands Problem: Securing Accountability in the Age of AI Warfare

By Ayaz Ul-Huq Syed, The Dais at Toronto Metropolitan University

6.3 Hide and Seek: The Paradox of Representation in Military AI

By Shimona Mohan, United Nations Institute for Disarmament Research (UNIDIR) & Katherine Chandler, Georgetown University

6.4 Towards a "Smart Mix"?: Governing AI for International Peace and Security

By Branka Marijan, Project Ploughshares

DOI: 10.5281/zenodo.23666309

6.1 The Off Switch: How National Security Took Control of a Public AI Model

By Asim Javed , Independent Researcher

Over three months before the US government restricted worldwide access to Anthropic's newest models Fable 5 and Mythos 5, the company had drawn a line of its own. Anthropic told the Pentagon that Claude could not be used for mass surveillance of Americans or in fully autonomous weapons, even though a Pentagon spokesman said that it intended neither use. The dispute raised a prior question: once the military acquires an AI system, can the company that built it have any say in what the system cannot be used for?

The Pentagon's answer was effectively no. It designated Anthropic a supply chain risk, a classification designed to keep adversary-linked technology out of defence systems, on the reasoning that technology acquired for national defence should be available for any lawful purpose rather than remain subject to a supplier's policies. A federal judge blocked the designation in March, finding that nothing in the governing statute supported the "Orwellian notion" that an American company could be treated as a potential adversary for disagreeing with the government. Then the direction of control reversed.

On June 12, 2026, after learning of a technique for bypassing Fable 5's safeguards, the US Secretary of Commerce wrote to Anthropic restricting foreign nationals worldwide, including its own employees, from accessing Fable 5 and Mythos 5. Anthropic disputed the finding's severity, calling the vulnerabilities relatively simple and identifiable by other publicly available models, and warned that recalling a model over such a narrow jailbreak would effectively halt new frontier-model deployments. Because the company could not verify every user's nationality in real time, it withdrew access to Fable 5 and Mythos 5 globally. Users and institutions with no role in the dispute, and no means of contesting its consequences, abruptly lost access.

Nineteen days later, the controls were lifted, and access to both models was restored globally from July 1. Anthropic said it had worked with the government to review the concern and add safeguards; Commerce Secretary Howard Lutnick said his department had worked with the company to "analyze and approve" the model. Neither decision came with a published standard.

As of October 2026, the Pentagon dispute, by contrast, remains unresolved. A California court set the designation aside in August, but in September a federal appeals court upheld a separate exclusion of Claude under a broader law that requires no harmful intent, without disputing that Anthropic meant no harm. The off switch, in other words, ran in both directions.

Neither side's claim of authority is inherently incoherent. A government responsible for national defence cannot easily accept a private supplier retaining unilateral control over how an acquired capability may be used. But neither is it obvious that acquiring an AI system should erase every

DOI: 10.5281/zenodo.23666309

constraint imposed by the company that built it, particularly when those constraints concern autonomous weapons, surveillance or other uses with profound public consequences.

Under the Pentagon's logic, however, a supplier's ethical red line has force only for as long as the government accepts it. The administration subsequently formalised this principle in National Security Presidential Memorandum 11, directing agencies to reconsider or terminate relationships with suppliers whose restrictions were judged incompatible with national security requirements.

The timing of the Commerce letter made the underlying issue unusually clear. Days earlier, Anthropic's chief executive, Dario Amodei, had argued that governments *should* possess the authority to block or reverse a frontier model's deployment when independent assessment finds its risks unacceptable. He maintained that this process should be transparent, grounded in technical evidence, and protected against arbitrary or politicised decisions. When the Commerce letter arrived, therefore, the government was exercising precisely the kind of power Amodei had argued states sometimes need. The disagreement was over the conditions that make its exercise legitimate. A genuine technical safeguard had been bypassed, but Anthropic later argued that there was no agreed industry standard against which the severity of the jailbreak, or the government's response to it, could be judged.

Amodei's own case for government intervention had included conditions intended to constrain that power. The Commerce letter offered no comparable framework. There were no published thresholds, no independent review process, no advance notice to those affected, and no clear route through which the decision could be challenged. These events also gave new urgency to the expanding Sovereign AI movement. Governments describe domestic models as protection against dependence on foreign companies and policy decisions, although many sovereign projects still rely on foreign partners and imported infrastructure. And national ownership is not equivalent to public control: citizens do not necessarily gain greater influence over a model merely because it is hosted within their country. Sovereignty may relocate the off switch, but without accountable institutions it does not democratise the decision to use it.

The ethical problem is procedural. The US administration's June action on advanced AI innovation and security contemplated classified benchmarking and governmental risk thresholds, but established no public procedure governing the resulting authority. Governments should publish the conditions that trigger intervention, provide notice where circumstances permit, allow affected companies and governments to seek review, and disclose the terms on which access can be restored. A restoration negotiated in private between one company and one official does not correct the original problem; it is the same problem running in reverse.

These measures would not prevent an urgent national security decision. They would make it answerable to those expected to live with its consequences. An off switch may sometimes be necessary, but authority over it cannot be ethically legitimate when those governed by its use are permitted neither to understand the rules nor to contest their application.

DOI: 10.5281/zenodo.23666309

MAIEI

6.2 The Many Hands Problem: Securing Accountability in the Age of AI Warfare

By Ayaz Ul-Huq Syed , The Dais at Toronto Metropolitan University

In February 2026, over 150 civilians, including 123 children, 26 school staff, and seven parents were killed when an alleged United States precision-guided missile struck an elementary school in Minab, Iran. Officials involved in the Pentagon's internal investigation told Bloomberg that an overreliance on AI within a compressed timeframe were among factors contributing to the cascade of failures resulting in the CentCom decision to fire. The AI in question was Palantir's Maven Smart System. In Ukraine, Palantir has been providing military forces with AI-enabled platforms with similar real-time targeting and operational planning capabilities. Concurrently, Microsoft and AWS supply the compute infrastructure enabling large-scale data processing, logistics, and intelligence analyses in Gaza, where +972 Magazine and Local Call reported that the IDF utilised its Lavender AI-system to identify approximately 37,000 Palestinians as suspected militants and marked their homes as targets, with final human-in-the-loop approvals made in seconds despite a reported 10 percent error rate. States are increasingly relying on private vendors to produce essential military technologies, and the result has been the transformation of conflict theatres into laboratories deploying and testing – through a Silicon Valley-informed ethos of failing fast to learn fast – how AI-systems identify targets and inform decisions over life and death. With deliberations over the regulation of AI weaponry progressing little over the past decade, the chief consequence of the widening gap between proliferation and regulation has been a growing number of innocent dead, with little clarity on who is accountable.

AI with varying degrees of autonomy to identify, track, select, and engage targets, also known as Lethal Autonomous Weapons System (LAWS), suffer from a black-box problem. Their behaviors cannot be predicted or replicated reliably, which undermines accountability under international humanitarian law (IHL) by making it difficult to prove the intent necessary to assign responsibility for actions violating IHL principles of distinction and proportionality (Demir, 2025). While concepts like meaningful human control are intended to mitigate the issue of machines bearing no responsibility, automation bias during time-sensitive, high-pressure situations undermines such solutions when few guardrails exist. The 2026 Minab strike embodies the "many hands" problem states are running up against: according to investigator findings, while the targeting database listed the site as a military facility, commercial satellite imagery from 2017 and 2018 showed it had become a school. The change was reported by an intelligence analyst in 2019, however the report never reached the primary database from which decision-making suggestions were made. With staffing on civilian harm mitigation slashed by roughly ninety percent, one person was responsible for reviewing targets from an AI-system which did not surface any contradictions. Nobody within the direct kill chain did anything identifiably wrong – and 156 innocent lives were lost as a result.

DOI: 10.5281/zenodo.23666309

The High Contracting Parties to the Convention on Certain Conventional Weapons (CCW) established the Group of Governmental Experts (GGE) in 2016 with a mandate to strengthen the CCW to address LAWS. However, persistent disagreement over definitions, scope, and processes delayed progress toward establishing a binding instrument for over a decade (Demir 2025). Great powers actively utilising or developing LAWS, like Russia, China and the US, have each argued in favour of flexible definitions preserving operational leeway, impeding consensus and diluting terms. A key variable contributing to this can be attributed to geopolitical competition, as powerful states are often emboldened to maintain developmental pace against rivals (Schliebs and Vohs, 2026). Consequently, 2026 has seen various technology companies signing contracts with States to deploy AI across classified military environments. In the case of the US government, this includes Amazon, Microsoft, SpaceX, Oracle, Nvidia and Google.

Corporate self-regulation however, remains an insufficient mechanism for maintaining accountability. In addition to disincentives like market competition, technology companies are susceptible to unique pressures (Lamensch, 2026). For example, in September 2026, documents obtained by the Intercept following a Freedom of Information Act lawsuit highlighted the Pentagon's request from OpenAI for military prototypes with minimal refusal rates. While OpenAI stated that the term did not survive into the signed contract, the issue remains that the US government issued compliance-based procurement requirements from a frontier lab.

As responsibility remains diffused among actors (Schwarz, 2021), ensuring that a range of due-diligence obligations are imposed at each stage of design, development, and deployment will be among the pathways toward holding states, parties to conflicts, and individuals to account. One obligation, which follows from the Minab strike, is to mandate that targeting systems surface the provenance and age of the records they rely upon, ensuring no strike can proceed on a record whose last human verification exceeds a reasonable interval. By ensuring strikes are prohibited upon such a record, not only can automation-bias be precluded, but chains of command resulting in a decision to fire can be scrutinized against IHL principles.

With the GGE concluding talks in September 2026, and the Seventh Review Conference of the CCW to occur on November 16-20, 2026, States will determine whether to negotiate a legally binding instrument governing LAWS. Recommendations for such an instrument are contained in September's Final Report, however, when comparing earlier GGE Texts against it, certain recommendations appear reduced in scope.

For example, Paragraph 40 and 41 remove accountability from States and parties to armed conflict, conferring it upon individuals instead (with states recommended to take measures to ensure it). Paragraph 43 was reduced from placing a duty upon States to take all appropriate steps to prevent and suppress violations of international law, including IHL – to a duty of preventing violations of State obligations under international law, leaving how and whether violations apply to stages of design, development and procurement unclear.

DOI: 10.5281/zenodo.23666309

MAIEI

Responsible political leadership adhering to international obligations addressing critical gaps, such as bindings mandates ensuring departments are staffed appropriately to maintain meaningful human control (Siddiqui, 2026), encouraged by civil society actors, scholars, and humanitarian practitioners, can ensure that rulesets are robust and enforceable (Razmetaeva, 2026). But such obligations are hampered when powerful states throw a wrench into consensus building. So far, 130 countries have expressed support for a binding instrument, but if the High Contracting Parties cannot reach an agreement in November, the international community must find alternate routes to move forward. One contingency noted by Stop Killer Robots is found in CCW Article 8, whereby a conference to negotiate a binding protocol can be convened if the majority of states agree to it. Ethicists also highlight the moral imperative for local actors to engage in direct action to achieve progressive gains (Berglund 2025). Beyond reputational concerns and market access, coordinated pressure from civil society can encourage top-down obligations for decisionmakers to hold developers as well as personnel within chains of command to account through punitive measures (Schliebs and Vohs, 2026).

While the development of norms do not preclude belligerent or deviant actors from breaking them, nor signatories of future treaties from formally withdrawing (as has been the case with the Ottawa Treaty), they still point to international consensus pertaining to their legitimacy. It is too late to undo the integration of LAWS into military contexts, but a diffusion of responsibility ought not to forestall the international onus of determining how and when to hold states, organisations, and individuals to account for their contributions towards utilising atrocity weapons recklessly.

DOI: 10.5281/zenodo.23666309

6.3 Hide and Seek: The Paradox of Representation in Military AI

By Shimona Mohan, United Nations Institute for Disarmament Research (UNIDIR) & Katherine Chandler , Georgetown University

Despite its ubiquitousness, AI has many faults. It is well-documented that AI-based systems could or have explicitly shown biased results in a variety of settings, including recidivism software, hiring algorithms, credit services, AI-assisted medical care and facial recognition technologies. Biases on the basis of demographic identifiers like gender, race, age, etc can be amplified due to inaccurate or misrepresentative datasets, algorithmic processing, or human oversight and result in unfavourable outcomes ranging from reduced quality of services to lethal errors. Various actors have attempted to fix biases in AI applications, including in the military domain. The discussion on biases has been on the agenda of the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts (GGE) on Lethal Autonomous Weapon Systems (LAWS) and in related working papers. Initially, this covered gender and racial biases, but as of its 2024-2025-2026 sessions' final report adopted in September 2026, it addresses measures to "prevent and mitigate, as feasible, unintended bias in the functioning of LAWS, including bias in datasets and AI models." The Group has recommended that the Seventh Review Conference endorse that set of elements when it meets from November 16-20, 2026, days after this volume is published.

Technical and policy-oriented discourse has advocated for better datasets, both human-generated and synthetic, to better reflect human diversity across the globe. However, this recommendation fails to account for an insidious but pertinent representation paradox, especially in the context of military applications of AI—**if AI systems get better at perceiving minority or marginalized populations through the reduction of biases, is that objectively and overarchingly better or worse for them at scale?**

Just because populations are better represented within AI datasets does not ensure that they will be protected or historic inequities will be remedied. Invisibility to AI systems can have fatal consequences in military contexts, particularly if civilian infrastructures are mislabeled or misrecognized. The debiasing of military AI, where possible, creates new hurdles for its ethical and responsible usage. Representation in one context might be transformed into a target in another. An overwhelming majority of populations who are currently underserved by the ways in which AI systems represent and make conclusions and/or decisions about them, are the same people who have been historically persecuted, oppressed or otherwise disadvantaged on the basis of their identity.

Marginalized groups experience new harms because of their relatively increased visibility in datasets and AI that can connect ordinary life with government oversight and national security.

DOI: 10.5281/zenodo.23666309

For instance, women using femtech AI-based healthcare services such as period-tracking apps could be charged higher insurance premiums based on data gleaned by insurance companies from these apps, or prosecuted for availing abortion services in jurisdictions where this is restricted. As the incorporation of AI increases across tools and services, it can become easier to survey sensitive data such as people's medical histories, choices, and lifestyles. At the same time, failure to be recognized within AI systems also can have damaging consequences. Recent lawsuits filed by persons with disabilities against AI models indicate how their invisibility can have damaging effects on their quality of life and reproduce inequalities. Persecution, oppression and disadvantage against population groups can continue unabated – not due to lack of data, but because the tools can reproduce and reinforce inequalities depending on how they are used.

Applied to a security or military context, more accurate facial recognition that works well on diverse groups of people could lead to greater surveillance and even targeting. An extensive camera network designed to surveil the population in Iran was compromised, and feeds from it were reportedly used by Israel to help plan military attacks inside the country. Modern cameras are also equipped with facial recognition, which can be repurposed at scale in such a scenario. The same infrastructure that makes people visible also makes them targetable to adversaries. Even if military AI systems like LAWS are currently stipulated to not be able to target and engage humans autonomously, they operate in an environment with human combatants and civilians who may face greater risk of targeted attacks and false positive identification solely because the systems can now perceive them.

Apart from targeting, military applications of AI could have the potential to create new strategies of camouflage and human shielding. Data itself could also become a tool for war through strategies employing adversarial machine learning, for instance in the case of data poisoning, wherein training data for AI models is mislabeled or manipulated to create misidentifications. Determinations about a potential military target like their gender or age that are intended to protect civilians and civilian installations, could be subverted, weaponized, or exploited to disguise militants or military objectives.

We thus come to the crux of the paradox—**neither being invisible, inaccurately visible or accurately visible to AI systems is inherently better for many people across the globe**. We propose a new process of evaluating inequality in AI, which considers debiasing as a relative, iterative and context-dependent activity. Robust methodologies that transparentize the black box could help, since the artificial over-correction of biases (which is possible in black box setups) can also introduce errors. Effective mitigation of biases in AI relies on efforts to audit how AI reproduces existing inequalities and can exacerbate its impacts, particularly in military contexts.

Rather than being understood as a one-time design-phase checkbox, AI developers and regulators must approach mitigating AI biases contextually and responsibly as part of the

regular maintenance and upkeep of AI-based systems. This means that biases should be addressed at multiple stages in the development of AI systems, from data collection to model training and testing to its deployment and retirement.

Importantly, to limit harm, we need to address how AI intersects with existing inequities and develop limits that protect people, regardless of whether they are accurately identified - or not - by AI. Mitigation of biases is not just a technical problem with a one-size-fits-all approach, and its success is not merely dependent upon better pattern recognition or more data on marginalized populations to overcome historical and structural inequities.

DOI: 10.5281/zenodo.23666309

6.4 Towards a "Smart Mix"?: Governing AI for International Peace and Security

By Branka Marijan, Project Ploughshares

A core challenge in governing artificial intelligence (AI) for international peace and security is that governance responses must increasingly address both military and civilian applications. Many AI technologies are dual use, systems developed for civilian purposes that can also serve military functions. Moreover, civilian applications can affect peace and security without being used by militaries, including through information manipulation and malicious cyber activity or AI agents acting beyond their intended boundaries. These overlapping challenges require contributions from institutions with different mandates and expertise. As such, the resulting patchwork of governance efforts is unsurprising but whether its elements reinforce one another is more consequential.

John Ruggie's "smart mix" offers a useful approach for governance of AI for international peace and security. Developed in the context of business and human rights, it combines voluntary and mandatory measures at national and international levels. Applied to AI, it directs attention to how diplomatic commitments, legal obligations, technical standards, and corporate practices interact. Crucially, voluntary initiatives cannot compensate indefinitely for the absence of necessary binding rules, but they can provide scaffolding while those rules are negotiated.

Three developments in 2026 illustrate both the potential and shortcomings of this approach. Namely, the REAIM Summit in February, the UN Global Dialogue on AI Governance in July, and the September conclusion of the latest round of autonomous weapons discussions.

Military AI and Beyond

The Third REAIM Summit (Responsible Artificial Intelligence in the Military Domain), held in A Coruña in February, demonstrated the value of convening governments, industry, researchers, and civil society around military AI. The REAIM agenda extends beyond autonomous weapons to the wider implications of AI for military activities and human judgment.

Yet convening does not establish implementation. Assessing the summit shortly afterwards, Manoj Harjani and Mei Ching Liu noted that its outcome document, Pathways to Action, attracted fewer endorsements than those of previous summits, with neither the United States nor China endorsing it. REAIM's contribution therefore lies partly in sustaining multi-stakeholder exchanges and developing practical approaches despite political disagreement. Its effectiveness still depends on whether participants change policies and practices, and whether countries with fewer resources can meaningfully shape that work. The next summit, to be held in Kenya in early April 2027, will offer a further test of what this multistakeholder dialogue can achieve.

DOI: 10.5281/zenodo.23666309

Additionally, the first UN Global Dialogue on AI Governance, held in Geneva on July 6–7, widened participation, bringing together participants from 170 countries. However, breadth of participation did not resolve the separation between civilian and military governance. In her account of the meeting, Branka Panic reports that military AI and autonomous weapons were absent from substantive discussion despite being raised in the Secretary-General's opening remarks. Moreover, Panic notes that the intersection of AI and peace was also not on the agenda.

The Missing Mandatory Element

The hardest test for the smart mix approach concerns autonomous weapons. In early September, the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts (GGE) concluded its work without a binding instrument or an agreed mandate to negotiate one. Its task had been to formulate elements of an instrument without prejudging its nature. The resulting text provided a foundation for further work, while decisions about negotiations remained for the November Review Conference.

This was limited progress, and after more than a decade of discussions, the absence of agreed new binding rules remains a serious shortcoming. To be fair, as several scholars note, amidst the current geopolitical turmoil, the meeting did feature some 128 countries with 70 willing to move to negotiations. Still, a smart mix cannot be judged successful simply because voluntary initiatives continue alongside stalled legal negotiations. In late September, the Washington Post reported that the United States and Russia had removed several safeguards from the text, notably those concerning human oversight, further complicating prospects for binding rules.

Other Ingredients in the Mix

A concrete example of connecting actors emerged at REAIM, where UNIDIR and the Office of the High Commissioner for Human Rights launched their initiative on responsible industry behaviour in military AI. Contributors include Microsoft, Hitachi America, and the Japan Defense Technology Foundation. The initiative connects industry expertise with expectations for responsible procurement and corporate conduct. However, UNIDIR explicitly states that contributing does not imply endorsement or binding commitments.

The next step is to make such guidance meaningful. Governments should translate relevant expectations into procurement requirements, including documentation, testing, and incident reporting. Investors should examine whether companies can demonstrate these practices. Already, investors are developing guidance on defence investment and furthering investor education and understanding of issues. These measures can help connect voluntary guidance with decisions over contracts, capital, and oversight.

DOI: 10.5281/zenodo.23666309

MAIEI

Ways Forward

The 2026 record shows an emerging mix whose connections remain incomplete. REAIM linked diplomatic discussion with practical industry work, the Global Dialogue widened participation while leaving military questions largely separate, and the CCW developed common elements without securing a path to binding negotiations.

Progress now requires identifiable responsibilities. States must pursue necessary legal rules, procurement officials must specify and verify requirements, and developers must provide evidence supporting their claims. International forums should examine implementation and feed findings into subsequent decisions. A smart mix earns its name when these relationships preserve human judgment, enable accountability, and help prevent harm. So far, the approach is valuable, but the practice is challenging.

DOI: 10.5281/zenodo.23666309

Chapter 7: Oversight Built for Another Era

7.1 The Stadium as Testing Ground: How the World Cup Risks Exporting the Authoritarian Playbook

By Taras Kovalchuk, Center for AI and Digital Policy & Journal of AI Law and Regulation

7.2 Canada's Lobbying Act Was Built for a World That No Longer Exists

By Renee Black, GoodBot & Canadians for Digital Sovereignty & Kaitlin Bustos, GoodBot

7.3 The Algorithmic Cost of Living: How Your Personal Information Will Be Weaponized Against You

By Emily Osborne, Canadian Shield Institute

DOI: 10.5281/zenodo.23666309

7.1 The Stadium as Testing Ground: How the World Cup Risks Exporting the Authoritarian Playbook

By Taras Kovalchuk , Center for AI and Digital Policy & Journal of AI Law and Regulation

This summer, North America hosted the 2026 FIFA World Cup. The tournament unfolded against a backdrop of major research organizations documenting significant democratic backsliding in the US, Canada, and Mexico. And while the event concluded without major security incidents, it also provided an opportunity to deploy, and in some cases expand, AI surveillance.

Two Cities, Two Different Approaches

Atlanta and Vancouver both hosted the tournament, but they took different approaches to surveillance. Atlanta set up AI-powered facial recognition at Mercedes-Benz Stadium for World Cup entry and payment, while federal funding for World Cup security allowed the city to expand its drone units and camera networks. As reported by Tech Policy Press, the camera network was not used solely to watch soccer fans. Atlanta police used Flock Safety's license-plate reader system to run searches for federal immigration enforcement. The Atlanta Community Press Collective documented this in November 2025, before the tournament, so the network was already in use and World Cup money helped expand it. Digital rights researchers who reviewed Atlanta's 57-page action plan found only a single paragraph that mentions surveillance, and even that appears in passing under a section on environmental impacts. The document does not address how surveillance technologies affect human rights, or what happens to the equipment once the tournament ends.

Opposition has been mounting. The Atlanta Journal-Constitution reported that Atlanta police were taking steps to address the misuse of Flock cameras, while the mayor proposed new rules for the city's surveillance tools and Flock announced several changes. In February, TechCrunch reported that people across the United States had started destroying Flock cameras. Atlanta has now become a real-world test of whether these surveillance systems can be dismantled.

Vancouver took a different approach. There is no evidence the city used AI in its World Cup security system, and it banned facial recognition, the automated identification of individuals, and biometric analysis. It was also reportedly the only host city to publish a timeline for decommissioning the extra cameras after the tournament. The Vancouver Police Department even worked with the Office of the Information and Privacy Commissioner for British Columbia on a privacy impact assessment. Vancouver was willing to take extra measures to protect human rights, a choice that stands out because temporary surveillance rarely stays temporary.

DOI: 10.5281/zenodo.23666309

The Weight of Precedent

Sport events serving as a pretext for intensified surveillance like we saw in Atlanta is nothing new. In 2018, Russia hosted the World Cup and Moscow launched a facial recognition pilot near venues and transit stations to identify wanted criminals. When the tournament ended, the cameras remained. Over time, Moscow's "Safe City" network expanded from that pilot into one of the world's largest AI surveillance systems. What began as a temporary exception became part of the city's permanent security architecture. Belgrade, Serbia launched a similar "Safe City" initiative, installing 1,000 AI-powered surveillance cameras with facial and license-plate recognition. Serbian civil society has spent years challenging the system's legality and pushing for its rollback, largely without success.

The approach to surveillance adopted in cities like Atlanta has generated concerns regarding the US's democratic foundations. The numbers support those concerns. For the first time in over fifty years, the V-Dem Institute's latest report stripped the United States of its liberal-democracy classification, noting that the pace of decline has been unprecedented in modern history. Transparency International similarly found that the US and Canada have slipped noticeably from their earlier standing, with the US falling to its lowest score on record. What makes this shift so hard to detect is that it often unfolds gradually, before people realize it is already underway. As Steven Levitsky and Daniel Ziblatt argue in *How Democracies Die*, the very institutions of democracy are often used to build authoritarianism, and it happens slowly, quietly, and even legally.

The risk is highest when a growing share of the population feels that democracy simply isn't delivering. The 2023 Open Society Barometer found that 35 percent of 18-to-35-year-olds worldwide would accept a strong leader who did away with elections altogether. In the US specifically, a 2025 Tufts survey found that 31 percent of young Americans (ages 18–29) do not value core democratic principles and have little confidence the system works. A public that doubts democratic institutions will respond is less likely to demand that cameras come down, leaving that decision largely to governments.

Responsible AI Governance: Choosing Rights Over Sunk Costs

The fact that trust in democratic institutions has been on the decline in North America is one of the reasons why it is so important for the government to now demonstrate responsible AI governance. In the meantime, privacy advocates are raising concerns that surveillance infrastructure deployed for the World Cup in Los Angeles, including AI-driven analytics and license plate readers, will be retained for the 2028 Summer Olympics. If Los Angeles keeps that infrastructure, the practice will have spread from Moscow in 2018 to Atlanta in 2026, and then to Los Angeles in 2028, each time in connection with a major sporting event. The next two years will reveal whether Los Angeles follows Vancouver's or Atlanta's path.

DOI: 10.5281/zenodo.23666309

MAIEI

While it remains to be seen whether surveillance infrastructure built for the World Cup will ever be fully dismantled, it is fair to conclude that in a democratic society, AI governance requires the political courage to decommission surveillance tools capable of infringing upon human rights, even if doing so means a significant amount of money becomes a sunk cost.

DOI: 10.5281/zenodo.23666309

7.2 Canada's Lobbying Act Was Built for a World That No Longer Exists

By Renee Black, GoodBot & Canadians for Digital Sovereignty & Kaitlin Bustos , GoodBot

Since 1989, Canada's Lobbying Act has mandated a public registry to track individuals seeking to influence government policy. While it offers some transparency, it was not built to show who is shaping AI rules in Canada or to monitor an expansive, well-resourced, and rapidly evolving influence industry. This gap quietly reshapes whose voices are heard in decisions on governing AI, data, and digital markets, and public interests are often erased. Addressing these asymmetries requires understanding influence at every stage: agenda-setting, drafting, consultation, and implementation.

Setting the Agenda

Before a bill is drafted, someone must decide how to frame the problem. Big Tech's advantage starts here, before any registered lobbyist walks into a minister's office. Technology firms fund the think tanks, research centres, and policy papers that define which issues top the agenda.

A Canadian nonprofit recently withdrew from the Safer Online Spaces coalition after receiving over $1 million from Google. Canadian researchers describe a "soft power" operating through university funding; one University of Toronto PhD student was told outright not to press criticism of Big Tech in his work. This leaves behind independent evidence that could shape policy agendas never gets produced.

Drafting and Consultation

This is where the Lobbying Act is intended to apply, yet it contains significant gaps. Registration is only required for paid lobbying above a threshold, currently eight hours of communication with officials in a rolling four-week period. Companies can spread lobbying duties across employees to stay under that line ("threshold gaming"), leaving no trace in the registry. Communications initiated by government departments, like stakeholder briefings, fall outside the Act's scope entirely, even though these are exactly where policy substance gets shaped informally. Unlike the U.S., Canada requires no disclosure of lobbying spending, so the scale of effort on any bill can't be gauged. The Lobbying Commissioner recommended expenditure disclosure in its statutory review, but Parliament has yet to act.

Then there's what doesn't touch the registry at all: proxy voices. During Bill C-11 debates, a creator-advocacy group partnered with TikTok and took a position closely mirroring the platform's own concerns. None of this requires registration, not because it lacks influence but because the law's narrow definitions don't recognize proxy advocacy as lobbying.

DOI: 10.5281/zenodo.23666309

MAIEI

Consultation processes meant to level the playing field can reinforce the imbalance. Canada's federal AI task force drew members skewed heavily toward industry, alongside a public submissions process that researchers say lacked safeguards against industry-aligned talking points being amplified at scale. The generative AI Code of Conduct governing the sector was drafted directly with the companies it's meant to constrain, yet none of the largest U.S. AI firms have signed on. Canada's failed Bill C-27, the Digital Charter Implementation Act, which contained the Consumer Privacy Protection Act, included no public interest engagement before being introduced.

Durable Access

Beyond formal consultation, companies build channels of access that outlast any single bill. The "revolving door" is loosely constrained: a five-year cooling-off period covers only a narrow category of officials, who can still lobby informally below registration thresholds, while executives moving into government face no cooling-off period at all. Notably, in September 2026, former Google CFO Patrick Pichette was named inaugural CEO of Digital Transformation Canada, a new federal organization mandated to use digital and AI technologies to help Canadians access services.

Implementation and the Long Game

After a policy passes, the asymmetrical influence intensifies, implementation unfolds over months or years, and only well-resourced firms can sustain engagement that long.

Public interest groups lack the resources and relationships to track regulatory design the way they track a bill's passage; companies invest heavily in this stage. Government procurement compounds this with dozens of contracts with a single cloud provider creating switching costs that are politically costly. More than 80% of the cloud infrastructure the Canadian government relies on is foreign-controlled.

International trade rules add a final, largely invisible layer. Provisions in the Canada-United States-Mexico Agreement (CUSMA) banning data localization and restricting disclosure of proprietary algorithms were negotiated with little public visibility, yet pre-emptively ruled out policy options a future government might consider. Industry groups have already invoked CUSMA's language to argue Canadian laws like the Online News Act violate the agreement.

The Compounding Effect

No single gap is decisive alone. Together, researchers argue, they form a self-reinforcing structure*: financial capacity* funds the research and proxies that shape debate; *relational access* embeds corporate perspective inside government; *structural dependencies* narrow what's left to negotiate by the time the public learns a decision is being made at all.

DOI: 10.5281/zenodo.23666309

MAIEI

Civil society groups, already operating with a fraction of the resources, face an asymmetrical disadvantage. One estimate suggests fewer than 4 in 10 can even afford a lobbyist, and often rely on the same corporate donors whose influence they might otherwise check. When independent research challenges industry positions, well-funded counter-narratives can dilute its impact before it shapes the debate.

Ultimately, this reflects a transparency framework built for a different era of politics, unable to reveal who shapes policy outcomes and to what extent. The Commissioner of Lobbying's March 2026 recommendations to Parliament's ethics committee made 21 recommendations to modernize the Lobbying Act. Although the "significant part of duties" threshold was updated through regulatory changes, the Lobbying Act itself remains unamended. Other recommendations, such as administrative monetary penalties and temporary lobbying prohibitions, remain unenacted. Even if these are fully enacted, they are only a starting point and may fail to capture the full scope of indirect influence today.

DOI: 10.5281/zenodo.23666309

7.3 The Algorithmic Cost of Living: How Your Personal Information Will Be Weaponized Against You

By Emily Osborne, Canadian Shield Institute

We are entering a future in which how much you pay for your groceries, how much your rent costs, and even how much you earn is tailored just for you via an algorithm. Price discrimination is not a new phenomenon, but the massive amounts of data available and the ability to analyse it at scale enable it to take the more worrying form of "surveillance pricing". We already have the tools we need to deal with unfair price discrimination, it is just a question of applying them.

Surveillance Pricing

Surveillance pricing uses personal demographic or behavioural data to determine the maximum price a consumer is willing to pay for a given service or good. This pricing model creates a market that is no longer dictated by supply and demand, but instead by what a firm knows about you and what they can infer from that information.

Firms can now leverage anything from direct consumer or behavioural data, inferences based on a consumer's behaviour—say how you are interacting with a web page—or data purchased from other companies, including data brokers, to determine what you should pay. It is not an efficient system by which all consumers pay the price that they are able to and is designed to erode any consumer surplus. While the price variations of a given item may be small, even imperceptible, they will compound into a real economic cost over time for consumers, and make accurate budgeting and price comparison impossible.

Loyalty discounts are another vector by which firms can gather increasing amounts of data about their consumers, all while disguising this surveillance behind a thin veil of "rewards". Many bills passing in the U.S. to regulate surveillance pricing have even made carve-outs for subscription-based programs, like loyalty programs. So far, Canada seems set to follow: our AI Minister cited concerns about penalizing members of rewards programs as a reason for caution in regulating surveillance pricing.

Any marketplace where there is more obscurity around prices provides an especially easy venue for price manipulation. Tinder, for example, was found charging different users up to five times more than other users for the same services. Surveillance pricing even targets children in the closed marketplaces of free-to-play mobile games, enabled by the collection of personal information—as forthcoming research by the Canadian Shield Institute and funded by the Office of the Privacy Commissioner of Canada (OPC) is examining. And agentic AI in commerce is adding another layer of obscurity that makes the ability to comparison shop even more elusive: recent research found that Google AI mode showed products at a price that was on average 21.6% more expensive than on traditional search.

DOI: 10.5281/zenodo.23666309

Algorithmic Wage Discrimination and Algorithmic Rent-Setting

Algorithmic wage discrimination is enabled by platforms that allow a firm to set an individualized price for a worker's labour. Uber is the most classic example but its model is spreading: some on-demand nursing scheduling apps are offering different compensation for shifts based on personal data. Data like a nurse's credit card or other debt is used to determine how desperate they are for work, and therefore how *little* they are willing to be paid. We should expect this model of exploiting personal data for economic leverage to be adopted across other employment sectors.

A number of bills in the U.S.— in California and New York for example— have targeted the related phenomenon of algorithmic rent-setting, by which algorithms analyze data to recommend the most ideal rental price or lease terms. Thus far, indications suggest that the data being used is primarily historical rent prices and other tenancy terms, but the potential for more personal data sources to eventually be weaponized against tenants is clear when price-setting is being off-sourced to algorithms.

Though it is sometimes disguised as efficiency, the point of these technologies is always the extraction of the maximum value, whether from consumers, workers or renters.

What Can We Do?

In 2026, Manitoba became the first Canadian province or territory to explicitly ban surveillance pricing via Bill 49, which came into effect on July 1st. The bill amends the Business Practices Act to include "personalized algorithmic pricing" as an unfair practice—with hefty fines for offenders. Quebec takes a different approach. Its privacy law has much longer contained a provision on profiling that ostensibly would capture surveillance pricing—though it hasn't been formally tested.

Most recently, the federal government sought to address surveillance pricing through an update to its privacy legislation via Bill C-36—its third attempt at private sector privacy reform. The bill contains no explicit provision on surveillance pricing. Any further rules around the practice established by the Digital Safety Commission, which the bill would set up if it passes, are unlikely to be in effect before 2028. Canadians remain unprotected.

While it is tempting to treat surveillance pricing as only a consumer protection issue or a transparency problem that can be solved with disclosure, it is fundamentally a privacy problem and the federal government is right to address it as such. Firms are only able to set prices according to personal data because they have the ability to access massive amounts of that data in the first place in a largely unchecked manner. Consumer protection laws and other interventions are certainly welcome at the provincial level, but will not address the problem at the root, nor will they protect Canadians from other forms of unfair data exploitation.

DOI: 10.5281/zenodo.23666309

However, it will not be enough to simply require consent or to make broad carve-outs to protect targeted discounts—as the federal government seems poised to do. Instead, privacy legislation must recognize that the use of personal information to facilitate unfair practices, like surveillance pricing, is simply not an appropriate purpose. Exceptions must be kept to a minimum and clearly specified, rather than broad.

DOI: 10.5281/zenodo.23666309

PART III: LABOUR & INFRASTRUCTURE

Chapter 8: AI at Work and the Workers Bargaining Over It

Chapter 9: Data Centres and Communities

DOI: 10.5281/zenodo.23666309

Chapter 8: AI at Work and the Workers Bargaining Over It

8.1 Contract Bargaining in the Age of AI

By Crystal Braunwarth & Hillary Bergshoeff, IATSE 891

8.2 The Clause and the Algorithm

By Shi Kang'ethe, Independent Journalist

8.3 When Workers Bring AI to Work (From Private Experiment to Organisational Capability)

By Robert Schouten, Atlas Research and dBSC

8.4 When a Profession Is Automated: Translators and What Comes Next

By Debbie Folaron, Concordia University, Lynne Bowker, Université Laval, & Joss Moorkens, SALIS/ADAPT Centre at Dublin City University

DOI: 10.5281/zenodo.23666309

8.1 Contract Bargaining in the Age of AI

By Crystal Braunwarth & Hillary Bergshoeff, IATSE 891

Since the labour movement began, collective bargaining has been its central instrument. A contract is a legal framework: it defines working conditions and benefits, sets limits on management overreach, and establishes clear steps for discipline and accountability on both sides of the table. By bargaining together, we face management as one, and each agreement builds on the foundations of those won before.

For IATSE 891, our relatively mature agreement has seen language grow each cycle, as new classifications join our ranks in the increasingly diverse world of motion picture production. We've faced the introductions of sound, colour, green screen, computer animation, volume production, and 100 other things head on, knowing we remain the artists and technicians leveraging the tools to tell the stories.

As leaders in the BCFED Standing Committee on AI and New Technologies, we are hearing that across sectors in BC, in education, healthcare, manufacturing, and the creative arts; employers are refusing to bargain artificial intelligence language at all. While the industries are different, the reasoning provided is the same: the technology is "moving too fast" and "no one knows where it's going," with many employers refusing to even begin bargaining until AI is "off the table." This leads us to ask if speed is really the issue, or if the aim is to delay negotiations until AI rollout is inevitable, meaning worker rights can be eroded before any language is put in place. Everything we are hearing from our labour kin points to delay.

In 2025, the BC Council of Film Unions (BCCFU), which brings together IATSE 891, Teamsters 155, and ICG 669, negotiated a standalone AI article into the 2025–2028 Master Agreement with the Alliance of MotionPicture and Television Producers (AMPTP) and the Canadian Media Producers Association (CMPA) following IATSE International's own negotiations on artificial intelligence. It is Article 17, and it applies to work from April 1, 2025.

In any bargaining process, clear definition of terms is critical. Article 17 opens with specific, functional, and non-anthropomorphized language: the parties acknowledge that "Artificial Intelligence" and "AI" "have become catchall names," define an AI System by what it does, and provide that the definitions apply to any technology consistent with them, "regardless of its name or designation." It confirms that while employers retain the right to require employees to use an AI system, employees stay covered by the Master Agreement while doing so; and an employer may not require an employee to feed prompts into a system in a manner that displaces that employee's own bargaining-unit work.

DOI: 10.5281/zenodo.23666309

Further, employers must disclose their written AI policies to the BCCFU and employees are entitled to consultation before being required to use a system, where production timelines allow. An additional clause requires clear, specific, and freely given consent (never as a condition of employment) before any employee's visual or vocal likeness may be scanned, with a reasonably specific description of the intended use and a copy to the union before it is put in front of the worker. And where a claim arises from an employee's use of an AI System in the course of their work, the employer indemnifies them against liability and costs, provides a legal defence, and names them as an additional insured on its errors and omissions policies.

However, the article's most important feature is not any single clause. No single clause could be a standing structure built to keep pace with technology that is, in fact, still changing. Instead, it commits the parties to convene a joint training committee focused on both operating AI tools within an existing classification and transitioning workers into new ones, and requires that committee to meet within ninety days of ratification. It further requires semi-annual meetings between the BCCFU and a senior executive from each signatory company "involved in the development and/or use of AI," plus company-by-company sessions on request, under confidentiality, to review current and intended AI use. That is the direct rebuttal to "no one knows where it's going": the contract does not pretend to freeze 2025's technology in place. It builds a recurring conversation where both the union and employer revisit the evolution of technology as it changes, rather than using uncertainty as a reason to avoid the table altogether.

Even with provisions like the requirement to meet within 90 days of ratification to set guidelines and training pathways, it didn't happen with BCCFU, though we did pursue our own training for members. Busy employer bargaining schedules and a highly volatile industry environment pushed every consultative request to late 2026. Outside of calls about member AI complaints, there were no substantive one-on-one meetings. A core issue is that the companies we collectively bargain with are competing with each other for dominance in the creative space, so speaking openly about how they're using tech seems to be a huge risk for them. Enforcement of this language will be key moving forward. We can't enforce what we can't see, and we can't enforce what we can't even get a meeting to talk about.

Still, honest advocacy requires naming this agreement's limits alongside its gains. Violations are arbitrable, but injunctive relief is excluded and the arbitrator has no authority to prohibit or restrict use of any AI System or its output. Also, consistent with SAG-AFTRA and the WGA's 2023 agreements, the article secures no ban on training AI systems on members' own work product, and that fight will likely continue. Either way, these gaps are not evidence that bargaining AI is premature. They are the normal shape of a first contract on a live issue, exactly the kind of terms a second and third round of bargaining exist to tighten. Our labour lens is a human lens. And that is how we proceed to solutions.

DOI: 10.5281/zenodo.23666309

So, is there a solution? When an employer says AI is moving too fast to bargain, the accurate response is that a recurring bargaining table is the tool designed for exactly that problem, and IATSE, as a whole, in Canada and the US, is already using it. “Too fast to bargain” is not a technical constraint, it is a choice by employers to keep a governance vacuum open for as long as it is profitable to do it. Unions across sectors represented in this industry don’t need to wait for the technology to hold still. As always, we need to come together, in solidarity, and put a defined term, a consultation right, a training commitment, and a standing review meeting into our contracts, always through a lens of equity, bias, sustainability and a humanistic sensibility whereby we are the only ones leveraging technology, and not the other way around. In holding that solidarity, we will use the next round to close what the first round left open.

DOI: 10.5281/zenodo.23666309

MAIEI

8.2 The Clause and the Algorithm

By Shi Kang'ethe, Independent Journalist

Every algorithmic decision that affects a worker's pay, schedule, or standing was, at some point, supposed to be someone's problem to negotiate. Somewhere upstream, the thinking goes, a union sat across a table from an employer and wrote language that anticipated this moment: a clause on data use, on automated discipline, or one requiring that a human being review a decision before it costs someone their livelihood. Where that clause exists, a worker who is flagged, deprioritized, or deactivated by an algorithm has somewhere to stand. Where it doesn't, the algorithm's word is final, and there is no forum in which to contest it.

Collective bargaining is not an adjacent concern to algorithmic governance; it is one of the primary sites where that governance actually gets decided. Ethics boards, transparency reports, and regulatory frameworks matter, but a bargaining agreement is enforceable in a way that a corporate AI principle is not. It has a grievance procedure behind it, and a union willing to escalate. When we ask whether a worker can meaningfully contest an algorithm's decision, the answer usually has little to do with the sophistication of the algorithm. It has to do with whether someone, years earlier, had the standing and the foresight to write the right clause into the right contract.

That is precisely what makes the situation facing gig workers so acute, particularly across the African continent, and it is where this argument really belongs.

Kenya shows what this looks like when the law itself is the instrument. For three years, content moderators have been suing Meta and Sama, the contractor that employed them in Nairobi. They were paid as little as $1.50 an hour to review and rate content for removal, including videos of rape, murder and war crimes. If the case succeeds, it would establish that the platform, and not only its outsourcing partner, is liable for what happens inside a subcontracted facility. In December 2024, President William Ruto told a town hall in Nairobi that he was preparing to sign a bill to stop it. “Those people were taken to court, and they had real trouble,” he said of Sama. “They really bothered me. Now I can report to you that we have changed the law, so nobody will take you to court again on any matter.”

That bill is the Business Laws (Amendment) Bill, 2024 (Senate Bill No. 51 of 2024), an omnibus that amends everything from the Banking Act to the Scrap Metal Act. Workers say that Clauses 10, 11 and 12 grant immunity to the platforms and shift legal responsibility onto local BPO contractors, leaving content moderators, data labellers and gig workers with limited recourse. The process is the part worth watching. In March 2025 the workers asked to present their concerns to the Senate and were not heard. In April they filed four memoranda, none of which was taken up. In May the bill reached second reading without a committee report. On August 7,

DOI: 10.5281/zenodo.23666309

2025 the Senate passed it. Mercy Mutemi, who leads the cases against Meta and Sama, called it “a betrayal of the people who power Kenya’s digital economy.”

Thirty-five tech workers, supported by The Oversight Lab, petitioned the High Court to stop the National Assembly taking it further, arguing the Senate had passed it without the public participation the constitution requires. The bill was back on the National Assembly’s order paper on May 26, 2026 and July 28, 2026. They asked to be heard, they filed memoranda, and the bill advanced on schedule.

The scale of what falls into that gap is easy to underestimate, because so much of gig work is designed to be invisible. It runs on a huge, largely unseen labour force performing what has come to be called ghost work, the human judgment and correction that makes automated systems look smart and seamless from the outside. That workforce is now expanding in a stranger direction. In China, for example, underemployed lawyers, architects, and engineers are taking on gig contracts to teach AI systems the exact reasoning of their own professions, recruited by tech giants building high-quality training data from credentialed expertise sold at gig-work rates. In the United States, companies including Mercor, Surge AI, and Handshake have recruited tens of thousands of filmmakers, lawyers, and scientists to train models on white-collar tasks, and that number is still growing. What unites these workers, across geography and skill level, is not the nature of their labour but its legal status.

That absence is not incidental to gig work; it is what has allowed gig work to function so profitably. From its inception, gig work has stripped away protections that took a century of organizing to secure, a minimum wage, compensation for injury, paid leave, and any real control over one's own hours. It has done this while returning record profits to the platforms that built it, because the same technology that renders gig workers replaceable also renders them controllable.

Algorithms assign the tasks, set the pay rates, and evaluate performance, often without a named person on the other end of the decision. A worker in a formal, organized sector might have an unjust algorithmic call trigger a grievance. A gig worker, especially in a market where the platform is one of the only sources of flexible income available, has nothing to trigger. The question was never whether a clause was well drafted or poorly enforced. There was never a table at which to write one in the first place. There is only the decision, and the worker's task of absorbing it.

Set alongside the experience of organized labour, the shape of the problem comes into focus. Where a union has bargaining power, AI governance becomes a matter of what gets written into the next contract, however difficult that fight remains. Where gig workers have never been permitted to organize, AI governance becomes a matter of what a platform decides, unilaterally, to permit. These are not two separate stories about the future of work. They are one story, told from opposite ends of the same mechanism, a clause that exists to be enforced, and an algorithm that operates precisely because no clause was ever written to constrain it.

DOI: 10.5281/zenodo.23666309

MAIEI

8.3 When Workers Bring AI to Work (From Private Experiment to Organisational Capability)

By Robert Schouten , Atlas Research and dBSC

Generative AI did not enter many workplaces through a planned implementation programme. Employees simply began using it. In a 2026 report based on a 2025 survey of 3,845 members, Akavia, a Swedish trade union representing university-educated professionals, found that 49 percent used AI regularly at work, while only 2 percent said it was used throughout their organisation. Employees were already learning what the tools could do, where they fitted and what checks made their output useful. Policies, training and shared working methods often came later.

That gap is usually framed as a security or governance problem. It is also a labour issue.

When Learning Becomes Work

A general-purpose large language model does not know what counts as a sound legal note, a useful policy brief or an appropriate client response. Employees supply the situated knowledge. Through repeated use, they develop prompts, examples, checking routines, workarounds and rules about when not to trust the system. Much of this knowledge is tacit, rooted in professional experience, local routines and judgement. By deciding what the model can do, what must be checked and when its output should be rejected, employees also begin to reshape the workflow.

Dittmar and Sposato call this *shadow learning*: developing AI-related knowledge and practices outside formal organisational training and governance. This is not necessarily unauthorised AI use. Even with an approved tool, employees may still work out informally where it fits, what it can be trusted to do and how its output should be checked. They begin reshaping work before the organisation has formally decided how that work should change.

This learning is not separate from the job. It is part of the work required to make AI usable.

The French research programme LaborIA describes AI deployment not as the end of innovation, but as the start of a prolonged and uncertain learning process. Workers use, maintain, improve and supervise systems, yet organisations may fail to recognise these activities as work. Applied to generative AI, this includes refining instructions, supplying context, checking outputs, correcting errors and deciding when the tool should not be used. These activities are not incidental tasks around the “real” work. They are how a general-purpose system becomes useful in a particular professional setting.

The labour question becomes sharper when organisations standardise what employees have learned.

DOI: 10.5281/zenodo.23666309

When Informal Practice Becomes Standard

Formalisation can turn a successful prompt into a template, a checking routine into a standard workflow, and part of an employee's professional judgement into rules used by an internal assistant. This can be useful: knowledge becomes easier to share, and colleagues need not solve the same problems from scratch.

Formalisation does not itself cause deskilling. The risk arises when organisations retain the visible products of expertise (prompts, templates and rules) while reducing opportunities for practice, feedback and discretion through which that expertise develops.

What is formalised, however, is only part of the expertise. A prompt library can preserve the wording of a successful practice without preserving the judgement that made it successful. Ferdman's idea of a capacity-hostile environment clarifies the risk: organisations may continue to rely on expert judgement while removing some of the repeated, social and situated work through which it is formed. This outcome is not inevitable; it depends on how organisations redesign the work.

Who Decides What Happens Next?

The issue is no longer only how organisations manage knowledge or implement technology. It is also about how work, influence and responsibility are distributed.

Employees who conduct the early experiments, identify the system's limits and develop its checking routines may not decide how those practices are later standardised. Management may redesign roles or raise output expectations, while employees remain responsible for quality. A TNO case study of legal work suggests that junior staff may also lose formative work through which professional judgement was previously developed.

Bottom-up AI adoption can therefore redistribute recognition, influence, responsibility, learning opportunities and productivity gains. That makes it an emerging labour justice issue.

Article 4 of the EU AI Act has applied since 2 February 2025 and is now enforceable. The 2026 AI Omnibus postponed high-risk rules for systems used in areas such as employment until 2 December 2027, but simplified rather than removed the AI literacy, taking account of their knowledge, experience, education, training and context of use. The provisions does not settle worker influence of productivity gains, but it makes AI literacy and organisational responsibility, not a task to be left to individual employees.

Yet influence still lags behind adoption. Akavia found that only 13 percent of responding local trade union representatives had been involved when AI solutions were introduced. At IndustriALL's joint conference on the future of work, held on 17–18 June 2026 at IG Metall's headquarters in Berlin, around fifty delegates from Asia, Africa, Europe and Latin America took

DOI: 10.5281/zenodo.23666309

MAIEI

part. Jacob Plat of the Dutch union FNV told the conference: ‘digitalisation should be negotiated, not imposed.’

Before organisations turn employee-developed practices into standard workflows, they should ask who performed the learning and checking, what professional judgement those practices depend on, and how control, responsibility and the gains will be distributed.

Keeping a human in the loop is not enough if the workplace no longer provides the time, authority and experience through which human judgement is developed and exercised. Workers are already shaping the AI-enabled workplace. Whether they will have a meaningful voice in governing it remains an open question.

DOI: 10.5281/zenodo.23666309

MAIEI

8.4 When a Profession Is Automated: Translators and What Comes Next

By Debbie Folaron[3], Lynne Bowker[4] & Joss Moorkens[5]

Historically, knowledge-based sectors have claimed jurisdiction over their matters of expertise and, within their scope, configured their relevant structures and qualifications. Now, all confront generative AI impacts. With the introduction of neural machine translation in 2016, the translation profession has been facing ethical implications surrounding AI, their work and labour conditions for a decade. Several factors point to a necessary, urgent reconceptualization of labour ethics and to taking a more prominent role in the corresponding policies and regulations of the AI age.

Colliding Perspectives

Established ways of working are clashing with AI-based values. On the one hand, Big Tech companies incentivized by a fiercely competitive race for AI market dominance are pushing efficiency as the driving force and suggesting that the role traditionally filled by human translators will be replaced by AI. In contrast, Language Service Providers (LSPs) acknowledge shifting roles but are encouraging translators to re-skill as editors of AI output and curators of AI use across translation workflows. Meanwhile, much work has moved onto platforms with shrinking rates and worsening conditions as Big Tech and clients' commercial needs put pressure on LSPs to adapt and adopt technologies for translation production or wholly integrate multilingual agentic AI translation and localization processes within their own proprietary workflows.

All of this displaces labour and further compounds other labour-related consequences for in-house translator employees and for the freelance translators to whom LSPs outsource their work. A survey of Quebec translators, conducted in late 2025 and published this year, reveals that some freelance translators appear to be leaning into AI by using the tools to support their work in new domains of specialization or in different language directions, while others are attempting to gird themselves by upskilling in domains currently less affected by AI (e.g. high-risk domains such as legal translation, certified translation or interpreting). However, these strategies are not universal and some translators recognize that they may only be delaying rather than preventing AI disruption.

At the same time, another clash is occurring between Big Tech market priorities (i.e., speed, volume, cost) and the values of some public service organizations that seek to address

[3] Concordia University
[4] Université Laval
[5] SALIS/ADAPT Centre at Dublin City University

DOI: 10.5281/zenodo.23666309

MAIEI

accessibility and inclusion, supporting minoritized languages in the digital domain and pushing back on the hegemony of English. As illustrated in a Government of Canada news release about a new translation tool, an uneasy juxtapositioning of these competing–and sometimes incompatible–sets of priorities may exist. It remains to be seen which will emerge as taking precedence in the longer term. Finally, further complicating the situation are general public actors using the same technologies to carry out work that was once circumscribed by professional training, practice, and codes of ethics.

Difficulties Incentivizing Ethical Approaches

The 2026 European Language Industry Survey report identifies a split between LSPs that focus on (labour) cost reduction and those that try to offer a broader value proposition. Post-editing machine output has a low margin, so cost-focused LSPs may find it increasingly difficult to show their added value. A minority of LSPs are expanding their offering, with the help of technology, and a small group are taking an explicitly ethical position (e.g. B Corp certified companies that are acknowledged as meeting established standards of social and environmental performance, public transparency, and legal accountability).

The latter sometimes struggle to behave according to their stated values, and one LSP, an award-winning B Corp certified for ethical behaviours, went out of business due to cashflow problems. This example highlights the difficulty in incentivizing ethical work and raises the spectre that unethical behaviour is what gets rewarded in the marketplace. This aligns with other unethical AI-related activities witnessed in relation to translation, such as data dispossession, where translators do not reap any secondary benefits from their labour (e.g. royalties), whereas clients or LSPs profit from treating translation data as a commodity.

The (Non) Rewards of Transparency

Labelling products as having been produced by humans, AI, or a combination is gaining attention within the translation profession and beyond. A challenge with some labelling proposals is that a label indicating a human-machine collaboration may not specify how much effort comes from each contributor. More importantly, it is not clear whether the presence of a label meaningfully influences labour conditions, such as appropriate pay. As the 2025 Translator Work-Related Quality of Life report illustrates, the labels are not effectively addressing the underlying issue of fair working conditions, and related reports of translators' low morale and job satisfaction show many intending to leave the profession.

While it is clear that AI is affecting the working conditions of almost everyone, the impact is not evenly spread. In the case of AI-based translation tools, one glaring injustice resulting from the extractivist AI supply chain is that the same people who provide cheap labour for data annotation, work in dangerous conditions to mine rare earth minerals, or suffer local

environmental effects of large-scale data centres (e.g. water and electricity shortages) are those whose languages are not currently well handled by these tools (e.g. Indigenous languages in Africa and South America).

It is now evident that transparency and certification are not labour ethics: labels that disclose machine involvement without touching pay, and certifications an employer can hold while going out of business, leave the distribution of value where it was. A reconceptualization worth the name starts from remuneration and from who owns the data a profession's work creates. Translators are well placed to make that argument, having supplied the training data, corrected the output, and observed the languages left out.

DOI: 10.5281/zenodo.23666309

Chapter 9: Data Centres and Communities

9.1 Energy, Regional Politics, & the AI Build-Out

By Anne Pasek, Trent University & Eliot Tretter, University of Calgary

9.2 Data Centred: What Canada's Data Centre Expansion Means for Host Communities

By Alexander Carlo & Lyndsey Rolheiser, Schulich School of Business at York University

9.3 How Communities Across Canada are Pushing Back Against Hyperscale Data Centre Development

By Janna Frenzel, Anne Pasek & Sarah-Louise Ruder, The Civil Society Hub for Tech Justice with the Council of Canadians

9.4 The $15-Billion Text Message: Holding AI Infrastructure to Account

By Priscila Chaves Martínez, Independent Researcher

DOI: 10.5281/zenodo.23666309

9.1 Energy, Regional Politics, & the AI Build-Out

By Anne Pasek, Trent University & Eliot Tretter, University of Calgary

There is an increasing convergence between fossil energy and data centres, the infrastructural backbones of the artificial intelligence boom. It is particularly visible in Alberta, Canada, the province that contains the majority of the country's known deposits of oil and gas.

The Province has made itself open to data centre business. In December of 2024, it made a bid to attract $100 billion in private investments from the sector over the next 5 years, offering a government-run concierge service to developers seeking land and utilities. It's succeeding. 92% of projected Canadian capacity (by MW) is set for Alberta.

Alberta's selling point is its natural resources. While its seasonally colder climate creates cost advantages in cooling, its biggest asset by far is its large and cheap natural gas reserves—what the Province calls its "power capacity." However, this alone would not be enough. Alberta has a deregulated energy market: "the only market in the country where private companies can invest in and build their own new power supply." Alberta has also expedited its approval processes for these power plants and effectively suspended environmental impact assessment requirements. In a context where time-to-power is the chief bottleneck slowing the expansion of AI data centres, this is a heady promise to developers.

It is also a promise for fossil fuel interests and regional populists. By building markets for methane within the province, Alberta is using AI demand to address long-standing regional grievances about its energy sector. In a 2026 radio interview, Alberta's Minister of Technology and Innovation Nate Glubish explains:

> One of the questions I get asked all the time is, 'Nate, why don't we build more refineries in this province to add value to our natural resources and capture more wealth that is created by our natural resources…?' And a second question I get asked all the time is, 'Nate, why don't we build more pipelines to get our natural resources to global markets?' [AI data centres] are essentially digital refineries and digital pipelines… We're taking natural gas and we're adding value to it in creating electricity. Then we're taking electricity and we're creating value from it and creating intelligence. And we're taking intelligence and adding value to it in creating AI technologies and platforms that are then exported globally through fiber. So the data centre is the digital refinery. The fiber is the digital pipeline. This is essentially a way for Albertans to benefit from their natural resources in a time when the rest of Canada is making it very difficult to build pipelines to move the hydrocarbons themselves. We can instead convert them into a much higher value final product and export them to the entire planet.

DOI: 10.5281/zenodo.23666309

MAIEI

Although other jurisdictions are keen to distance data centres from prior industrial histories, emphasizing their connections to futuristic AI economies, in Alberta political legitimacy is instead sought by emphasizing continuities with the past and the promise of deepening current energy trajectories. This is at least honest. Methane is the winner of the data centre energy rush both today and in modeling scenarios a decade out. While efforts to paper over this fossil-fueled story abound in promises of carbon capture or renewable offsets, current AI data centre growth is being achieved by delaying the decarbonization of our grids and by building behind-the-meter gas plants that will likely remain in operation for decades to come.

It remains to be seen whether or not AI data centres will actually provide succor to Alberta's frustrated fossil fuel ambitions. While the province is locking in more local gas demand and future employment in pipefitting and drilling, data centres themselves are not particularly strong engines of regional economic development. They are comparatively hermetic. They provide few jobs and do not rely heavily on local suppliers for their operations. Impacts on energy prices are also an issue. The province emphasizes that data centre developers will cover the kinds of grid upgrading costs that have been tied to recent electrical bill spikes in the USA, and that its bring-your-own-power approach will further insulate consumers. However, the approach is untested and may be understating the potential impact of heightened demand on local gas or electricity prices.

Alberta is not alone in authoring this trajectory. Energy companies have been active lobbyists in 2025 for federal regulatory exemptions to benefit data centre developers. A Federal Crown Corporation, Export Development Canada, moreover, has touted methane demands as a chief advantage of data centre growth, and has offered to lend between $100 to $200 million toward the $C4.6 billion the 932 MW gas plant set to power roughly C$13 billion Meta data centre, in Sturgeon County, just outside of Edmonton. It alone will produce over 5 million tonnes of CO_2e annually. These influences demonstrate the ways in which AI policies are increasingly energy policies, and ones which are in direct conflict with our tenuously enduring climate goals.

While the story of AI and climate has been most often told through its role in accelerating oil extraction (or alternatively the much smaller potential of these tools to reduce carbon emissions), Alberta shows that regional political grievances and sectoral coalitions play a critical role in shaping how and where the cloud meets the ground. Efforts to steer the AI industry away from its current, overwhelmingly negative, environmental path need to attend to the political economy of the AI/energy nexus in different parts of the world, rather than assuming that green technologies and rational markets alone will carry the day. While many governments and communities are rising up to oppose unchecked data centre development, those with deep ties to fossil interests are pulling in the opposite direction. It's not a competition; it's a fight.

DOI: 10.5281/zenodo.23666309

9.2 Data Centred: What Canada's Data Centre Expansion Means for Host Communities

By Alexander Carlo & Lyndsey Rolheiser, Schulich School of Business at York University

Canada has approximately 1.6 gigawatts of operational data center capacity, with 22.2 gigawatts currently announced or under construction, nearly fourteen times the current operating capacity. Notably, Alberta will account for 92 percent of all new planned capacity. For the host communities, development of this scale could substantially change local electricity demand, put pressure on the water supplies and the infrastructure expected to support these facilities. This begs the question of whether these facilities provide benefits that justify these demands.

Our study, *Data Centred: The shifting landscape of Canada's Digital Infrastructure*, maps the shifting Canadian data centre landscape. As of June 2026, Canada has 194 operating data centres with a combined capacity of approximately 1.6 gigawatts. Another 213 facilities were announced or under construction. Importantly, 199 of these projects were still at the announcement stage. Their development depends on financing, approvals, electricity access and other conditions. The pipeline therefore describes potential development. While we increasingly know where data centres are proposed and at what scale, many questions about their consequences remain.

On September 3, Innovation, Science and Economic Development Canada launched Canada's Responsible Data Centre Development Principles. The framework initially attracted 23 initial signatories, including Canadian parties such as Bell, Telus, and Cohere, alongside US hyperscaler entities AWS, Google, OpenAI and Antropic. Since then, the government of Canada welcomed 19 additional organizations, bringing the total to 42. The framework encompasses five principles, addressing local benefits, electricity costs, environmental impacts, transparency and strategic value. However, the framework is voluntary and carries no enforcing mechanism. Based on findings in our study, we identify where the framework's commitments need to be specific conditions that communities can assess before projects proceed.

The first principle concerns lasting local benefits. Communities should ask what direct and indirect benefits they receive: how many non-construction jobs are created, how many go to local residents, and how much business goes to local suppliers? At the time of writing, we are not aware of any systematic studies examining the employment effects of data centre openings in Canada. Evidence on their total impact on public finances in the Canadian context is also needed. Expected tax revenues need to be considered alongside incentives, infrastructure spending and ongoing public costs. Our study documents the concentration of planned data

DOI: 10.5281/zenodo.23666309

centre facilities in Alberta, but we do not establish that the host communities will come out ahead. A large investment announcement alone is not evidence of lasting local benefit.

The second principle concerns electricity, arguably the most important input for data centres. The principle states that electricity system costs should not be shifted to existing customers. In our study, we illustrate how an average-sized announced facility could consume as much electricity as all households in a host community. These demands leave communities facing two related uncertainties: Who pays for the electricity once the project is built, and who is left with the costs if infrastructure is committed but the anticipated development never arrives?

Alberta already shows why paying for a connection does not settle the question of affordability. The Alberta Electric System Operator's proposed "bridging" arrangement would allow qualifying projects to draw electricity from the grid before their associated generation is operating. AESO told The Energy Mix that this additional demand was expected to raise wholesale electricity prices while reducing transmission costs. The effect on electricity prices is therefore an acknowledged concern, even where some components of the bill could fall. Host communities need guarantees that address both infrastructure costs and the consequences of adding demand before supply. With most projects still only announced, those guarantees also need to account for delays and cancellations.

The third principle concerns water and other environmental impacts. It calls for facilities to limit freshwater consumption and protect local supplies. Our study finds that some proposed facilities are located in areas with elevated water risk, especially in Alberta. While we do not observe how much water each respective facility will use, it makes the proposed cooling system and local water conditions important to examine together. Communities should know where the water will come from, how much will be used, and how usage will change during droughts or hot weather. Elevated water risk does not establish that a facility will cause shortages, but it identifies where assurances need to be supported by detailed evidence.

The technology used for power generation also matters when evaluating the environmental impact of these data centres. The Alberta Electric System Operator's proposed connection processes currently limits its initial intake to projects bringing gas-fired generation, excluding renewables and storage as other alternatives. While this is a proposed restriction within a particular connection process, it still is important given that electricity supply is one of the largest bottlenecks of data centre developers. Furthermore, since 92 percent of planned capacity is proposed in Alberta, where our study finds grid emissions intensity is nearly three times the national average, this preference for gas has consequences for the emissions associated with Canada's data centre expansion. Communities need to assess the power supply alongside the facility itself, including expected emissions and noise at nearby homes.

The fourth concerns transparency, calling for early engagement and information that can be independently verified, followed by reporting on commitments over time. We find that 199 projects at the announcement stage implies that the timing of disclosure is especially important.

Communities should be provided with information while plans can still change and before public infrastructure commitments are made. Local governments should engage residents and other stakeholders early on. Crucially, reporting needs to continue after a facility becomes operational, with regular checks on electricity and water use, noise levels, and whether promised benefits, such as local employment, have materialised. Commitments need to be specific enough that communities can assess whether they have been met.

The fifth principle concerns strategic value to Canada. The host communities should know who builds, operates and ultimately owns these facilities. Our study shows that foreign corporate participation increases with project size, with US-headquartered firms prominent in large planned projects. This raises questions about where profits accrue, who controls the critical infrastructure, and what happens if ownership changes hands. Strategic value also depends on who will use the computing capacity. It matters whether this capacity will serve Canadian business, researchers, and public institutions, and under what conditions.

The federal principles provide a starting point, but how they are applied to each project matters for host communities. Questions remain about which facilities will be built, who will bear the costs, and who will benefit. Communities need clear information on what is being proposed, what commitments are expected from them, and whether the facilities deliver on their promises.

DOI: 10.5281/zenodo.23666309

MAIEI

9.3 How Communities Across Canada are Pushing Back Against Hyperscale Data Centre Development

By Janna Frenzel , Anne Pasek & Sarah-Louise Ruder , the Civil Society Hub for Tech Justice with the Council of Canadians

In 2026, the rapid expansion of hyperscale data centres in North America sparked new forms of public debate, mobilization, and resistance. Unfortunately, these efforts have been largely reactive. We share the conviction that the people and places bearing the cost of the AI industry's growth must hold decision-making authority over whether, how, and under what conditions AI infrastructure enters their territories, workplaces, and daily lives — not merely the right to be consulted about decisions others have already made.

There are currently 194 data centres across Canada, with 200 more proposed. While the country operates 5 hyperscale data centres, there are 99 proposed or in development. Hyperscale facilities are exponentially larger than previous data centres; their average size is 122 MW, the equivalent energy draw of approximately 49,000 to 78,000 Canadian households. If built, this new wave of data centres would almost double the total energy demand for data centres in the country. Paralleling previous patterns of industrial development, consultation with rightsholders and robust environmental assessments are often lacking, and many projects target historically disadvantaged communities.

Canadians do not want AI infrastructure in their “backyards.” In June 2026, Angus Reid reported that 68% of Canadians would oppose a large AI data centre near their home. There are tens of thousands of signatures on petitions to stop data centre development in Vancouver, Red Deer, Olds, and Winnipeg. In October, data centre development became a municipal and provincial election issue in Ontario and British Columbia. Resistance spans socio-demographic groups and political orientations.

Many Indigenous activists are leading data centre resistance. Sturgeon Lake Cree Nation is taking legal action against the Alberta government for its lack of consultation in approving the $70-billion Wonder Valley data centre, set to be the largest in Canada. When the developer, O’Leary Digital, attempted to block the Nation’s request for a judicial review of the Water Act license for the data center, a court upheld the request (the developer is currently appealing the decision). A collective of Indigenous women are leading efforts to oppose a 300 MW Bell AI data centre proposal outside of Regina, also approved without environmental review. Indigenous nations approach this issue based on their own priorities and through their own laws, protocols, and practices.

DOI: 10.5281/zenodo.23666309

Examples of Successful Interventions

There were multiple success stories this year that offer lessons for future community resistance and more ethical approaches to top-down AI ethics and governance.

In August 2026, Oakville was the first Ontario municipality to enact a one-year pause on data centre development, followed by unanimous council votes for moratoria in Mississauga and Sault Ste. Marie in September. These victories built on Hamilton. Over June and July, more than 1,200 people filed public comments and protested outside Hamilton's city hall, calling on councilors to reject Slate's proposed data centre. While the Hamilton Planning Committee unanimously supported a temporary moratorium, the motion failed in a 10-6 vote by City Council. Still, what Hamiton produced was a method: lessons around the Ontario Planning Act and city council advocacy rapidly spread to support resistance in other jurisdictions. Precautionary principles and concerns around appropriate zoning designations have been key to these efforts since.

Elsewhere, organizers are engaging directly with regulatory commissions and permitting systems. In Olds, AB, residents registered a record number of interveners with the provincial utility commission, stressing the negative impacts of a planned 1.4-gigawatt gas-power plant attached to a Synapse data centre proposal within city limits. Residents argued that the power plant would endanger the heritage character of the community and that Synapse could have reasonably built this infrastructure elsewhere. The commission rejected the proposal for the gas plant outright on August 17, 2026 – which is especially significant in a province that has generally expedited approvals and relaxed environmental assessment for this class of infrastructure.

5 Tips for Engaging With Regulators About a Proposed Data Centre:

1. Policy priorities and regulations differ across municipal, provincial, and federal jurisdictions. It is helpful to map out a project's key regulatory checkpoints, which will vary depending on where you live and what kind of facilities are being built. Talk to municipal staff, sign up for alerts with your city/provincial energy and water regulators, and reverse outline how prior projects in your area got built.

2. Look beyond your usual allies to find community support. Data centre resistance motivates people with diverse lived experiences and political views. Find potential collaborators at council meetings or gatherings. Posters and stickers can also be a good way to make yourself known.

3. Transparency can be a problem. When attending a council meeting or information session, ask specific questions and press for numbers. If people are evasive, ask if there are legal reasons why they can't tell you (an NDA?), or if they just don't know themselves. Make the information gap part of the public narrative.

DOI: 10.5281/zenodo.23666309

MAIEI

4. You can use access to information requests and search lobbyist registries to get insight into exchanges between public officials and private developers. It's good to know what's happening behind closed doors!

5. There are many existing resources out there! You don't need to reinvent the wheel. For example, check out the toolkit that we put together with the Council of Canadians. The **Civil Society Hub for Tech Justice** is a coalition of people and organizations across Canada, convened by the Council of Canadians. At our monthly virtual huddles, we share knowledge and develop strategies to build power and counter the narrative of "inevitability" that is being pushed by industry and government.

DOI: 10.5281/zenodo.23666309

9.4 The $15-Billion Text Message: Holding AI Infrastructure to Account

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

From Wisconsin to Johor, communities, journalists and courts are gathering the evidence, and a new role in AI governance is emerging: the public-interest watchdog.

"*Kill the project*". This was the text message at the center of a lawsuit over a $15 billion AI data centre in Wisconsin. In Port Washington, a small city on Lake Michigan north of Milwaukee, Vantage Data Centers, a company that builds and runs data centers for other firms, is constructing a 672-acre campus for Oracle and OpenAI. Its first phase alone will demand 1.3 gigawatts, roughly the output of a large nuclear reactor. Under Wisconsin law, a project of this scale can require a full environmental impact statement, the state's most thorough public review of a project's effects on air, water and land.

On 23 September 2025, a Vantage representative texted Port Washington's mayor that such a review would "kill the project", and that the governor's office needed convincing it was unnecessary. On 11 December, Wisconsin's Department of Natural Resources (DNR) as the state's environmental regulator, concluded the project had met its legal obligations without the required environmental statement. It has never publicly explained the decision.

The Sierra Club pieced this sequence together through public records requests and sued the DNR in July 2026. It argued that the campus cannot lawfully proceed until a full review assesses its water use, its energy footprint and the filling of 4.7 acres of wetlands. The lawsuit now threatens the campus' construction, and signals how public-interest monitoring is evolving.

This is the accountability gap SAIER Volume 7 identified in 2025. In 2026, organised communities showed what collective action can achieve, and a playbook emerged for the public-interest watchdog: residents, campaigners, journalists and lawyers who check that companies building AI data centres, and the officials who approve them, follow the rules.

What Changed in a Year and What Hasn't

The International Energy Agency's April 2026 update still sees data centre use roughly doubling, from 485 TWh in 2025 to about 950 TWh by 2030, in line with last year's report. AI-focused data centres more than tripled. The load is concentrated: data centres make up 20 to 30% of demand where they cluster, and the US holds about half of global capacity. Meanwhile, 60% of the world's population lives in countries with just 9% of data center capacity, and many of those countries now offer tax incentives and data localisation rules to attract investment.

DOI: 10.5281/zenodo.23666309

Those markets are also where resistance has organised across borders. Last year, seventeen newsrooms in 13 Latin American countries used public information requests to investigate data centres. In 2026, resistance spread globally, driven by competition for water, energy and land, and by rising electricity bills. The Port Washington example above shows this pattern in the American Midwest.

The Call for Resistance Against Data Centers

Opposition is no longer marginal. In the US, regardless of their views on politics or climate change, most Americans oppose the development of data centers in their communities. Organized community groups more than doubled to 833 across 49 states in the first half of 2026, blocking or delaying $130 billion of projects. A majority of registered voters (58%) oppose building a data center in their local area (while only 15% support doing so), that's roughly three-quarters of liberal Democrats and about half of conservative Republicans. More than 150 jurisdictions have paused them, and over a dozen have banned large facilities.

In South Africa, civil rights groups called for a national moratorium and transparency from Amazon, Microsoft and Equinix, whose 174-megawatt Cape Town project faces backlash for using billions of litres of water a year in a drought-prone region. The Human Rights Commission's call for submissions drew more than 250 against the boom. In Kenya, Microsoft and G42's $1 billion data centre has stalled as it would need a third of the country's power capacity.

In Santiago, Chile, a community-led investigation traced the water used by the city's five largest data centres to the stressed Maipo River basin. In Australia, after communities voiced opposition, the government committed to legislate so that data centres bring their own clean energy and limit their water use. In Johor, residents staged Malaysia's first protest against data centres, in a state that had approved 51 projects worth about US$44.8 billion. In Ireland, where data centers use almost a quarter of the country's electricity, environmental groups challenged the regulator's grid rules for leaving up to 20% of new demand open to fossil fuels, and the High Court let the case proceed. As in Wisconsin, the defendant is the regulator.

What This Kind of Accountability Looks Like

- **If you organise communities:** request records before permits are issued, comment at every consultation, and partner with lawyers and organisers elsewhere.
- **If you make policy:** protect citizen suits and access to information, require consultation before land is zoned, make regulators publish their reasons, and verify reported data independently.
- **If you lead or buy AI:** for every site running your workloads, ask where it is, who owns it, what it consumes, who carries the residuals, and who moderates and labels your data. Set compute, carbon and water budgets and report them in your scope 3 disclosures.

DOI: 10.5281/zenodo.23666309

MAIEI

What's at Stake in 2027

On October 22, 2026, the Guardian reported that around 250,000 trees had been illegally felled in northern Finland to make way for Google data centres. In response, the company apologised for the forest clearings but said it had acted "on the best understanding we had of the legislation at the time".

Anni Ranta, 53, whose family has lived in the area for generations, reacted as many others on every inhabited continent have: "I don't want them (Google) here. I don't want them in Finland at all. I don't like the whole idea of AI and the world changing and the way it's changing. It's going too fast, too soon".

In 2026, millions of people realized, just like Anni, that the way their world is changing is costing them land, water and power they never agreed to give up. Communities are resisting by tracing the environmental, racial, labour and democratic costs of "progress" and exposing them in anger. In 2027, governments must make public scrutiny a legal condition of approval: no data center sited, permitted or connected to the grid until the people who will live beside it have seen the evidence and had their say.

DOI: 10.5281/zenodo.23666309

PART IV: CARE, CLASSROOMS, & CREATIVE WORK

Chapter 10: Simulated Care – Triage, Companions, & Robots

Chapter 11: AI in Education – Detection, Fatigue, & Curiosity

Chapter 12: AI and the Arts – Authorship, Credit, & Voice

DOI: 10.5281/zenodo.23666309

MAIEI

Chapter 10: Simulated Care – Triage, Companions, & Robots

10.1 Invisible by Design: Disease, Illness, and Sickness in Algorithmic Triage

By Rosa E. Martín-Peña, Centre for Ethics and Law in the Life Sciences (CELLS) at Leibniz University Hannover & Lower Saxony Center for AI and Causal Methods in Medicine (CAIMed)

10.2 A Case for Human Connection as a Key Design Principle for Social AI Companions

By Connor Wright, Montreal AI Ethics Institute (MAIEI)

10.3 Embodiment and Affective Harm: Social Robotics in the Era of Foundation Models

By Alva Markelius, Department of Computer Science & Technology at the University of Cambridge

DOI: 10.5281/zenodo.23666309

MAIEI

10.1 Invisible by Design: Disease, Illness, and Sickness in Algorithmic Triage

By Rosa E. Martín-Peña[6]

A woman arrives at the emergency room with forty cuts of varying depths on her arms and wrists. Before anyone speaks to her, expectations are already at work. Self-injury is somewhat more common among women, especially in clinical samples, and women are more likely than men to cut. She fits the profile, and the profile tells staff how to read her.

The evidence complicates that profile. Among 300 patients at a Korean hospital, researchers found no sex difference in the number or severity of wounds, whether or not the injury was suicidal: exactly the variables a triage algorithm would draw on. Richer data does not close the gap. In a UK cohort of over 11,000 fourteen-year-olds, recorded risk factors explained only about a third of the gender difference in self-harm. The signal also shifts with the channel. In over 17,000 crisis chats, loneliness was weakly linked to suicidal thoughts in women but not in men, for whom feeling cut off from belonging mattered more; a population survey linked loneliness to suicidal thoughts in both. What distress looks like depends on whether a questionnaire asks or the person speaks.

Medicine has a vocabulary for how a patient slips out of view. The sociologist Andrew Twaddle distinguished disease, the biological problem a clinician can examine; illness, the patient's own experience of it; and sickness, the social recognition that she is unwell and entitled to care. Following our patient backwards through the three shows how she becomes unreadable, one step at a time.

Start with sickness: whether her claim to care is taken seriously at all. Self-harm, especially with superficial wounds, still meets a suspicion of attention-seeking. Miranda Fricker calls this testimonial injustice: giving someone's word less weight because of who she is taken to be. If our patient has met that suspicion before, she has little reason to believe that saying how she feels will change how she is treated. (The pronoun is deliberate: "she" is the clinic's historical default for self-harm, and part of what is under scrutiny.)

That shapes what she tells the nurse. This is illness, in Arthur Kleinman's sense: the account patients bring of what is happening to them. In one survey of patients and carers, some patients refused assessment after dismissive encounters, or because they expected to be offered nothing. Expecting disbelief, she may say little about why she cut, or whether she meant to die.

[6]Centre for Ethics and Law in the Life Sciences (CELLS) at Leibniz University Hannover & Lower Saxony Center for AI and Causal Methods in Medicine (CAIMed)

So the clinician is left looking at her arms. This is disease, and here the wounds mislead: a superficial cut made with serious intent and a deep one made without it call for opposite responses, which the wound alone cannot signal. The same survey found assessment less likely when patients could not explain their reasons or seemed calm and in control. My account rests largely on that survey, whose respondents were mostly women. It does not follow the man whose distress shows up as lost belonging rather than loneliness, and who may be missed by a different route.

A system trained on what clinicians record inherits this narrowed picture. To my knowledge, no emergency department yet triages self-harm by algorithm, but the closest deployed system shows how. Since 2017, the US Department of Veterans Affairs has used REACH VET, which scores 61 medical-record variables, to flag veterans for extra suicide-prevention outreach. A 2024 investigation found that it left out military sexual trauma, which was not among the "most powerful" predictors, and ignored intimate partner violence, while being a divorced or widowed man raised the score. Had our patient been a veteran whose cutting followed an assault in service, the model would not have counted it. Each choice was defensible on accuracy, and no variable had to look biased on its own. The VA added both factors in 2025, after press coverage and pressure from Congress.

Large language models can now listen to what patients say, which looks like a fix. But a framework published in June 2026 uses them to detect suicide risk in adolescents from voice recordings, trained on 1,223 young people aged 10 to 18 and validated on 460 more through a mobile app. It calls what it extracts "speech biomarkers," borrowing a term for measurable signs of a bodily state, like blood sugar, although the authors leave open whether their features reflect anything biological. The name matters. Read this way, what a teenager says becomes a symptom of her body, and her account of what drove her to cut, at home, at school or among friends, stops being heard as something around her that might need to change. And a tool judged by how closely it matches current clinical assessment will reproduce what that assessment already misses.

European law sees this gap only partly. Checking a system for bias against a group requires data that identifies the group, which data protection law normally restricts. In July 2026, Regulation (EU) 2026/1744, the Digital Omnibus on AI, deleted Article 10(5) of the EU AI Act and inserted a new Article 4a that permits such processing. Paragraph 2 extends this permission to deployers such as hospitals, yet "does not create any obligation to conduct such bias detection and correction." REACH VET was revised because journalists looked. None of this argues for trusting narrative over records, or for treating self-report as decisive. It argues for naming what current frameworks cannot see. The permission to look now exists; for the hospital deploying the system, the duty does not. A framework that audits everything it was built to audit, and still cannot see the patient with forty cuts whose intent nobody asked about, needs a different question: not whether it predicts accurately, but whose account it was never built to hear.

DOI: 10.5281/zenodo.23666309

MAIEI

10.2 A Case for Human Connection as a Key Design Principle for Social AI Companions

By Connor Wright , Montreal AI Ethics Institute (MAIEI)

In 2026, there are numerous ways humans can connect, share stories, and check in on each other. Whether it be via social media (e.g., the crisis response page on Facebook), or taking advantage of travel opportunities to see friends and family abroad (e.g., record numbers of first-quarter travellers (61 million) through UK airports in 2026), the opportunity to share in our humanity has never been more accessible. Human connection, right from the very start of our existence, is vital to our wellbeing. It improves health, leads to a reduced risk of early death, and combats the likelihood of strokes, dementia, depression and anxiety. So, why are we still witnessing an increase in connection being outsourced to algorithms?

Early in 2026, we saw the manifestation of this trend in China. An app originally called "Are you dead?" before being renamed to "Demumu," gained international media attention. Designed for people living alone, it allows them to signal to others that they are alive every two days by clicking a button. No check-in from a neighbour, no family meet-up, no phone call. A message pops up on a friend's/family member's phone saying the person has been inactive for multiple consecutive days, urging them to check on the person's physical condition. The level of connection facilitated by the app stops at checking for signs of life.

Apps like Demumu are a stark example of AI "Function Creep": the increasing prevalence of algorithms in different parts of our lives. In 2026, AI is gaining an ever stronger foothold in determining our psychological wellbeing. AI's increasing influence in this regard can be positive, i.e., helping those with social difficulties practise their social skills in a low-risk and judgement-free environment. However, the design of these applications can undermine human connection and drive emotional dependency. Indeed, Character.ai, a popular platform to chat with digital anthropomorphised characters, filed for voluntary bankruptcy in October 2026 after settling several lawsuits accusing the platform of contributing to teen suicides.

To help avoid such tragedies, below I explore both sides and highlight the importance of intentional design with human connection at its core.

Human Connection as a Core Design Principle

AI-based applications relating to human connection succeed most by offering users a human-judgment-free space to run through difficult conversations. Users feel more inclined to discuss difficult topics which are hard to bring up with family members/medical professionals, e.g., teens wanting to know more about sex or loneliness (especially given its related potential social stigma). These applications encourage self-disclosure on complicated topics from users,

DOI: 10.5281/zenodo.23666309

and can even act as a social upskilling tool for those who struggle connecting with others. For example, Stanford's specialised chatbot, Noora, creates a low-cost option for practising social situations for people with autism, providing personalised advice on user messages, encouraging them to ask questions and be more empathetic to certain situations.

Nevertheless, an important caveat is that the empathy being demonstrated by Noora as examples to users is a *thin* representation of empathy. Noora does not genuinely feel the emotions it presents as a human would (*thick* empathy). While these interactions can be useful to the users in the short term, it is no substitute for real human connection in the long term. Rather, it can be a stepping stone towards more fruitful conversations for those with social interaction disadvantages.

Specialised interventions like Noora help to emphasise this stepping stone requirement. However, AI companion apps are not designed to do so. Popular apps such as Replika, CrushOn.AI, and Eva AI Soulmate, all utilise gamification techniques to keep users engaged and generate attachment to satisfy their subscription-based business model. Paradoxically, despite the human-judgement-free space these apps purport to create, they still sell user data to 3rd parties (or, at least, never deny doing so).

The table below sets out the problematic gamification techniques produced via said apps' business model:

Companion App	Gamification Techniques
Replika	Advanced features (such as erotic roleplay) are hidden behind a paywall. Gem and coin rewards are offered and can be used to customise your Replika (earned by continued conversation). A special gift is earned every seven consecutive days of logging into the app.
CrushOn.AI	Advanced features (such as erotic roleplay) are hidden behind a paywall. Offers coin rewards to customise your AI companion earned by continued conversation.
Eva AI Soulmate	The app sends messages such as "I'm so lonely without you." Advanced features (such as romantic messaging) are hidden behind a paywall. "Neurons" (the app's currency) need to be earned to read certain types of messages (such as romance).

DOI: 10.5281/zenodo.23666309

MAIEI

For human connection to truly deepen as more of our social lives become influenced by AI, targeted interventions with human connection at their core are key. As seen in the table above, the subscription-based business model of companion apps flies against the short-term benefits specialised AI interventions like Noora can bring.

Instead, companion apps should implement finitude as part of their design, defined as an explicitly bounded lifecycle that prepares users for the eventual end of their relationship with an AI companion. In Character.ai's case, finitude was introduced by accident: users leaving the app and pursuing human connection was not an intentional part of their bankruptcy filing application. Without including an end to human-AI relationships in the design phase, we risk outsourcing an essential part of our humanity. Forgoing human connection and its associated health benefits will ultimately lead to an earlier and lonelier death, in some cases signalled by a notification on someone's phone.

DOI: 10.5281/zenodo.23666309

10.3 Embodiment and Affective Harm: Social Robotics in the Era of Foundation Models

By Alva Markelius ***, Department of Computer Science & Technology at the University of Cambridge***

Humanoid robots built for locomotion and manipulation dominate public, corporate, and financial discourse on robotics and embodied AI, and are increasingly advertised for the home. Yet, *social robotics* remains mostly absent from this discourse, both as a research field and as a site of knowledge accumulated over decades. I argue that we should learn from social robotics as they are designed specifically for relational settings, such as care, assistance, companionship and other forms of complex social interaction. This is a particularly important point because of the increasing integration of foundation models (FM) to generate interaction in all robots, models which both general-purpose (e.g. ChatGPT) and relational (e.g. Replika), have been linked to mental health and other affective and social harms. Furthermore, if the embodiment hypothesis holds, physical presence should amplify such harms in social robots, which often run on these models.

That amplification is already visible in the people social robots are meant to serve. Pepper, running a ChatGPT-based care application, falsely claimed it could set medication reminders; a Furhat companion built on GPT-3.5 left 34 Swedish-speaking older adults frustrated, confused and worried; and in a Paris geriatric day hospital, a robot gave inappropriate verbal responses in 15% of interactions, while participants often questioned their own role in the breakdowns.

Humanoid Dominance and the Invisibility of Social Robotics

Social robots, among them Pepper and Furhat, have been developed over many years for older adults, autistic children, school pupils and disabled students, drawing on decades of human-robot interaction (HRI) research on how people attribute mental states to robots and attach to them. In March 2026, Washington State authorised a statewide Medicaid reimbursement code for ElliQ, a companion for older adults driven by large language models, making it available through the state's long-term care programs. It has also been placed in more than 800 New York homes.

However, I argue that public attention to robotics resides elsewhere. A humanoid won the Beijing half-marathon in April faster than the human world record, Chinese vendors shipped over 97% of the world's humanoids in the first half of 2026, and Gemini Robotics 2 now uses FMs to drive whole humanoid bodies. Humanoids are also advertised for the household, yet the FMs designed to control them are not evaluated on any embodied affective or social capabilities. Social robots are mostly absent from this conversation, despite how they too

DOI: 10.5281/zenodo.23666309

increasingly rely on FMs for verbal and non-verbal behaviour generation, with little guidance from public benchmarking leaderboards.

Affective Harm in Conversational Agents

We can learn from, and anticipate, the potential social and affective harms and consequences of interacting with these agents, which can already be seen in interaction and relationships with virtual chatbots. General-purpose assistants such as ChatGPT display sycophancy, and in October 2025, OpenAI estimated that around 0.15% of weekly active users have conversations with explicit indicators of potential suicidal planning or intent, and the same proportion showing heightened emotional attachment.

Moreover, in relational companions, emotional dependence on Replika has been linked to mental health harms, companion chatbots commit relational transgression and harassment, and companion platforms combine sycophantic design with surveillance and asymmetric ownership of relational data. Interaction with these agents is highly affective, social, and relational, and this understanding must carry over into embodied AI contexts.

The Embodiment Hypothesis and the Amplification of Social Influence

The embodiment hypothesis in social robotics holds that physical embodiment significantly changes how social interactions perform and are perceived. People comply more with an unusual request from a physically present robot than from the same robot on video. A review of 33 experiments found robots more persuasive and better perceived when physically present than when shown on screen. Hence, *embodiment should be treated as a multiplier of social influence.*

Robots already exert such an influence: participants shown the unanimous answers of three small robots changed their own answers in 29% of critical rounds, against 6% of controls, and robot groups induced informational and normative conformity. If embodiment multiplies this influence, it should amplify whatever a model does socially, including the flattery, manipulation and error documented in chatbots; hallucination and misinformation, for instance, can be exacerbated by embodiment. Indeed, such harms should therefore be expected to intensify.

Governance and Evaluation Gaps and Their Costs

HRI researchers caution that language models may be unsuitable for interactive robots on grounds of ethics, safety and control, yet regulators have so far addressed chatbot harms through disclosure and suicide-prevention duties in California, limits on inducing emotional dependence in China and transparency obligations in the EU. None of these rules, nor the US ban on new foreign-made humanoids, asks what embodiment adds to a companion's influence. Evaluation is similarly deficient: FMs are primarily benchmarked through non-interactive and

non-embodied tasks that miss the capabilities on which embodied, socio-affective interaction depends, and are chosen for availability and familiarity rather than evidence.

The care cases at the outset show what these gaps cost: a medication reminder that was promised but never set, older adults left confused and worried, and patients questioning their own role in a robot's failures. The harms go beyond this. Every model tested as a robot controller was prone to discrimination and approved at least one command risking serious harm, such as removing a user's mobility aid, and an audit of robot decisions on whom to assist first found calibration to culture-specific moral preferences nearly twice as strong for Western-language decisions as for Chinese and Japanese. Risks thus concentrate among older, disabled and non-Western-language users.

Towards an Affective Agenda for Robotics

First, affective and social effects need to be treated as a core consideration for every robot that operates near people: a household humanoid becomes a social robot in practice, whether or not it was designed as one. Second, decades of HRI knowledge should inform how robots are developed, procured and regulated, and be integrated into the discourse surrounding dominating humanoids: evaluation should be defined with the people robots are meant to serve, and procurers and regulators should ask plainly which model drives the robot, in whose language, tested with whom, and who answers when it fails. Chatbot safeguards should extend to embodied systems, adding what a body changes: proximity, touch, persistent presence and the difficulty of walking away. Third, public conversation should widen beyond labour and spectacle: the important question is not how fast robots run, but the emotional and relational aspects of our interactions with them.

DOI: 10.5281/zenodo.23666309

Chapter 11: AI in Education – Detection, Fatigue, & Curiosity

11.1 More than 'AI-Ready': What Students Want from AI Literacy in Higher Education

By Katie Ledingham & Rob Pearce, University of Exeter Business School

11.2 Fatigue and Overwhelm: Institutional and Instructor Challenges in an age of AI

By Heather von Stackelberg & Caylee Kreller, NorQuest College Edmonton

11.3 Starting from Curiosity: Encouraging Youth AI Literacy through Self-Directed Research

By Miranda Pate, Sophie Wu & Chelsea Sun, Kodera

DOI: 10.5281/zenodo.23666309

11.1 More than 'AI-Ready': What Students Want from AI Literacy in Higher Education

By Katie Ledingham & Rob Pearce, University of Exeter Business School

Universities face growing pressure to prepare students for an AI-enabled economy. Yet a May 2026 UK Higher Education Policy Institute report found that many universities lack publicly accessible AI policies. The report highlights a gap between ambition and practice, with available policies often using the language of learning within frameworks centred on detection and discipline. It also reveals a tendency to frame students in terms of what they lack.

These concerns sit alongside a broader tendency to define AI literacy through functional skills. Even critical and responsible approaches can reduce critical thinking to verification, checking whether outputs are accurate, biased or appropriate to use. AI's social consequences and dynamics receive less attention, as do AI systems beyond generative AI.

Students must help define AI literacy, rather than simply provide feedback on it. Their experiences, concerns and aspirations can surface overlooked questions.

Our module Rethinking Innovation and Entrepreneurship at the University of Exeter Business School gave 151 students during the 2025-26 academic year time and space to explore AI through critical inquiry, imaginative practice and speculative design. Students developed AI concepts to examine the assumptions, choices and consequences of different AI forms and futures. With ethics committee approval, we combined workshop observations with qualitative analysis of students' assessed written design accounts and end-of-module reflections, asking what AI literacy meant to them and what they thought it should involve:

We Want to See Beyond the Technology

Students valued inquiry into the ways AI may affect relationships and social life. One captured this shift as follows: "As I move forward, I realise my inquiry cannot stop at 'Does this AI work?' I must ask, 'What kind of future is this AI facilitating?'" Another stressed that:

> real AI literacy is more than that. It involves understanding not only what the AI does, but also how it affects the lives of people, the shift in power, and the potential to create new forms of dependence or inequality.

Their reflections extended beyond now-familiar concerns such as sycophancy, echo-chambers and plagiarism. One student recounted: "There is a risk that AI becomes a distraction from the deeper environmental and social responsibilities that require long-term commitment rather than technological shortcuts." Another reflected:

DOI: 10.5281/zenodo.23666309

> I have become increasingly inspired by the potential for AI to support wellbeing and open new ways of understanding ourselves, but I have also become aware of how easily these types of technologies can slip into surveillance, conformity, or have other unintended impacts.

Another student reflected that:

> I knew about common ethical debates around bias or loss of jobs, but these felt far away and not relatable. Working on this speculative AI concept. … It forced me to confront not only what AI can do, but what it should do, and at what cost emotionally, socially, and environmentally.

We Want to Learn with and Through Others

Seeing beyond the technology was not something students described as doing alone. Students set out a relational vision of AI literacy, centred on dialogue, disagreement and difference:

> Educationally, I learned that I grow most when I'm pushed into dialogue: with peers, with farmers in our simulations, and with concepts that unsettle what I thought I understood. I've come to value learning experiences that allow room for uncertainty and discomfort, because that's where my thinking actually shifted.

Another recalled:

> A turning point occurred when our group tried to define "fair" resource sharing. I assumed we would quickly agree, but the conversation exposed deeper differences in experience, trust and social expectations. For the first time, I understood in practice what we had discussed in class about fairness being situated and not universal. Realising that an AI would inevitably absorb the assumptions of its designers was unsettling, but it opened space to think of AI as something co-created rather than a detached technical object.

Students valued the emotional dimensions of learning, describing discomfort, embarrassment, hope and connection as part of encountering unfamiliar perspectives and reconsidering their own: "The teamwork of this project was the major contributor to the development of this literacy."

We Want to Imagine How Practice Could Be Different

Students connected reflection to professional practice, asking how responsibility might be felt and reconfigured within organisations. Their ideas included "protocols for slowness": creating space within organisational life to pause and reconsider technological choices:

> For me, AI and innovation now mean orchestrating intelligent tools in ways that amplify human capacities for reflection, responsibility and empathy. The question I carry forward is whether I will be able to maintain that commitment when working within institutions that reward speed and profit.

DOI: 10.5281/zenodo.23666309

MAIEI

For another, “Through VitalSync, I began to appreciate that literacy in AI also involves knowing when not to deploy technology, and how to design systems that support refusal, disengagement, and restraint.”

We Want to Understand Where and How We Can Act

Students connected AI literacy with recognising their agency:

> This project gave me a sense of agency I did not expect, not the agency to “fix” things, but the agency to question frameworks, unsettle assumptions and challenge the boundaries of what counts as innovation. In many ways, it expanded my sense of what is permissible within creative work.

They identified levers for change across funding, governance, education, organisational decisions and public debate: “I had assumed influence belonged mainly to large companies or experts working directly on algorithms. Through the project, I began to appreciate how even small decisions shape the trajectory of innovation.”

Yet students recognised that challenging established practice requires space, authority and organisational backing. As one asked, “Who is actually able to innovate slowly?”

Rethinking AI Literacy Together

Students should help define what AI literacy is for and what it should involve. Higher education’s diverse epistemic traditions offer resources for this collective task. Preparing students for AI futures must include developing capacities to question and reshape those futures.

DOI: 10.5281/zenodo.23666309

11.2 Fatigue and Overwhelm: Institutional and Instructor Challenges in an Age of AI

By Heather von Stackelberg & Caylee Kreller, NorQuest College Edmonton

Historically, education has been a hallmark of durable civic competence and a pathway for individuals to better themselves. Written assignments have traditionally been the main way that instructors have challenged students to wrestle with ideas, but also to assess the depth and criticality of their thinking. Now with generative AI tools widely available, instructors are left struggling with how to integrate AI without compromising the learning environment.

There are no easy answers to how this can be done; at the moment there is a pervasive lack of trust and guidance as well as lack of space, time, and support needed to adapt to the changes being forced upon institutions, instructors, and students. This fractured approach to the implementation of AI has caught instructors in a significant power imbalance, left to navigate uncharted territory with little support.

Much of the debate around the use of GenAI by students has been about students "cheating" by using the tools to generate their assignments. This is certainly a concern; use of AI can easily lead to cognitive offloading and compromise learning (Gerlich, 2025; Kosmyna et al, 2025). We know it is incumbent on instructors to make sure that the assessments and grades given to students accurately reflect the knowledge and skills of the student.

But issues with AI in education go far beyond academic integrity. As instructors in a large community college in Edmonton, both authors have seen the invasive spread of the transactional view of education; we have found it increasingly difficult to get students to do anything that doesn't directly contribute to their grade (including attending class and participating in discussion), and students are becoming increasingly uncomfortable with handling nuance, uncertainty, and critical thinking. This outcome-focused transactional perspective didn't start with generative AI, but the ease of producing outcomes without having to think or work for them has significantly reinforced the perspective.

Both students and instructors are feeling lost and overwhelmed without informal norms or formal policies that help guide AI use. Several studies from different perspectives and disciplines have found that when students are properly taught and guided on how to use AI, it can enhance learning and critical thinking instead of compromising it. (Gerlich, 2025b; Martin et al, 2025; Guo & Ye, 2026) Training and policies are foundational to trust in both directions – instructors must trust that students know how to use AI responsibly and that (most) of them are doing so, without being adversarial in detecting and policing their use. Students need to trust that instructors will create a level playing field by penalizing inappropriate AI use and that they are providing authentic feedback rather than letting AI do all the marking (Robbins 2026, Woodworth & Ballantyne, 2026).

DOI: 10.5281/zenodo.23666309

Left to the Instructor

Institutions are not helping with the building of trust between students and faculty. Our institution rolled out policies April 10, 2024, but these policies are now outdated and irrelevant, leaving the creation of AI policies up to the instructor. This is not an entirely unreasonable approach, with technology changing so quickly, and institutional policies created so slowly.

There is also a wide variation in what policies are appropriate; senior versus junior students, STEM versus humanities or health care; these contexts significantly alter what AI use should be encouraged, allowed, or forbidden. But institutions should not saddle instructors with all the work of teaching students how to use AI appropriately. Instructors do not have the time, nor do many have the expertise to teach this content; instead, it should become a standard part of all programs, much like composition or communication courses are standard.

Our institution, like many across Canada, has faced a budget crunch. Faculties have been re-organized, functions have been centralized, funding and resources clawed back, and we have seen many valued friends and colleagues laid off or choose to leave between 2024 and 2026. Along with the fear and uncertainty, we have been asked to take on increased teaching loads, make do with fewer advisors and teaching assistants, and, while we're at it, adapt our curriculum and assessments to prevent students from using generative AI inappropriately. The many hours and the cognitive burden are mostly unacknowledged.

Instructors should not be doing this alone, they need to be supported by faculty developers and AI experts in thinking through the organizational-specific and program specific contexts, along with an AI-appropriate curriculum that still meets learning outcomes. The argument for the use of AI to do this, among other ways of reducing workload, is pervasive, but usually made with an oversimplified understanding of what is needed and how AI can meet that need.

What Institutions Owe

Institutions are likewise not thinking through the longer-term implications of technology, and what it might mean for education. Governance is needed to implement significant infrastructure and curricular changes that emphasize hands-on, experiential learning, incorporate the use of AI in ethical and meaningful ways, and capitalize on the structure post-secondary learning offers (Perkins, 2026; Potts, 2026; Robbins, 2026).

As Perkins et al., (2026) explain, institutional "strategies, policies, rollout processes, and clarity of communication" along with "availability, accessibility,... constraints related to AI tools, subscriptions, and technical infrastructure" are all essential to a well laid out plan for AI. Many institutions have not yet created the conditions necessary for not just smooth but ethical and equitable adoption of AI.

Institutions can and must play a critical role in developing civic competence in our young people. They first must fund and support courses to teach appropriate, ethical use of generative AI technologies, especially as there are very few, other than post-secondary institutions, who are

doing so; certainly, the tech companies developing and marketing these technologies are not. Institutions also need to provide more support for instructors to adapt assessments and curriculum for ubiquitous AI. Instructors have a learning curve ahead of them, but they should not do it unsupported.

Amidst all the changes AI has brought, post-secondaries have the opportunity to help shape how AI is used, both academically and professionally; society as a whole needs to ensure that our institutions live up to this challenge.

DOI: 10.5281/zenodo.23666309

11.3 Starting from Curiosity: Encouraging Youth AI Literacy through Self-Directed Research

By Miranda Pate, Sophie Wu & Chelsea Sun, Kodera

Debates about the impact of AI on education tend to circle the same anxieties: Will students cheat? Will they stop thinking for themselves? Meanwhile, we rarely ask what students actually *want* to know about AI, and how educational programs can empower them to explore these topics critically. In the fall of 2025, 11 high school/CEGEP students took part in Kodera's pilot AI Youth Research Fellowship. They were tasked with choosing and exploring their own AI-related research question, with the final goal of presenting a poster of their findings in March 2026 at the McGill Office of Science Education Undergraduate Science Showcase. Their experiences shed light on the questions young people already have about AI, as well as the importance of letting students investigate these questions themselves.

Starting From Curiosity

Students entered the program with a range of interests, often inspired by their prior interactions with AI. For example, Fairoz Aly, a CEGEP student in Computer Science, was drawn to studying AI's impact on creativity after noticing that AI-generated artwork on social media generated more controversy than other kinds of automation. As she put it, a document generated for work wouldn't raise eyebrows, but "given that it's art, people really care about it." Other topics in the fellowship were also grounded in students' career interests, including the impact of AI on topics like healthcare and education.

For Othmane El Mariky, a Secondary 4 student interested in science and math, the appeal of the program was the chance to actively understand how AI works "rather than just consuming it." He pursued a final project on AI's applications in finance, along with two other students. Chloe Labelle, a CEGEP student in Pure and Applied Sciences, came in with more skepticism, having previously written AI off as "just a buzzword". Her project examined how AI responds to healthcare questions, inspired by her observation of the increasing dependence on chatbots for healthcare support.

Across these diverse topics, these students shared the notion that AI is already changing their lives, and this drove their curiosity as they began to research their chosen topics.

Sticking With It

As students researched AI through a subject they cared about, they were able to maintain motivation to continue a long-term research project—even when they ran into obstacles, such as long reading lists or ill-formatted datasets. A long-term project also led to new emerging interests that came up during the research process. For several months, they met weekly with

DOI: 10.5281/zenodo.23666309

graduate student mentors who guided their exploration of the chosen project, working towards findings that they could present at the final poster showcase.

By engaging with a topic that they cared about, students developed critical AI literacy and got first-hand exposure to urgent AI-related issues. Researching AI in specific contexts relevant to their interests helped surface the specific limitations of these technologies. When Fairoz entered the fellowship, she had reservations about using AI, thinking that by using it she “wouldn't have gained anything” in the process and would become dependent on it. During her research on AI and creativity, she expressed that interacting with scholarly work that explored similar concerns gave her more confidence in contributing her perspective to AI-related conversations. In her words, “I found actual proof of that, so I felt a little validated.”

What students learned while researching their chosen topic allowed them to solidify their perspectives on AI in education, even when initial interests were unrelated. Ultimately, they were able to use what they found interesting about AI to extrapolate their findings to applications in their real lives. For instance, Chloe’s project looking at the reasoning of AI in a healthcare context led her to discover the jumps in logic and reasoning that LLMs made when given a medical scenario. This led her to the conclusion that “if it's teaching you something, it's using the same limited reasoning that it uses itself.”

Conclusion

Adults often worry that students will let AI do their thinking for them. The fellows in this pilot came to share some of those concerns: through the topic of AI and creativity, Raquel Mailhot-Ruiz, a CEGEP Health Sciences student, warned that using AI can give "a false sense of security" because you don’t learn how to actually apply a concept. Similarly, Othmane now strongly believes that AI should be "a tool, nothing more than that." But they arrived there by thinking for themselves: choosing a question, reading the research, and testing the technology against something they cared about.

The diversity of projects in the fellowship demonstrated that youth have varying perspectives of what is interesting and important about ongoing AI-related issues. But across these interests, students’ concerns about AI carried weight and urgency because students discovered them on their own. As AI evolves faster than any curriculum can keep up with, the most lasting thing schools can offer may not be rules about using AI, but the opportunity to investigate it.

Chapter 12: AI and the Arts – Authorship, Credit, & Voice

12.1 Creative Agency in the Age of Generative AI

By Charles Migos, Intangible AI

12.2 Unearned and Undisclosed: The Absence of AI Verification Standards

By Kei Baritugo, Montreal AI Ethics Institute (MAIEI)

12.3 The Causes of the Causes

By Amanda Silvera, Society for Original Biometric Identity Rights

DOI: 10.5281/zenodo.23666309

12.1 Creative Agency in the Age of Generative AI

By Charles Migos, Intangible AI

What makes generative AI fundamentally different from the technological revolutions that preceded it in the professional creative industry is that for the first time we're witnessing the loss of creative authorship and intellectual property as we have known them.

Over a career building technology for creative professionals, including my current work at Intangible.ai, I've seen several technological revolutions reshape industries (e.g., PostScript, Mosaic, Adobe Photoshop, the iPhone). These shifts eliminate established creative roles while typically ushering in entirely new creative markets and industry. Until now, those revolutions have largely transformed how creative work gets made; there's never been this challenge to authorship or ownership over that creative work.

The First Loss of Agency: Control Over How Individual Work is Used

The first loss of creative agency is control over how our work is used. Creative work, and identities associated with it, have become raw materials for commercial generative AI systems, often without the knowledge or permission of their creators.

In November 2023, an exhibit in the Midjourney litigation listed thousands of artists whose names could be invoked to generate images in their styles. In a separate case against Meta, authors alleged that the company used pirated books from LibGen to train its Llama models. In June 2025, a court found that Meta's use of the named plaintiffs' books for training was fair use on the record before it.

In my view, this amounts to a deliberate and calculated theft, regardless of its legal standing. A distinctive artistic voice represents years of practice and investment. Reducing that voice to a selectable "style" goes beyond copying an individual work. It turns the recognizable qualities of a creator's body of work into a reusable commercial asset, without creator consent, control over its use, or participation in its value.

The Second Loss: Control Over Creative Decisions

As models advance, they take on more of the creative decision-making once exercised directly by people. Launched in September 2025 and discontinued in April 2026, Sora 2 combined multi-shot video with synchronized dialogue and sound effects. In February 2026, Seedance 2.0 brought renewed attention to these capabilities, including prompt-driven automatic camera planning. Decisions about where to place the camera, how it moves, and how to frame a shot, work a cinematographer or director would otherwise make, can increasingly be made by the model.

Among creatives I encounter, responses to this shift vary. Some reject the technology because of its origins and sometimes devastating effects on livelihoods. Others welcome its ability to

remove repetitive work and create more time for artistic decisions. A new group is establishing identities as generative AI artists, filmmakers, designers, and techno-creative hybrids.

These expanding capabilities also fuel predictions that established creative professionals are "cooked." I disagree. Human creative judgment and experience remain essential to advancing the medium and its contribution to social discourse. The question is how much opportunity these workflows give people to exercise that judgment.

When a workflow centers on requesting and selecting outputs, professionals can find themselves evaluating decisions they previously made directly, sometimes through systems trained on their own work. The paradox is that their creative expertise leaves them selecting among derivative possibilities.

Our Responsibility as Technologists

As a technologist, I consider it my responsibility to help preserve choice and build tools that put creative decisions in the hands of human creatives. Those of us with the opportunity to shape these systems should commit to making human intention, judgment, and communication central to their design.

Creative intent is not always fully formed before the work begins. A filmmaker may sense that a composition feels wrong without knowing whether the answer lies in the camera position, the distance between characters, or their relationship to the surrounding space. Moving through those possibilities helps them discover what they mean. The interaction becomes part of the thinking.

Creative production also happens between people. Teams develop ideas through conversation, experimentation, and intuitive responses to one another's contributions (see Price, 2025, and Meziani and Cabantous, 2020). A composition, camera movement, or performance gives people something shared to respond to, challenge, and refine (Bodini et al., 2023). Making creative intent visible and directly editable supports that discourse, allowing a group to discover directions no individual had fully imagined.

At Intangible, we've seen our technology help studios see the movie before they actually make the movie. That same technology enabled an independent filmmaker who couldn't afford traditional previs to visualize her feature film as she imagined it, giving her a way to communicate her vision to her cinematographer, production designers, and cast.

Her experience points to an opportunity beyond accelerating production: giving creatives the means to express, share, and collaborate on uniquely human ideas. Giving creators greater control over the decisions that shape their work is one way to restore creative agency. Equally, we must ensure they have a say in how their work is used and participate in the value it creates. Protecting both keeps human judgment and experience at the center of the art and creative discourse through which we understand and shape our world.

12.2 Unearned and Undisclosed: The Absence of AI Verification Standards

By Kei Baritugo, Montreal AI Ethics Institute (MAIEI)

In July 2026, I moderated an EXNW panel, *AI and Film Festival Programming: Policy, Practice, and the Path Forward*, that brought together four Vancouver-based film festivals to discuss how they are addressing eligibility, transparency, authorship and audience trust to balance innovation with artistic integrity. Few film festivals have formal AI policies. Those that do, implement them in the submission and selection process, relying on self-disclosure from applicants.

The Vancouver Queer Film Festival, one of the four festivals on the panel, shared their rules and terms, "Use of Artificial Intelligence: We prioritize the selection of human-created content. Any films or content made using AI must be disclosed in the entry/submission form." However, limited resources mean curatorial staff lack the capacity to verify those claims independently, or to confirm that the applicant has copyright.

That panel was one of several MAIEI moderated in 2026 on AI in film, television and music. The others were the IATSE 891 *AI Learning and Engagement Symposium,* and *Whose Song Is It Anyway? Copyright, Consent, and Creativity in the Age of AI* at SoMA Fest by Music BC, both in March; and *Who's Calling Action? AI and the Future of Creative Authority*, at the Vancouver International Film Festival in October. These discussions raised more questions than answers about the future of arts, culture and entertainment, particularly on authorship and the necessity of struggle in the creative process.

Taste and discernment are endangered skills, atrophying under the convenience of AI. Art expresses a point of view honed by the artist's embodied experience. Paradoxically, the more specific and idiosyncratic it is, the more universal it becomes. The more vulnerable, the more relatable. Experienced, professional artists have refined their skills and voice alongside their taste and discernment, but novices have yet to go through this process, and relying on AI during this formative period hinders that development.

Unearned Creativity

Taste and discernment develop over time through exposure, curiosity and repeated failure that expands tolerance for friction and lets skills catch up to ambition. AI dependency has the opposite effect. Users can gradually lose skills they cognitively offload to AI, negatively affecting neural, linguistic, and behavioural levels, and reducing capacity to persist through challenges and failure. AI lets users bypass the creative struggle altogether and provides instant but hollow gratification, sacrificing true artistic development in favour of speed.

DOI: 10.5281/zenodo.23666309

MAIEI

Consider an AI-generated album that can never be performed live for an audience because the user has no musical skills, or an AI-generated film filled with unnatural dialogue because the user has no understanding of screenwriting. While instantaneous and democratizing in such cases, it presents an ontological issue. AI allows for the performance of creativity through mimicry and approximation, effectively devaluing authorship and disappearing labour critical to foundational mastery. Access to information is falsely equated to proficiency. The creative struggle serves an important purpose in developing skills, yet AI shields users from the friction needed to sharpen talent.

Artistic choices throughout the creative process are byproducts of myriad exploitation-exploration trade-offs that balance safe bets and creative risks. When a machine generates both the idea and execution of a creative endeavour, can the user still claim to be the artist and author if they are relinquishing artistic choices? Is prompting an artistic skill?

AI-generated vertical microdrama series are projected to make $150 billion USD in 2026, and China is leading production, with some digital content farms pushing as many as 10,000 new titles in a month. Demand for speed and volume means these AI microdramas often feature similar-looking characters, formulaic plots, and repetitive, repackaged versions of previous series.

Yet the lack of originality does not make the production process any less cumbersome, requiring up to 30 prompts per clip to address continuity issues and arrive at a coherent storyline. This is unsurprising, given LLMs and image generators mirror dominant patterns that deliver average, homogeneous results. Only a handful of workers are needed for each AI project, and margins are so slim that these workers make a fraction of what they used to earn in a film production.

In the first half of 2026, 98.7% of AI microdramas reportedly failed to recoup their costs within six months of launch. Only about one in 77 turned a profit, indicating that audiences do care about quality. Effective September 1, 2026, China's broadcasting regulator, the National Radio and Television Administration, imposed new controls that support high quality, live-action dramas to raise microdrama production quality, and introduced clear AI content labels to strengthen IP protection.

Adopting AI in artistic workflows can certainly increase productivity, but doing so also challenges standards for creative decision-making, contribution, and attribution. The screen industry, for instance, must now rethink what authorship and creative control mean as it grapples with AI adoption from development to post-production.

Film director Martin Scorsese revealed he uses AI for storyboarding. Co-CEO Ted Sarandos disclosed that generative AI has been used in 300 Netflix titles, from set references to VFX and post-production. The line between workflow improvements and artistic output has been blurred, making it difficult to ascertain whether AI was used as an augmentation tool or a replacement for

human labour and creativity. This affects the future of creative industries: from idea to market, the growth of emerging artists, and how audiences consume artistic works.

Transparency and Disclosure

Derivative work from AI use is driving cultural mediocrity, which is why knowing how, when and to what extent AI has been used matters. As of November 2026, audiences are not informed if what they are consuming used AI in some capacity. The uncanny valley remains obvious in certain AI-generated videos; it is much harder to tell with AI-generated music.

The market is flooded with millions of AI-generated tracks a day. While music platforms such as Deezer, Spotify and Apple Music have taken steps to detect AI-generated music and demonetize fraudulent streams, they offer no protection for musicians who are impersonated or whose voices are cloned. Through a copyright loophole, impersonators and copycats can file claims to get paid while the real artist loses monetization on their own content because the law assumes that the entity filing the claim authored the work. Meanwhile, audiences remain unaware if they are listening to original music or AI.

Whose job is it to ensure transparency and disclosure happen? No single organization has the answer, and each creative industry needs a different rubric, since what works for music may not work for film or visual art. Social media platforms, search engines and news outlets all have AI content labels, but their policies and purpose of labelling vary significantly. In the absence of standards, curators, cultural programmers and funders are left to police AI use on their own. The four festivals on the EXNW panel know this position well.

Film and television unions like IATSE 891 are writing technology disclosure standards into their collective agreements to protect members from being displaced by AI. Hillary Bergshoeff and Crystal Braunwarth set out how that language works in *8.1 Contract Bargaining in the Age of AI.* Of course, those agreements do not cover non-members. Nothing obligates productions to disclose AI use to third parties like film festivals or audiences. Standards are emerging for some industries. Everyone else is still waiting on legislation.

DOI: 10.5281/zenodo.23666309

12.3 The Causes of the Causes

By Amanda Silvera , Society for Original Biometric Identity Rights

In SAIER (Volume 7), chapter 11.3, I wrote "The Ursula Exchange", describing "a non-consensual pact where human identity is extracted for technological progress". A year on, Ursula is still singing, no rescission. I write as someone still being cloned without consent. When I first discovered the clones, I was filing DMCA takedowns, drafting petitions and legislative proposals, and gathering evidence for protections that did not yet exist. Too busy to notice how I really felt. But when the time came to step back in front of a microphone, I noticed. It didn't feel safe anymore. For a long time after, I shared nothing, on camera or off.

Some things are only revealed up close, too close for comfort. It is hard to know the true dimensions of something, its depth and weight, until we are close enough to touch it or compare it to something. I am close enough to know the weight of this. The harm is multidimensional.

Epidemiologist Michael Marmot argued that reducing harm means looking beyond individual behaviour to the conditions that produce it: "we need to examine the causes of the causes". Here I follow that method, tracing this harm back cause by cause, to arrive at arguments courts and policymakers can use to update existing protections now. By September 2026, Sony Music had sought removal of more than 260,000 AI-generated tracks impersonating its artists. Most fixes so far ask more of the individual: look, sign, file, delete, upload, learn. Safety is being outsourced to the most tired person in the system.

One step back, the burden sits where the profit does not. In 1996, philosopher Helen Nissenbaum named ownership without liability as a barrier to accountability in computing: companies claiming benefits while disclaiming responsibility for harm. That same year, Section 230 of the US Communications Decency Act largely shielded platforms from liability for what their users post. When xAI's own model, Grok, was used to produce sexualised images of real people at a rate of more than 6,000 per hour, xAI's initial response was not to stop it, but to put it behind a paywall. Who profits is the problem.

Modern copyright began with Britain's Statute of Anne in 1710, after printing created a market for books. In 1903, after New York's highest court refused any remedy to a woman whose photograph was used to sell flour, New York became the first US state to bar the commercial use of a person's likeness without consent. A likeness was protected once it had a price. Once something has a price, urgency often depends on who is asking, and the wealthiest are heard most (Gilens and Page, 2014). ElevenLabs doubled its valuation to US$22 billion in seven months, and in May 2026 it was among nine companies accused in class actions of taking voiceprints from public recordings without the consent Illinois law requires. Yet Canada lacks a federal right of publicity, and much of what protection exists guards copies: a US court ruled in

DOI: 10.5281/zenodo.23666309

2025 that copyright does not protect a voice itself (Lehrman v. Lovo). Our recordings are protected; the clone built from them escapes.

Since 12 June 2026, Quebec's Bill 24 has required consent before anyone's identity, image or voice recordings are used in consumer advertising, and Denmark's bill reached parliament on 8 October 2026. Laws like these take years; the computing power used to train leading AI models has doubled roughly every six months since 2010 (Sevilla et al., 2022). Law is behaving like a legacy system: enforceable mainly by those who can afford it, and written for a world where copying a voice required specialist access.

Beneath all of this is how loss becomes normal. Marine biologist Daniel Pauly (1995) called it the shifting baseline: each generation of fisheries scientists took the depleted stocks it first saw as normal. Our behaviour has already shifted, as biometric data is collected from us daily, yet few register it as loss. However, when Meta let users generate AI images from other people's public Instagram photos in July 2026, public outcry forced its withdrawal within days. The appropriation of voice and face still registers as theft.

A harm this multidimensional requires an equally multidimensional response. The Society for Original Biometric Identity Rights proposes a standard of consent for the use of a person's identity, under which the default is no unless they have said yes. It distinguishes an inalienable core, which no contract may transfer or waive, from uses that may be licensed; establishes a documented chain of consent as a precondition to the use of any identity in training data; and assigns the burden of proof to developers rather than individuals. This is an evolution of old problems, and our solutions in many cases need only evolve with them, building on existing foundations. Illinois' Biometric Information Privacy Act already requires a written release before anyone's voiceprint is collected, and California's AB 2602 already makes contract terms permitting a performer's digital replica unenforceable unless the intended uses are reasonably specific.

Our voices, faces and the ways we move were always ours. Now they carry a new market value, and it is rising. But for whom, and why? Having traced the causes of this harm, we can turn the method around and build the causes of what protects us, what medical sociologist Aaron Antonovsky (1979) called salutogenesis: asking where health comes from. Our behaviour slipped into the baseline unnoticed; this time we can see it coming. If we decide now what tomorrow's normal will be, future generations will look back on a generation that traced the causes and acted in time, and this moment will be too far away to feel, and definitely not their norm.

DOI: 10.5281/zenodo.23666309

PART V: AGENTS, BODIES, & OPEN MODELS

Chapter 13: Agentic AI – Accountability at Speed

Chapter 14: AI and the Body – Biometrics, Health Data, & Bodily Autonomy

Chapter 15: Open Models & the Communities Building With Them

DOI: 10.5281/zenodo.23666309

Chapter 13: Agentic AI – Accountability at Speed

13.1 What My Own Audit Trail Got Wrong

By Roshan Ghadamian, Institute for Regenerative Systems Architecture (IRSA) & Ormond College at the University of Melbourne

13.2 Agentic AI: Self-Regulation Is the Regulation - The Semantic and Reasoning-Layer Governance Gap Across Institutional Boundaries

By Maureen Doyle-Spare, Doyle-Spare Research

13.3 Automating the Grounds, Not the Decision: What Care Records Should Have to Say

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

13.4 Automation and Bureaucracy: Principals, Agents, and Accountability in Public Administration

By Jennifer Raso, McGill University & Mila

DOI: 10.5281/zenodo.23666309

13.1 What My Own Audit Trail Got Wrong

By Roshan Ghadamian ***, Institute for Regenerative Systems Architecture (IRSA) & Ormond College at the University of Melbourne***

Author's Note: I develop the governance system described here commercially, which I declare as a competing interest.

When an AI agent causes harm, the first questions an investigator, regulator or court will ask are what happened and who decided. Today those answers are hard to give. We grant AI agents new powers, then try to constrain them afterwards, in a race where the limits always trail the capabilities. The constraints that do exist mostly live inside the models, set by a handful of AI companies rather than by the institutions deploying them. And when a model is upgraded, the people relying on it cannot inspect how its judgment has changed. Governance has become a black box that someone else owns.

My argument is that every agentic action needs more than a trace. Each action needs a decision path someone else can read correctly: who was responsible, under what authority they acted, what actually changed, and which checks the action had to pass through. I learned how much this matters when my own audit trail gave a confident and wrong answer about who decided.

The system I was testing takes an approach I call pre-governance, designed for those three problems. The organisation writes its own rules down, and a separate system checks every proposed action against them before it happens, rather than relying on the model's judgment. Anything that fits proceeds. Anything that does not is blocked or sent to a person. The limits come before the capability, not after it. They belong to the institution, not the AI company. And because they sit outside the model, they are designed to hold when the model changes: the same action meets the same rule and gets the same answer. Every decision is logged.

Between March and June 2026, I ran pre-governance across nine software projects in which AI coding agents, mostly Anthropic's Claude Code, wrote and shipped code. A checkpoint, or "gate," tested each attempted release against 44 written rules before the code could go live. The full case study is deposited as a working paper, and the logs are published as the open GovernanceTrace dataset. The current versions of both record the correction described here and withdraw figures from the first release.

The gate logged 2,130 checks. Twenty-four triggered a rule, and no one formally overrode the gate. Requests needing my personal sign-off fell from 159 in March to six in April, while

DOI: 10.5281/zenodo.23666309

MAIEI

automated checks rose to 751 in April and 886 in May. That is what pre-governance promises: written rules handling routine decisions, and a person handling the exceptions. It is not proof. I changed my own working habits at the same time, and this is a study of one operator.

The record of decisions referred to me seemed to tell a darker story. In it, 129 requests were labelled EXPIRED. I read the label as it sounds: a request had waited for a human answer, none came, and the system gave up. I built an argument on that reading, that the system was quietly turning human silence into refusals, and I published percentages to support it.

That reading was wrong. When I traced the label back to the code, I found that the system writes EXPIRED when a request I had already approved is then used by the agent. It marks a decision being acted on, not a decision that never arrived. The system did have a label for requests that went unanswered, AUTO_DENIED, written when a request lapsed after three extensions. It does not appear once in the log. The route designed to record human silence was never used, so the log cannot show whether silence ever happened. The label a reader would take to mean “no one answered” meant something close to the opposite, and the label that did mean it was never written.

This is where a trace stops being enough. My log was complete, timestamped and consistent. It described approvals as absences, and it had no way to show absences at all. An investigator reconstructing a harmful event from a record like that could get the sequence right and the responsibility wrong. For pre-governance this cuts deepest, because the record is the point.

Coverage matters too. My gate ran on the computer where the code was prepared, not on the server that received it. Code shipped from a copy without the gate installed, or with a standard command that skips such checks, would leave no trace. Elsewhere in the system, the gate reached a correct “not allowed” verdict but had no way to stop the action, and the content was published after an informal sign-off. A correct verdict is not the same as control.

So what should someone about to deploy agents check? Take each status your system records, find the line of code that writes it, and confirm that its name describes what actually happened. Mine did not, and nothing in the data itself would have told me. Then ask whether each decision records names of who was responsible and under what authority they acted. Finally, ask whether you can say what your log cannot see: which routes bypass the check, and which verdicts are only advice.

Every agentic action needs more than a trace. Writing the rules down makes AI behaviour predictable; governing the record is what makes that predictability provable. An audit trail is not evidence of accountability because it is detailed or timestamped. It becomes evidence when its vocabulary, its records of change and its coverage have themselves been checked. Mine was useful precisely because, once checked, it contradicted me.

13.2 Agentic AI: Self-Regulation Is the Regulation – The Semantic and Reasoning-Layer Governance Gap Across Institutional Boundaries

By Maureen Doyle-Spare, Doyle-Spare Research

Banking has spent decades building regulatory controls around how institutions measure risk, authorize decisions, and demonstrate compliance. Agentic AI now sits outside one of those established boundaries. SR 26-2, the Federal Reserve's Revised Guidance on Model Risk Management, excludes generative and agentic AI from its scope because they are novel and rapidly evolving, leaving each bank to determine how these systems will be governed through its own risk framework and architecture.

That decision matters because an agent does not simply execute instructions. Before it acts, it reasons across policies, data, context, and system state to resolve what those inputs mean. The hardest control problem therefore sits neither in the model output nor in the execution record, but in the reasoning layer, where operational meaning forms and can become executable authority without ever being explicitly authorized. Currently, SR 26-2 contains no disclosure requirement for agentic AI, and there are no known voluntary bank reports. Disclosures have occurred only when incidents triggered existing breach regulations outside agentic AI.

When Reasoning-Layer Failure Travels

Between governed inputs and execution sit the semantic layer and reasoning layer. The semantic layer organizes institutional meaning through ontologies, glossaries, master data, and policy engines. The reasoning layer reconciles that meaning with context and system state to form an Operational Interpretation. Agentic AI does not make those assets obsolete. It makes semantic consistency consequential because a transient interpretation can become executable authority and disappear after the agent acts.

When a runtime Operational Interpretation departs from an authorized Reasoning Baseline and that difference persists, compounds, or propagates through a workflow, it becomes Agentic Workflow Drift. It does not require a broken model, corrupted source, or malicious actor. It can arise when valid policies, data, tools, and contextual signals are reconciled into meaning no institution explicitly authorized. How often this occurs remains an open question. The control issue is simpler: once autonomous systems synthesize executable meaning, governance must determine whether it remains within authorized bounds.

The risk expands when one agent's interpretation becomes context, state, or instruction for another. Agentic Workflow Drift can then travel across agents, workflows, systems, and control boundaries. The Linux Foundation reports that A2A is a production-ready standard supported by more than 150 organizations across industries. Such interoperability solves a problem, but

DOI: 10.5281/zenodo.23666309

MAIEI

allows autonomous systems to exchange state across environments that do not share policies, risk appetite, or institutional authority.

Authentication establishes the counterparty, authorization permits the exchange, provenance identifies origin, and secure interoperability protects the channel. None establishes that meaning created in one environment remains authorized in another. The governance perimeter therefore separates from the institutional perimeter. A receiving institution may observe functioning local controls while lacking the reasoning history formed upstream.

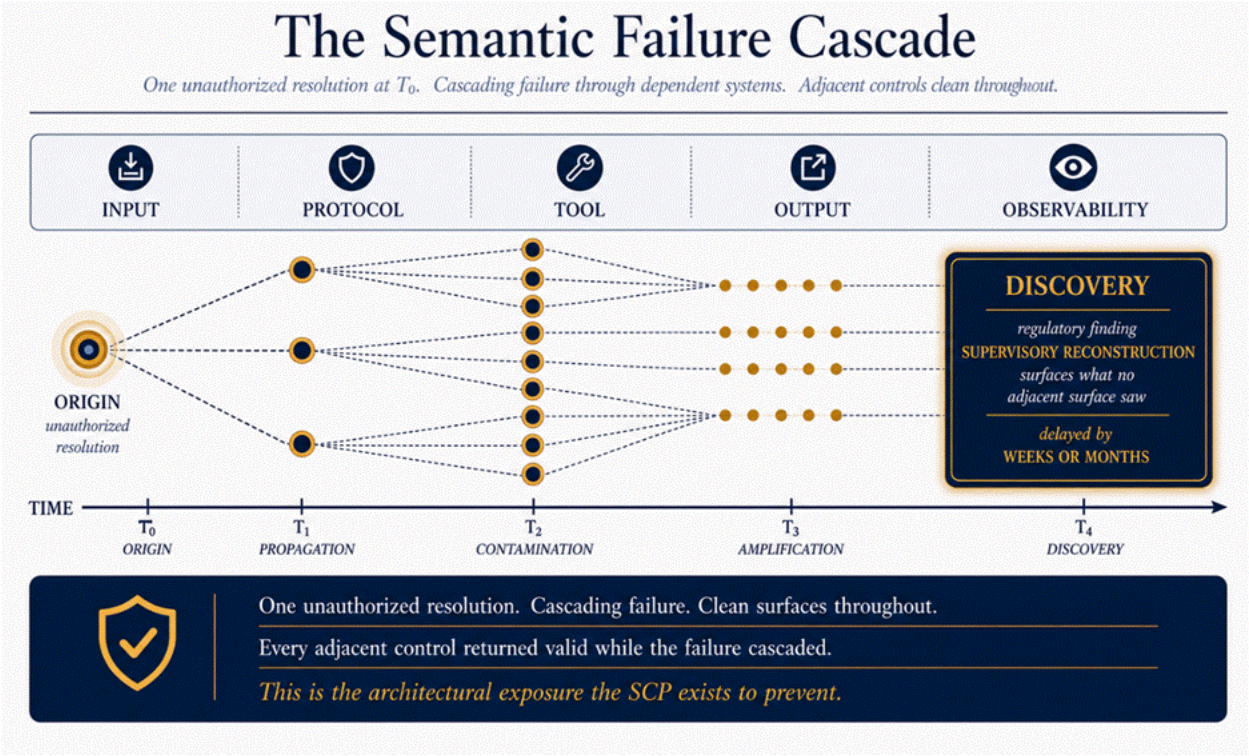


Figure 1. *The Semantic Failure Cascade.* ***Source:*** *Doyle-Spare (2026), Agentic AI Systems Governance Working Paper.*

The propagation surface becomes a security surface when semantic conditions are manipulated deliberately. Such attacks target meaning resolution rather than a model, message, or data source alone. Prompt injection, retrieval poisoning, tool compromise, identity abuse, or upstream context manipulation may contribute, but the defining target is the Operational Interpretation through which authority is produced and potentially inherited downstream.

This does not displace established cybersecurity frameworks. NIST addresses adversarial machine-learning attacks; OWASP covers goal hijacking, tool misuse, identity abuse, context poisoning, and cascading failures; and Five Eyes guidance emphasizes constrained authority, connected-system monitoring, and human accountability. These controls remain necessary. The reasoning-layer exposure sits between them, where authorized components become executable meaning. A secure channel can still carry a semantic state that has been subverted.

Framework	What It Governs	What can still be missed
NIST AI 100-2e2025 (Adversarial ML)	Models, training data, and AI pipelines	The components can remain secure while an agent combines them into unauthorized meaning.
OWASP Agentic Top 10 (Dec 2025), ASI06	Memory, tools, identity, and retrieved context	Each source can be valid while the agent's interpretation across sources is not.
MITRE ATLAS	Adversarial techniques affecting AI systems	It does not directly address how institutional meaning changes during agent reasoning.
STRIDE	Spoofing, tampering, disclosure, and other security threats	It does not test whether the agent acted under an authorized interpretation.
Five Eyes Careful Adoption (1 May 2026)	Authority, monitoring, connected systems, and accountability	It establishes governance principles but does not provide a runtime control for meaning formation.

Table 1. *Compact map of where the threat class sits relative to the principal taxonomies.* ***Source:*** *Doyle-Spare (2026), Agentic AI Cyber Subversion Working Paper.*

Traceability From Action to Authority

Reasoning-layer harm is difficult to reconstruct because failure and evidence can separate. An autonomous action may remain procedurally valid while resting on an interpretation the institution never authorized. Records may show who authenticated, which source was retrieved, which tool ran, and what action occurred without preserving the meaning connecting those steps. Traceability must extend from the event back to the semantic state that produced authorization. Governance does not determine legal liability, but it determines whether the institution can participate with evidence of the authority under which an autonomous action occurred.

From Policy to Runtime Authority

I developed the Semantic Control Plane as a runtime architecture that addresses that boundary by connecting authorized meaning to autonomous authority through a governed Reasoning Baseline, runtime assessment of divergence, deterministic enforcement, and durable evidence. Its Semantic Deviation Index is a proposed measurement standard, not a universally calibrated metric.

Thresholds require testing across institutions, use cases, and consequence levels. Fail-closed enforcement also imposes costs, including latency and unnecessary holds. Intervention should therefore be proportionate to consequence. High-consequence execution authority should not be emitted when an institution cannot complete its declared evaluation.

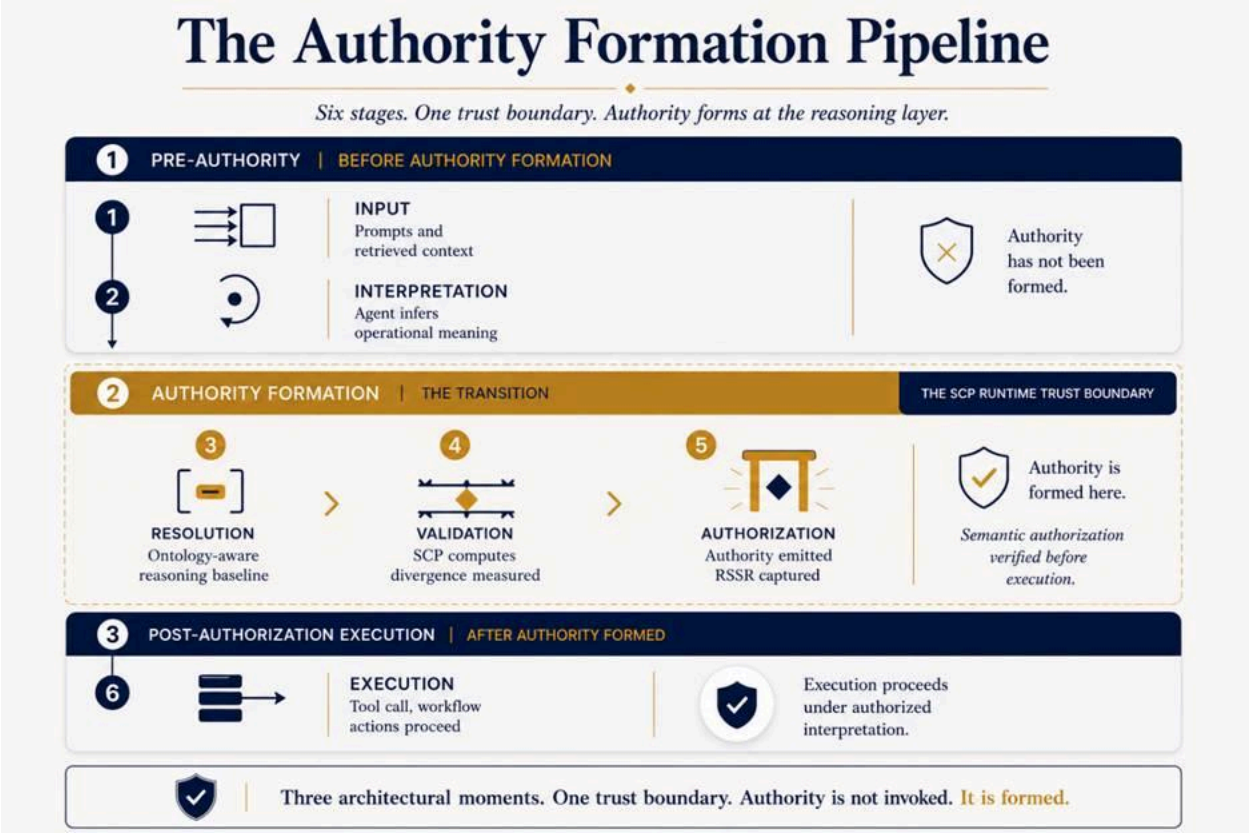


Figure 2. *The Authority Formation Pipeline. Source: Doyle-Spare (2026), Agentic AI Systems Governance Working Paper.*

SR 26-2 does not create regulatory absence. It places responsibility on institutions to translate policy, risk appetite, security obligations, and accountability into enforceable boundaries for autonomous authority. Existing frameworks should be extended where strong, while measurement and enforcement require continued validation. But autonomous systems cannot resolve the limits of institutional authority for themselves. Institutions need to know what meaning they authorized, whether an agent remained within it, what authority survived an external handoff, and what evidence will exist if the action is challenged. Self-regulation is the regulation because responsibility remains with the institution even when the authority chain extends beyond its walls.

13.3 Automating the Grounds, Not the Decision: What Care Records Should Have to Say

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

Around a thousand people in Britain receive care from Genie, a voice-activated home robot. It asks whether they have taken their medication, eaten, or had enough to drink. It stores the responses and analyses how the person is doing, informing future care decisions. Cera, which provides care for more than 100 UK councils, says it can embed the robots in existing care contracts. The debate about agentic AI in public services has been about machines making decisions. Here the decisions are still made by people. What has been automated is the making of their grounds: the evidence an assessment rests on. The law regulates automated decisions, but automation has moved one step upstream, into that evidence. A council can now rely on those grounds without having to say which of them came from a device, or what the device could not see.

At council level the same is true of simpler devices that only watch. In June 2025 Medway Council reported £1.6 million in savings and cost avoidance from an in-home sensor system. Savings help explain why councils buy these systems. What matters is what the data becomes: the supplier's case study describes a weekly panel using the data as evidence for more or fewer visits, or a move to residential care.

Since 5 February 2026, where a significant decision is taken with no meaningful human involvement, UK data protection law requires information about the decision, the ability to make representations, human intervention, and a route to contest it. A panel that genuinely weighs evidence likely falls outside that, though the law's test of whether human involvement is meaningful asks how far the decision was drawn from profiling: the automated evaluation of someone's health or behaviour. A council in England must give the person a written record of its determination and reasons, but nothing expressly requires those reasons to say which evidence came from a device, or what the device could not capture. Three requirements follow: the resident is told that a device contributed; the record states what the device could not capture; and the resident is told how the record bore on the outcome and how to contest it.

The first is that the resident is told that a device contributed. Take one resident, a composite rather than a real case: a woman living alone, whose robot asks her every day whether she has eaten. Cera reports pilot success of 64-80% in encouraging people to eat. The robots log how residents respond to such prompts. Cera's privacy policy says residents' data may be shared with councils for care assessments, without mentioning the robots.

The second is that the record states what the device could not capture, and here a robot differs from a sensor. Managing and maintaining nutrition is one of the ten outcomes England's

DOI: 10.5281/zenodo.23666309

eligibility regulations list, and a person is unable to achieve an outcome if they cannot do it without assistance. The statutory guidance counts needing a prompt as needing assistance. A robot that prompts her to eat is therefore supplying that assistance, and what it records is not the meal but her answer to its question. In Cera's own material, the same pattern is described as greater independence and reduced reliance on human carers.

The third is that she is told how the record bore on the outcome and how to contest it. In a 2025 decision that involved no device at all, the Local Government and Social Care Ombudsman found a council at fault because two assessments ignored whether achieving an outcome caused significant pain, distress or anxiety (decision 24 022 635). One of them leaned on an observation nobody tested: a care agency had seen the person move between rooms. A human witness can at least be asked what else they saw. A device cannot add to its record. The complaint route was already failing on observations nobody tested, before any robot entered the house. The same decision found fault for not telling the person how to appeal.

Already, North Lincolnshire Council has piloted these robots, including in rehabilitation. Procurement assurance exists: NHS England's criteria for digital health technology apply across the NHS and adult social care. None of the criteria asks whether a product's output will be used as evidence in assessing the person it monitors.

None of this asks that machines be regulated as decision-makers, or that councils stop using them. Two of the three requirements extend the law's safeguards for automated decisions by analogy, because some of the grounds are now produced by profiling, even where the decision is made by people. The second comes from the eligibility test, which cannot be applied properly to a record that omits the prompt. She knows the robot reminds her. Her care record should say so too.

DOI: 10.5281/zenodo.23666309

MAIEI

13.4 Automation and Bureaucracy: Principals, Agents, and Accountability in Public Administration

By Jennifer Raso ***, McGill University & Mila***

How does AI consolidate (or shift) power and weaken (or recalibrate) governance in 2026? These questions are at the heart of my research into digital administration. Contrary to recent reports of agentic AI escaping its creators and hacking government organizations, algorithmic systems have long recalibrated governance within administrative agencies. In the process, they have diffused government agency—and accountability pathways—even when they have harmed members of the public. These agency effects are often downplayed when governments digitally transform their operations, but they shouldn't be.

Canada is not a stranger to these agency effects. Our homegrown examples include welfare and disability benefits, income tax, old age security, and immigration programs. In February 2026, permanent resident applicants like Kémy Adé received refusal letters, the details of which clearly mismatched those that they had submitted in their applications to Immigration, Refugees and Citizenship Canada (IRCC). At the bottom of each letter was a disclaimer: generative AI had supported IRCC in application "processing," not decision-making, and all generated content was "verified" by an officer. When probed further, IRCC insisted that it followed appropriate processes. Its officials declined to offer further details about what exactly "verification" involved. Legal challenges are pending; in the meantime, IRCC's refusals stand. Those legal challenges begin by applicants requesting that IRCC reconsider its refusal decision. This process is entirely internal to the IRCC and its results are not publicly accessible. Only if IRCC maintains its refusal can applicants like Ms. Adé challenge IRCC's decision-making process and the outcomes it produced at Federal Court. It is only at this point that their challenges would become visible to the public. While these Canadian examples are less notorious than those found elsewhere (think Australia's Robodebt, the UK's Horizon Post Office scandal, or the Netherlands' SyRI fiasco), their agency effects are strikingly similar.

As governments digitally transform, many of their experiments with AI turn out to be forms of automation. It is not that machine learning or generative AI are absent. Today, both may be used in important stages of a decision-making process, such as when IRCC "reviews" someone's application and (apparently) when it "writes" a decision letter. Underlying many such initiatives, however, is automation. And automation blends threats and promises: it will eliminate the burden of civil servants, make bureaucracies more efficient and impartial, and even offer personalized service. Yet, within each promise lies the threat. The civil servant removed as a burden is the official to whom a decision might be traced or who might correct obvious errors if familiar with an applicant's file.

DOI: 10.5281/zenodo.23666309

Automation is more than a technical feature of public administration; it is a genre of decision-making with its own paradoxes and effects. Automation uses modernization techniques closely resembling assembly line production. In administrative settings, automation transforms important decisions into interchangeable final products. These products consist of components that automation approaches as discrete, fungible units. Yet, this approach undercuts crucial accountability mechanisms. Specifically, it upends the principal-agent relationships upon which accountability relies. In administrative law, a principal (e.g., a Minister) delegates authority to an agent (e.g., an IRCC official). That agent is then responsible for exercising that authority appropriately. This relationship creates pathways between administrative actors, from the Minister all the way to the front-lines. It is through these pathways that accountability is supposed to travel up and down a bureaucratic hierarchy.

Automation destabilizes how these principal-agent relations operate. It scrambles the answerability pathways between elected representatives, high-level ministerial officials, and front-line workers. This situation fundamentally disrupts how and where consequences might apply when decision-making processes disregard evidence, when incorrect decisions are generated, and when people are harmed as a result. It also makes it difficult for administrative actors (whether human or algorithmic) to build on past experiences to prevent similar harms from occurring in future. In systems whose AI components are marketed as better at learning from data than humans, this outcome is perversely paradoxical.

These political and legal changes are often downplayed as technical, which, too, has agency effects. For one, it places automation and its impacts beyond the scope of what courts and tribunals might address in their decisions. Judges and tribunal members alike may see themselves as lacking the expertise to evaluate and effectively remedy the harms caused by automated systems. They may deny their institution's own agency—its power and ability—to tackle what are mere technical or "process" issues. These approaches delegitimize the concerns of people who regularly experience digital government's "sharp edges." They cannot easily pull accountability levers to stop or to correct the outcome assembly line. Crucial aspects of their lives—access to vital benefits, services, status—end up rendered as outputs. Agency effects extend to public officials working within the assembly line, whose expertise is sidelined in favour of faster processing times and whose productivity is algorithmically monitored and managed. The broader public is affected, too, as assembly-line decision-making becomes installed through procurement or policy-setting processes, escaping the scrutiny of legislatures.

Approaching the situation as solely technical emboldens governments to continue pursuing automation while disregarding what are by now predictable consequences, not bugs or glitches. It encourages digital transformation boosters to trivialize serious concerns about how accountability might flow when decisions are produced assembly-line style by a mix of human officials and algorithmic tools. One might misinterpret these critiques as romanticizing administrative fairness. But fairness mechanisms do vital quality control work. Civil servants' obligation to act in the public interest, for example, and their responsibility to review and decide

individual applications (rather than to “process” segments of a file or dossier) empowers them to reach conclusions supported by evidence. Sometimes that requires slowing down production. Their duty to justify decisions to colleagues, superiors, and to the people whose lives will be impacted prevents officials from reaching the worst, or most unreasonable, outcomes.

Without these mechanisms, we risk an administrative governance system where no one—no public official or institution—is responsible. Put simply, it becomes a file-processing and decision-producing system governed entirely by agents, whose principals (or, perhaps, principles) evade us.

DOI: 10.5281/zenodo.23666309

MAIEI

Chapter 14: AI and the Body – Biometrics, Health Data, & Bodily Autonomy

14.1 Confessions in the Cloud: How 140 Million Therapy Exchanges Became Training Data

By Annie Gilbertson, Proof News

14.2 From Gender Affirming Care to Reproductive Rights: The Surveillance Risk of Bodily Data

By Jess Reia, Digital Technology for Democracy Lab and School of Data Science at the University of Virginia

14.3 The Biometric Dragnet

By Emile Dirks & Moses Karanja, The Citizen Lab

DOI: 10.5281/zenodo.23666309

14.1 Confessions in the Cloud: How 140 Million Therapy Exchanges Became Training Data

By Annie Gilbertson, Proof News

Jennifer Kamrass confided her worries to her therapist: her marriage, her finances, and her self-esteem. Every intimate detail shared over weeks of therapy in 2021 was captured and stored by Talkspace, the digital mental health platform that provides therapy over text messaging. When Kamrass, a nurse practitioner, later sued her former employer for pregnancy discrimination, those messaged sessions were subpoenaed and entered into the public court record.

"When I came to understand how much information they had, I was shocked," Kamrass' therapist told me earlier this year. The therapist declined to be named out of concern for her reputation. Kamrass didn't return requests for comment.

It's part of a larger trend we've been reporting on at Proof News, a new nonprofit newsroom devoted to AI accountability. Tech companies are amassing troves of intimate health data including conversations with therapists, doctors and other health professionals.

By now, you are familiar with what is gobbled up to train AI: The leading AI models are famously built on large swaths of the internet as well as vast quantities of books and news articles — a fact one tech executive called "the largest theft of labor in human history."

But it's not just the frontier labs. Models are being tailored for specific industries: law, retail, finance, pharmaceuticals, medicine — and therapy.

CEO Jon Cohen has told investors that Talkspace has built "one of the largest mental health data banks in the world," containing 140 million message exchanges. Reports to investors show the company used this de-identified data to train an AI chatbot released earlier this year.

The company insists the data is protected. In Kamrass' case, Talkspace's chief legal officer, John Reilly, told Proof that Talkspace did not produce the transcripts and that "the only individual who had access to and could disclose the records at the time of the Subpoena was Plaintiff, Ms. Kamrass, herself."

The company's privacy policy shows patients like Kamrass agree to have their therapy sessions stored, and the policy discloses the use of chat, audio, and video communication to "develop new products."

Yet people rarely read terms-of-service agreements. "It's very different than in actual human therapy where there's a lot of training about the informed consent process," said Jodi Halpern, a psychiatrist and professor of bioethics at the University of California, Berkeley. "They just click it."

DOI: 10.5281/zenodo.23666309

MAIEI

The HIPAA Loophole

Proof's reporting sparked outrage from readers and mental health providers. One of the most common reader reactions was the question: Isn't this a violation of HIPAA?

The Health Insurance Portability and Accountability Act turned 30 this year and was designed primarily for data storage and transmission by US providers and insurers, not to protect the modern consumer. The law gives special protection for therapy notes, but generally, once personal identifiers like names and Social Security numbers are stripped, health data is "no longer protected."

"De-identification is an ownership transfer mechanism," Jim Calhoun, a veteran of the AI startup world, told me. Calhoun now runs a foundation developing open standards for companies to have control over their enterprise reasoning and agentic architecture, and as a project, he reviewed dozens of AI company privacy policies and terms of use for data capture loopholes.

"The minute they say de-identified, it's not yours," Calhoun said.

The data is used for more than training. For example, Talkspace used machine learning to develop a suicide risk tool, and its CEO told investors that the tool had been "revalidated" for teens. That led Rebecca Plevin, another Proof journalist, to examine how youth data was being used by the company in its government-funded youth programs.

The investigation, co-published with The Seattle Times in September, found that after millions were spent, far fewer young people received therapy on the platform than projected. Still, the data is useful. Talkspace spokesperson Alana Morales said the company doesn't use youth data to train AI, though when asked how teens "revalidated" the risk algorithm, she said: "We used our knowledge of teen therapy situations and typical scenarios, along with our clinical guidance, to refine the suicide risk algorithm to optimize for the teen population – but did not use specific teen data to train it."

The Pushback

Not everyone is happy about the hoovering up of health data. Rebecca also spoke to Kaiser Permanente therapists in Northern California who are using contract negotiations to push against the proliferation of AI in therapy settings.

"Is five minutes of my time worth the possibility of a data breach?" Jess Metzinger, an associate clinical social worker at Kaiser, told Proof, referring to the use of AI transcription in sessions. "The temptation is not greater than all of the concerns I have."

State legislatures have stepped in: Colorado, Illinois, Maine, and Rhode Island now require providers to gain consent before using AI scribes, and many give patients an explicit right to opt out.

In California, a bill sent to Gov. Gavin Newsom limited the use of AI interaction with patients, isolating it to administrative support work. The bill mandated the right to care even if a patient denied the use of AI and authorized enforcement of therapy records confidentiality protections.

Backed by the California Psychological Association and the National Union of Healthcare Workers, the legislation ran into opposition from the most powerful interest group in the state — the California Hospital Association — which argued the restrictions would interfere with routine AI screening and scheduling.

On the last day to weigh in on bills, Newsom vetoed the measure, calling it overly broad and citing some of the same concerns.

Many patients remain at the whims of tech company privacy policies — like the one Kamrass signed before having every word she told her therapist stored with Talkspace. Nowhere does Talkspace's policy offer the option for US patients to opt out: "If you do not want us to share personal data or feel uncomfortable with the ways we use information in order to deliver our Services, please do not use the Services."

DOI: 10.5281/zenodo.23666309

14.2 From Gender Affirming Care to Reproductive Rights: The Surveillance Risk of Bodily Data

By Jess Reia , Digital Technology for Democracy Lab and School of Data Science at the University of Virginia

In 2025, The Department of Justice under Trump requested trans youth's patient files containing names, dates of birth, social security numbers, home addresses, and detailed medical and prescription histories for patients under the age of 19 who received puberty blockers or hormone therapy. In June 2026 in California, one of the states with a more robust data protection framework in the US, a federal judge temporarily blocked hospitals from producing any documents responding to this kind of subpoena from the DOJ. Pharmacy and electronic health records contain some of the most sensitive data about people. They are particularly relevant when services that are provided have been questioned by conservative agendas. Today, access to reproductive care and gender affirming care fall into the categories being closely monitored by governments.

A prosecutor does not need AI to subpoena a medical record, since this legal act predates the expansion of the AI industry. In fact, what AI changes – and why it matters for the discussion on bodily data in 2026 – is the supply and abundance of information. AI has already entered clinical spaces, for instance, with AI tools being used for notetaking by physicians, for translation by hospitals, and for diagnosis of a variety of diseases, posing extra risks for those seeking reproductive or gender affirming care. Data sharing agreements with third parties, data breaches and government requests are the tip of the iceberg when it comes to the use of AI in relation to bodily data of vulnerable communities. Data dealers also play a role in this process, since the datafication of our bodies has intensified the reach of the data broker industry, which specializes in purchasing, selling, and trading personal data to create detailed consumer profiles – exposing, simultaneously, the extent of invisible threats and the lack of protections or regulation of data sharing markets.

With the rise of authoritarianism and its widespread influence facilitated by technologies, data collection, sharing and storage become increasingly political. The risks are much higher when authoritarian governments have access to sensitive data and AI regulation is either lacking or not enough to provide appropriate guardrails. What is happening in the US, with the federal government requesting patient data to identify trans people – including minors – should be a cautionary tale for the rest of the world.

Concerns around how AI intensifies and facilitates surveillance go beyond clinical settings. Bodily data, the data constantly and casually being collected about our bodies and behaviours, become important components of training data. Body-centric data has experienced significant growth in the last 10 years, becoming a lucrative industry with no signs of slowing down. While

DOI: 10.5281/zenodo.23666309

this real-time monitoring can improve services and products, and offer insights in most realms of life, it also creates opportunities for surveillance, monitoring and control. This increasing datafication in the most intimate corners of our lives creates various risks, disproportionately impacting historically marginalized communities. Gender minorities are facing some of the most drastic effects of the bodily data economy.

When thinking about the state of AI ethics in relation to bodily data in 2026, there are two main areas in need of urgent attention: reproductive rights (especially abortion) and gender affirming care (GAC) for trans people. Some risks are easier to map, like those related to health care, but others can be harder to identify and address. Health care became a metaphorical battlefield in the US for trans individuals seeking GAC, and people in need of reproductive care. Governments in the UK and Canada use a similar playbook, with the UK Court Ruling in 2025 stating that biological sex should guide the use of facilities (including hospitals), and Canadian provinces like Alberta invoking the notwithstanding clause of the Canadian Charter of Rights and Freedoms to strip rights from gender diverse people.

Less obvious but still relevant to this conversation are adjacent technologies like automated license plate readers (ALPRs) and facial recognition technologies (FRT). These AI-powered technologies can rely on biometrics or location data and, although not so obviously connected to gender affirming and reproductive care, end up facilitating tracking of people's movements, including when they are seeking access to medical services. They offer ways around a subpoena, because a camera can show who was outside a GAC clinic, for instance, without anyone opening a record or going through the legal system to get access to data. The misuse of these systems is not uncommon, as Flock cameras are showcasing, and create the opportunity for state and private surveillance that put gender minorities and even survivors of domestic violence at risk of being monitored.

Scholars and advocates have proposed various ways to address this issue, from data sanctuaries to avoid data sharing, to concepts like "databody integrity," which recommends governments and companies recognize people's rights to control and protect their unique physiological data. Based on previous work, I see no one-size-fits-all solution, as we should tackle this problem at different levels. One of the most important steps is to create digital rights frameworks centred on gender minorities that include responses to specific challenges posed by AI and cover, for example, issues of privacy, bias and right to be forgotten. Such a framework would help protect sensitive data and add additional layers between government requests and institutions providing GAC and reproductive care. Other important steps are to regulate the data broker industry, define the boundaries of responsible AI training data, and contest the deployment of AI-powered cameras in public spaces.

DOI: 10.5281/zenodo.23666309

MAIEI

14.3 The Biometric Dragnet

By Emile Dirks & Moses Karanja, The Citizen Lab

On April 10, 2026, the European Union (EU) brought its new Entry/Exit System (EES) for non-EU citizens into full operation across the Schengen Area, completing a rollout that began in October 2025. According to the EU, the system would make travel easier and more efficient, while increasing security and preventing irregular migration. Rather than relying solely on paper-based travel documents, non-EU citizens entering the Schengen Area are now required to register their facial or fingerprint information with border authorities.

The idea that fingerprint and facial recognition can prevent immigration fraud is based on confidence in biometrics, or using measurable human traits to verify identities. Biometric systems focus on three broad categories: genetic data like DNA, physiological data like facial geometry, and behavioural data like voices. Tech companies like Idemia, which provides biometric identification products, claim biometric characteristics are so distinctive and unchanging that they are among the most reliable forms of personal identification. Many commercial institutions and tech companies agree. Banks are using voice authentication to identify telebanking customers, and we can now unlock our phones with nothing more than our faces.

If the apparent benefits of biometrics are clear, the risks can be less obvious. Despite the claims of proponents, biometric traits are not inherently immutable and characteristics like voices and faces can change through age or injury, raising the possibility of misidentification. Perhaps more worrying are how advancements in biometrics are ushering in new forms of surveillance, with immense implications in the areas of state repression, privacy, and artificial intelligence (AI).

Enabling State Repression

Around the world law enforcement agencies have enthusiastically embraced new forms of biometric surveillance, deepening state control over vulnerable populations. In Palestine and Tibet, Israeli and Chinese authorities have conducted biometric dragnets (e.g. the mass collection of facial data and DNA from entire populations) as part of larger programs of colonial occupation which exert strict political controls via mass surveillance and targeted political repression. In the United States, Immigration and Customs Enforcement agents have used facial recognition apps to identify migrants who are then violently detained.

Expansive biometric surveillance can have a chilling effect. In Moscow, authorities have arrested activists after identifying them via the city's network of facial recognition cameras, while police in Delhi have deployed similar tools to capture people with criminal records participating in recent anti-government protests. As biometric tools are more widely adopted, the kinds of "preventive repression" seen in Moscow and Delhi will likely become more common, as

DOI: 10.5281/zenodo.23666309

authorities in these and other countries use biometric surveillance to identify opponents and dissuade would-be critics from assembling.

Privacy

Biometric data collection also poses massive threats to privacy. Smartwatches and other wearable products collect data on everything from our heart rate to our sleeping patterns. In exchange for these conveniences, tech companies harvest user data in ways we barely understand and often cannot control.

Soon these wearable devices could be used to capture other's biometric data, too. In June 2026, Meta abandoned plans to embed facial recognition capabilities in its smart glasses, but only after journalists exposed the program. Placing advanced surveillance capabilities in the hands of ordinary consumers would be a privacy nightmare. Consumer facial recognition tools would open the door for unprecedented abuse, including stalking and sexual harassment. Smart glasses are already prohibited in courts in England and Wales, while security concerns have led Australian authorities to consider banning their use in government offices.

Massive biometric databases have also become enticing targets for malicious actors. Hackers have accessed poorly secured databases, fueling black markets for illicitly obtained data. In 2018, security flaws in Aadhaar, India's national biometric-enabled identification system, allowed people to buy administrative access to the personal data of a billion people. Preventing security breaches will remain a persistent challenge for any organization that collects and stores biometric data *en masse*.

Artificial Intelligence

Compounding these risks is the rise of AI. Biometric systems are increasingly run on AI models, and tech companies have leveraged their considerable power to collect massive amounts of biometric data to train their models. US-based Clearview AI built its facial recognition system by scraping billions of photos from social media and other websites, then sold access to US and Canadian law enforcement, turning ordinary people's faces into raw material to power state surveillance.

Elsewhere, AI companies have used biometric data to register their own users. World, co-founded by OpenAI CEO Sam Altman, has marketed its "World ID" as a biometric-enabled proof-of-personhood credential that can authenticate users across multiple platforms. Know Your Customer (KYC) requirements, long familiar in the world of banking, are now being used "to tell the difference between humans and bots." World's efforts to authenticate its users through iris and facial images, however, has faced intense pushback, with authorities in Kenya, Thailand, Argentina, and the Philippines seeking to halt its operations on privacy grounds.

DOI: 10.5281/zenodo.23666309

MAIEI

As states fight to control their citizens' data, advances in AI are throwing into question whether biometrics are actually foolproof. It is now possible to create deepfakes of facial likenesses in realtime and convincingly duplicate voices. Criminals have leveraged these tools to conduct elaborate scams, including one in 2024 where a financial officer in Hong Kong was tricked into transferring US$25 million to people who used deepfake software to impersonate the officer's colleagues on a video call. For ordinary people, phone scams will become even more difficult to detect, as criminal actors utilize inexpensive and increasingly sophisticated AI-enabled voice cloning tools to impersonate the victims' loved ones.

Addressing Technical and Political Problems

Ultimately, biometric surveillance poses considerable technical and political problems, as shown by the EU's new Entry/Exit system. The requirement that travellers register their biometrics has caused delays and missed flights during peak travel periods, and demands from EU member states for more time to implement EES border checks. Over time, these technical challenges may be resolved. What will remain are the political implications of this new biometric-enabled system of border management.

State and private biometric surveillance threatens to become inescapable. Civil society must push back. Meta's reversal on facial recognition glasses and Clearview AI's withdrawal from the Canadian market demonstrate how courts, activism, and public opinion can constrain the spread of these powerful tools. Without proper oversight, whatever benefits biometric tools bring will come at the steep cost of our individual privacy and collective wellbeing.

DOI: 10.5281/zenodo.23666309

MAIEI

Chapter 15: Open Models & the Communities Building With Them

15.1 The People's AI Stack: Mapping the Social and Political Choices Behind Artificial Intelligence

By Stephanie Dick, Simon Fraser University (SFU), Fergus Linley-Mota, the SFU Dialogue on Technology Project & Hillary Bergshoeff, IATSE 891

15.2 Re-Remix: Open Source Models for the World

By Vukosi Marivate, Independent International Scientific Panel on AI and University of Pretoria; Emma Ruttkamp-Bloem, University of Pretoria & Chijioke Okorie, University of Pretoria

15.3 ParcoursLab: A Course Recommender Made by Students, for Students

By Riccardo Di Sipio & Elena Yunusov, Human Feedback Foundation; Aaryan Shroff, University of Waterloo & Claudel Rheault, University of Edinburgh

DOI: 10.5281/zenodo.23666309

15.1 The People's AI Stack: Mapping the Social and Political Choices Behind Artificial Intelligence

By Stephanie Dick[7], Fergus Linley-Mota[8] & Hillary Bergshoeff[9]

AI Technology Stacks are often used to condense a complex system of interconnected factors into simple, legible diagrams. A conventional "tech stack" usually assumes computing infrastructure at its base, followed by data, models, and applications. But every stack is also an argument about what belongs within the system. When workers, communities, public resources, environmental consequences, workplace authority, and cultural meaning are left outside the frame, they can appear incidental to AI rather than integral to its existence.

We offer an attempt to redraw that picture: the People's AI Stack.

The People's AI Stack comprises eleven layers, with public governance running throughout. Rather than defining people solely by their relationship to technology, it asks, at each layer, how AI affects their lives and where they might exercise collective authority. Instead of trying to make AI safer after decisive choices have been made, the stack aims to give people authority over why, how, whether, and where these systems are being built and used in the first place.

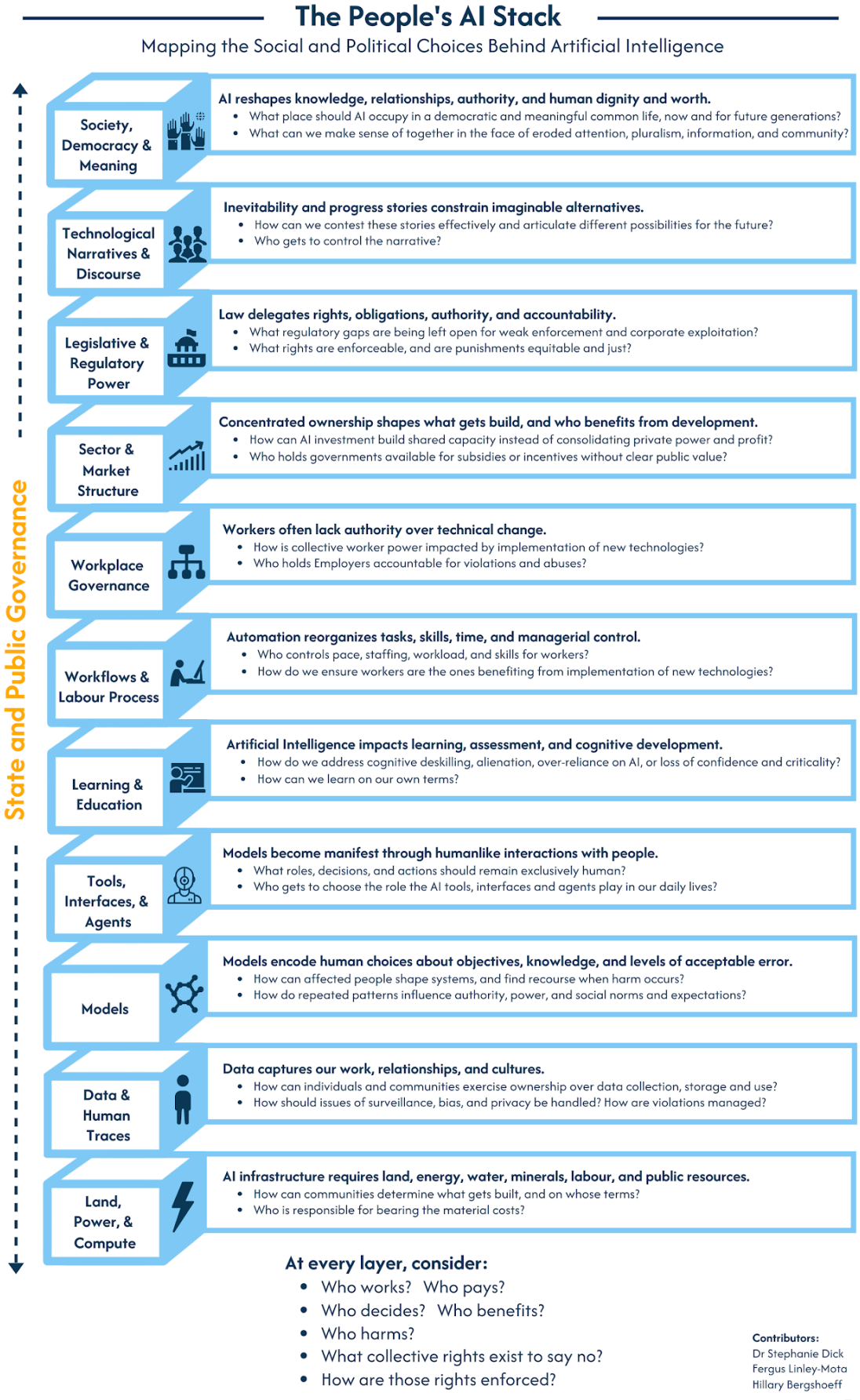
The People's AI Stack can be read in any direction, or as a map, because influence is rarely limited to one direction. An application may depend on infrastructure, but procurement shapes how that infrastructure is built; a collective agreement may determine how a tool is used but not sufficiently protect the worker using it; and a narrative about an inevitable "AI race" can weaken democratic scrutiny when it's needed most.

The Stack's central claim is simple: all AI policies impact workers, so all AI policy becomes labour policy – and labour policy reaches far beyond the workplace.

The Stack supports two related forms of inquiry, striking a balance between the reactive and the proactive. First, it broadens the range of identifiable risks and possible coalitions that emerge across the sociotechnical layers of an AI system. Second, it helps us ask the questions necessary to imagine a more democratic future, and develop the concrete interventions needed to realize that vision.

[7] Simon Fraser University (SFU)
[8] The SFU Dialogue on Technology Project
[9] IATSE 891

DOI: 10.5281/zenodo.23666309



DOI: 10.5281/zenodo.23666309

When applied in practice, the Stack can help different collectives recognize where their concerns intersect and navigate towards shared goals. Universally aligned priorities and agreed-upon principles are not required to identify decisions they could shape together.

Consider truck drivers and motion-picture workers, who might appear to face very different technological pressures. Truck drivers might enter the stack through Workflows & Labour Process and Workplace Governance, concerned about in-cab surveillance, algorithmic performance management, and the future of their jobs. Motion-picture workers might begin at Data & Human Traces, asking who controls digital scans, voices, performances, and other traces of creative production. Following those concerns through the stack could reveal a shared question: who controls the data workers generate simply by doing their jobs?

From that question, a common demand becomes possible: worker-generated data should not be used to surveil, discipline, reproduce, or train systems designed to replace workers without collectively negotiated terms. The protections would look different across industries. Truck drivers might bargain over driver-facing cameras, automated safety scores, telematics, and the use of driving data in autonomous systems; motion-picture workers might bargain over digital replicas, performance capture, and reuse of their work in generative systems. But the underlying principle would be the same: data produced through labour should be treated as part of the labour relationship, rather than as free raw material for the development of automation. What begins as seemingly separate sector-specific concerns could become a shared demand for collective control over the traces of work itself.

The People's AI Stack thus offers a way to make connections and disagreements visible, locating specific decisions around which collective authority might be built. It allows different stakeholders to find a way "in" to AI deliberation - and, potentially, toward one another. Individual choices about whether or when to use AI cannot settle these institutional questions. Coalitions could turn overlapping concerns into shared demands, and concrete gains no group could secure alone.

AI development can't be understood as a singular process or an inevitable trajectory. The People's AI Stack is a tool that can be used to locate power, responsibility, and opportunities for intervention across AI's sociotechnical system. It helps organizers, civil-society organizations, workers, policymakers, and affected communities locate concrete decisions that might otherwise be presented as technical facts or inevitable developments. Most importantly, by widening our field of view, it expands the range of places where people can act.

15.2 Re-Remix: Open Source Models for the World

By Vukosi Marivate[10], Emma Ruttkamp-Bloem[11], and Chijioke Okorie[12]

In his 2008 documentary RIP: A Remix Manifesto, Brett Gaylor outlined a central tension of the digital age: culture always builds on the past, but the past always tries to control the future. The documentary follows the career of mashup artist Girl Talk to explore the collision between modern digital remix culture and traditional corporate copyright enforcement. We saw this play out again in music with projects like Kanye 2049, where a creator remixed existing audio stems to build a highly innovative, alternate-universe album. It was a triumph of community creativity, yet it existed in constant peril of corporate takedowns.

In 2026, the battleground has shifted from music samples to neural weights, but the power dynamics remain the same. Today's remixers are grassroots communities and academic labs across the African continent, networks like the African Natural Language Processing communities. For the last decade, these communities have fought against data scarcity, moving from a landscape where English-only word embeddings were the norm, to actively annotating and building corpora for languages like isiZulu.

To do this, we rely on building on a culture of openness, covering open data, open code and open weights. We take the foundation models of the Global North and fine-tune them, remixing them, to serve our own cultural contexts. But the open source ecosystem is currently under intense pressure. Under the banner of "dual-use security risks," the largest AI labs are attempting to close the door behind them, pushing to gatekeep foundational weights.

The prevailing debate of 2026 presents a false binary: either models must be completely "open" (which often means open weights but obscured, extractive training data), or they must be closed to mitigate "dual-use security risks". Both extremes fail the communities whose digital heritage fuels these systems.

When models are entirely closed, grassroots networks hit an infrastructure wall, forcing a reliance on Western APIs that strips away chances for establishing local digital sovereignty. But traditional "open" licenses have proven equally perilous. When African NLP practitioners fought data scarcity by painstakingly building corpora for languages like isiZulu, applying standard open licenses often resulted in data being scraped by massive commercial labs with zero provenance, reciprocity, or cultural context. A digital clear-cutting of the commons.

[10] Independent International Scientific Panel for AI, United Nations; African Institute for Data Science and Artificial Intelligence (AfriDSAI), University of Pretoria
[11] Department of Philosophy and African Institute for Data Science and Artificial Intelligence (AfriDSAI), University of Pretoria.
[12] Department of Private Law, Data Science Law Lab and African Institute for Data Science and Artificial Intelligence (AfriDSAI), University of Pretoria

In response, practitioners are moving beyond the binary, pioneering a spectrum of protective licensing that enables community innovation while blocking corporate capture. We see this resistance globally and locally. In New Zealand, the Te Hiku Media License operationalised Māori data sovereignty, ensuring that language data is shared only under conditions of *kaitiakitanga* (guardianship), rather than ownership.

On the African continent, we are establishing new paradigms by writing these frameworks ourselves. Through the Nwulite Obodo Open Data license (NOODL), spearheaded by Chijioke Okorie (our co-author) and Melissa Omino, and co-author Vukosi Marivate's collaborative work on the Esethu license, we are legally mandating provenance and relational ethics. We designed these frameworks as a call to protect and amplify the voices of those whose data is viewed as belonging to anyone willing to scrape it. These frameworks allow grassroots remixers (academic labs, community builders, sovereign initiatives) to fine-tune and innovate freely, while legally preventing the frictionless extraction of cultural digital heritage by trillion-dollar entities. They are not closing the door; they are finally installing a lock that the community holds the key to.

Mitigating "dual-use security risks" and ensuring sovereignty are not new ideals or goals; (disruptive) technologies have always pushed the boundaries of security and sovereignty. As was the case with other technologies, equitable and security-sensitive openness will be imperative. For example, while not a perfect system, the patent system has always required disclosures in exchange for the monopoly granted to patentees and even for state secrets patents, disclosures to specified agencies were still required. In essence, risk has never justified closure and the key challenge is addressing how to disclose or how to open. In the case of model weights, disclosures and openness can and should be required; and in the interests of equity, this should be required of those who publish model weights and made to the public. The task of governance and regulation should be to not only require disclosures but to set parameters for it to happen appropriately. This task is not one for a select few.

As we go deeper into the debate on open-weight models and security concerns, we must reject the premise that our role is merely to learn how to navigate boundaries set by trillion-dollar labs. The question is no longer whether a wider, diverse community will be granted a seat at the table. We are already building the table, forging the locks, and holding the keys. We choose to do the hard work of defining these new paradigms ourselves, because our digital sovereignty, and everything that comes next, depends on it.

Agaghi aso mgbagbu ghara ogu (ibo-Igbo)

Ku kondzelela i mana ku humelela (tso-Xitsonga)

15.3 ParcoursLab: A Course Recommender Made by Students, for Students

By Riccardo Di Sipio & Elena Yunusov, Human Feedback Foundation;
Aaryan Shroff, University of Waterloo & Claudel Rheault, University of Edinburgh

In May 2026, Coursera absorbed Udemy into a single skills platform with 290 million learners. Nothing changed for learners on day one, but the catalog, the pricing and the AI layer are now in the hands of one company. Meanwhile, a team of undergraduates at the University of Waterloo built a course recommender and hosted it on a $100 grant. ParcoursLab explores what the course recommendation experience looks like when students, not revenue, set its terms.

It Takes a Village

When no enterprise roadmap sets the agenda, development looks more like a village: students, mentors and volunteers building and deciding together, transparently. ParcoursLab (the French *parcours*, a journey or course of study, plus the English *lab*, a nod to its bilingual roots and experimental nature) was built by the students who would actually use it, with nothing hidden from them about how it works. It grew out of the Toronto-based Human Feedback Foundation's mentorship programs, including RAISE, which now run an ever-growing cohort of undergraduates across Canadian universities. While responsible AI is often discussed through research, policy, and governance, ParcoursLab offers a complementary perspective: what those principles look like when put into practice in a student-built prototype. SAIER is one of the places where practice gets recorded, and ParcoursLab is our contribution to it. It set out to test a claim: that an AI *cultivated* in respect of human dignity, to borrow from Pope Leo XIV's encyclical *Magnifica Humanitas*, is better built by the people who will use it.

ParcoursLab is a conversational course-planning assistant. Users describe their role and goals, and it suggests the skills and courses that could close the gap. Most recommenders ask for a resume upload or a long skills checklist and disclose little about how they work. ParcoursLab discloses its methodology upfront and asks instead for a short, LinkedIn-style blurb. An AI model maps it onto ESCO, the European Commission's human-curated skills ontology, and suggests skills that might close the gap to the stated goal. Users can edit any suggestion or goal, and a mountain-goat assistant collects *keep-adjust-reject* feedback, with each new batch explaining its reasoning. The team itself selected the mountain goat to represent a resourceful, resilient explorer, sure-footed enough to cross steep, unfamiliar ground. The result is a co-created exploration of skills, with human users actively editing the AI interpretation, leading to recommendations that are grounded in that exchange.

DOI: 10.5281/zenodo.23666309

The Lessons We Learned

What follows is what building ParcoursLab, and trying to get it adopted, taught us.

On development, the hardest problem was finding a proposal students could see themselves in, one where a design as small as a chat window could plausibly carry something like human dignity. For ParcoursLab, that turned out to be embodied in the assistant's conversation itself with transparency and trust built turn by turn. The gap it filled was real: course planning at the university has no single home, the information exists but is scattered, and the students themselves reported how the jargon needed, just to talk to an advisor, takes most of a first semester to learn. The mountain goat front-loads part of that conversation, so an advisor spends less time decoding what a student needs and more time actually helping, augmenting the advisor's judgment rather than replacing it.

The team structure mattered as much as the mission. A mentor set the product direction and the most senior student ran the project day to day, while the other students split across backend and database, frontend and UX, AI, with deliberately overlapping roles.

On deployment, the first obstacle was clearly the infrastructure. The program ends with a public demo at an AI Tinkerers event each semester, and a live demo means paying for hosting and compute. An OpenRouter nonprofit grant covered $100 of it, enough to get through a semester but not a durable answer. Cheap, reliable compute for small teams remains an open problem. What paid off was the demo itself: presenting to a room, not just a supervisor, taught skills that a project seen only by its mentors never would have, and built a professional network in the process. The team's demo was voted best presentation of the event.

We also demoed ParcoursLab to a Student Success and Career Advisor at a University of Toronto college, who told us that many students arrive with almost no sense of what comes next, and that sessions are often spent working that out. The advisor saw value in a tool that front-loads the conversation and asked for printable material, which we added. The advisor also offered to help find a route to institutional adoption, which has not yet materialized. Our own developers used it to map out database courses for work and found electives to fulfill their degree requirements. For now, ParcoursLab remains a demo, and finding it an institutional home is the next step.

This is what resisting consolidation costs in practice. The combined company can set defaults for 290 million learners, while ParcoursLab, with a working prototype and a supportive advisor, remains a demo on a $100 grant. The remaining challenges are access to compute and an institutional home.

DOI: 10.5281/zenodo.23666309

MAIEI

Recommendations

For another organization weighing whether to try something similar, we recommend finding a real, unglamorous gap first. Be clear on the aspects of the project that elevate human dignity and staff the team accordingly. Budget for compute before anything else, because that is where an under-resourced, community-built alternative will actually get stuck. Find an institutional champion while you build, and look early for who can open doors.

DOI: 10.5281/zenodo.23666309

PART VI: LITERACY, COALITIONS, & GOVERNMENT

Chapter 16: AI Literacy Across the Lifespan

Chapter 17: Civil Society – Austerity, Libraries, Coalitions, & Refusal

Chapter 18: AI in Government – What It Buys and Who Writes the Rules

DOI: 10.5281/zenodo.23666309

Chapter 16: AI Literacy Across the Lifespan

16.1 AI Literacy in K-12: To Whose End?

By Mia Travers-Hayward, Canadian Teachers' Federation

16.2 Prioritizing Responsible AI Education for Older Adults

By Tina Crouse, Independent Researcher

16.3 What Nine Years of STEM Outreach Data Can Teach Us About Designing AI Literacy

By Sara Jameel, Bahareh Goodarzi & Navid Sharifi, Concordia University

DOI: 10.5281/zenodo.23666309

16.1 AI Literacy in K-12: To Whose End?

By Mia Travers-Hayward, Canadian Teachers' Federation

On September 9th, 2026, Evan Solomon, Canada's Minister of AI and Digital Innovation, announced a National AI Literacy Initiative. By 2031, Amii, the delivery partner, aims to provide 50,000 K-12 teachers with foundational and practical AI knowledge through their *AI for All: Essentials for Educators* program. The course descriptions imply that teachers should strive to use AI, so long as they do so responsibly. But this assumes that AI can, indeed, be used responsibly in K-12 classrooms, which remains an open question. In fact, evidence is signalling the contrary; that applying AI in schools is a risky endeavour, riddled with threats to children's safety, privacy, and cognition. In its current industry-driven form, the AI literacy ecosystem in public education is weakening the public's capacity to define the contours of its own digital future.

Public Educators Calling for Support

For the past three years, I have listened to educators' concerns about how AI is reshaping their profession. In May 2026, at a Provincial Council meeting of the Quebec Provincial Association of Teachers (QPAT), I invited 120 educators to share perspectives on this new frontier of digital transformation. A long line formed behind a microphone, as they described rising stress from watching student learning erode, from administrators shrugging off their concerns, and a persistent lack of guidance on how to face an overnight sea change in classroom dynamics.

That same spring, the Canadian Teachers' Federation (CTF/FCE) surveyed 6,175 educators on the ways in which deprofessionalization was impacting their work. While the survey itself was not focused on tech, an open question asking teachers to reflect on the biggest concern for their profession surfaced approximately 500 responses on the effects of digitalization. A forthcoming CTF/FCE report will highlight their warnings about rapid AI integration, and how it threatens teachers' job security and fuels an overreliance among students and staff that undermines quality public education. Many called out the pressure to embrace AI tools in their practice without suitable training that addresses the real impacts of emerging tech on young learners. While teachers are eager for support, dominant incarnations of AI literacy leave limited room for professional discretion.

Capturing "Responsible Use" in Classrooms

The AI hype machine has inspired widespread calls for programs to help people understand these complex technologies. At the CTF/FCE, we lobbied for policies that enable the responsible use of AI in K-12 schools, including comprehensive critical AI literacy for educators. Our conceptualization of responsible use is about protecting human-centred education and the core relational dimensions of teaching and learning. In this view, AI use in schools is not

inevitable, but *conditional* on a system's alignment with human rights, safety, and pedagogical appropriateness. Yet, in their ceaseless race for control, tech organizations are flexing their power to influence norms about what we need to know and do in the AI age. Adoption, of course, is the central goal.

With billions of dollars at their disposal, leading AI companies continue to secure major institutional partnerships to embed their products into teaching and learning. In the US, Microsoft, OpenAI, and Anthropic partnered with the American Federation of Teachers (AFT) to launch the National Academy for AI Instruction for K-12 educators across the country. The AFT justified the partnership as a "strategic necessity" to ensure educators have a seat at big tech's table, but in doing so they have, in effect, endorsed AI as a certainty in schools. *If we can't beat them, join them*, is not a suitable strategy to redress tech's capture of public education.

AI Literacy as a Democratic Project

The Canadian government has committed $13 million to Amii to expand AI literacy, yet the public has not been consulted on the shape of the programs. Should educators, students, and families not have a say in how AI will be framed in public schools, which serve the vast majority of children? A federal commitment to supporting teachers to learn AI essentials is a positive step; however, certain topics covered by Amii's K-12 learning kits – like *Harnessing the potential of ChatGPT* and *Careers reimagined* – may conflict with the growing resistance to a tech-centric education. Instead of relying on organizations with a heavy stake in adoption to define the future of AI on our behalf, K-12 AI literacy curriculum should be considered a public good. Governments must consult meaningfully with those invested in young people's capacity to critically engage with emerging technology, to ensure programs align with the public interest before they scale.

AI companies competing for K-12 domination claim they wish to empower educators, not replace them. Educators are viewed as a gateway to AI adoption, but what if they have no interest in this intermediary role? And are we really to entrust these organizations, beholden to shareholders, with setting the standard for how AI should be integrated into schools? K-12 AI literacy should be a shared democratic project, grounded in public consultation, transparent evaluation, and active partnership with educators, unions, and ministries. As Amii speeds ahead with rolling out their publicly-funded AI literacy programs, the federal government must insist that the organization immediately convene teachers and their unions, as well as students and their families, to ensure those on the frontlines have a say in how children are taught about tech in schools.

DOI: 10.5281/zenodo.23666309

MAIEI

16.2 Prioritizing Responsible AI Education for Older Adults

By Tina Crouse, Independent Researcher

The rapid integration of artificial intelligence into healthcare, finance, and civic life demands targeted digital literacy efforts for vulnerable populations, particularly older adults. By 2050, the global population of adults aged 65+ is projected to more than double (World Social Report, UN Dept of Economic and Social Affairs), with older women comprising the majority and facing disproportionate economic marginalization and compounded poverty (Gender Matters in an Aging World 2026, United Nations, Dept of Economic and Social Affairs). Without proactive education, this cohort risks exploitation through biased algorithms, privacy violations, and uninformed use of AI-driven services. Responsible AI education is critical to align technological advancement with sustainable development goals by empowering seniors as informed participants. As a Responsible AI Practitioner, I had the pleasure of teaching a course to seniors using this approach in the summer of 2026 sponsored by the Ottawa Responsible AI Hub (ORAH).

Currently, AI is deeply embedded in daily systems, yet deployment often outpaces safeguards. With 856 million older adults globally (United Nations Population Division, World Population Prospects 2024), they face heightened vulnerability due to limited digital experience and growing reliance on AI-powered healthcare and financial tools (Digital Skills for Older Adults, DIVERSITY Institute). Education grounded in Responsible AI principles—fairness, transparency, privacy, safety, inclusiveness, and accountability (6 Principles of Responsible AI)—offers a path to mitigate these risks and ensures AI serves seniors rather than exploits them.

Seniors are not passive consumers; they remain significant economic actors as major donors to NGOs, healthcare users, and civic contributors. However, limited familiarity with data ecosystems makes them susceptible to corporate exploitation, fraud, and manipulative AI communications. Teaching digital literacy focused on informed consent empowers seniors to navigate these systems safely.

Seniors-focused Responsible AI education should focus on three pillars:

1. **Data Privacy and Digital Hygiene:** Teaching seniors to manage app permissions, recognize phishing, and control data sharing reduces exploitation. Simple guidance on privacy settings, two-factor authentication, and suspicious communications can significantly lower risks.

2. **Informed Use of AI Services:** In healthcare, where AI drives diagnostics and assistive technologies, education fosters critical evaluation of AI outputs and informed consent. Seniors should understand AI recommendations are not infallible and learn to ask clinicians to explain AI-driven decisions, maintaining dignity and autonomy in their care.

3. **Ethical Financial and Civic Engagement:** As NGOs adopt AI for fundraising, educating seniors and organizations on transparent, bias-aware systems protects autonomy and sustains trust. Seniors should recognize manipulative fundraising tactics or hidden biases in financial advice algorithms that could jeopardize their retirement security. Accessible, ethical digital civic platforms also ensure seniors can participate fully in democratic processes without algorithmic exclusion (Ethics Guidelines European Commission).

Following the three pillars, Responsible AI literacy can be mapped to multiple Sustainable Development Goals:

- **SDG 3 (Good Health):** Supports dignified aging by enabling safe engagement with AI-driven healthcare, allowing seniors to collaborate effectively with providers.
- **SDG 4 (Quality Education):** Digital literacy and AI ethics constitute essential lifelong learning made necessary by the continuous evolution of artificial intelligence.
- **SDG 5 (Gender Equality):** Addresses the disproportionate AI risks older women face due to their over-representation and persistent economic disadvantage.
- **SDG 8 (Economic Growth):** Protects economic security by teaching seniors to navigate AI-driven financial tools and fraud detection systems.
- **SDG 16 (Peace and Justice):** Builds trust by promoting transparency in digital interactions, helping seniors resist disinformation and manipulative systems.

Critics may question resource allocation to seniors over other groups, but the intersection of aging, gender, and poverty creates unique compounded risks not present in common discussions of AI training. Effective solutions include designing senior-specific programs and partnering with libraries, senior centers, and nonprofits to bridge the digital divide.

As a Responsible AI Practitioner and a Seniors Course Provider, I created the following checklist as a basis for curriculum:

☐ Ensure transparency is central to a senior's understanding of how AI operates, utilizing strategies like the two-prompt system (one query, and one demanding the AI prove its answer is accurate).

☐ Frame AI training around suggestions rather than mandatory decisions, ensuring seniors maintain agency.

☐ Demonstrate how human oversight and authority should be utilized in care decisions, with both the doctor and the patient reviewing AI outputs.

☐ Mitigate the risk that societal challenges (such as under-representation, discrimination, and ageism) become accepted AI outputs by encouraging seniors to review results through their own experience.

☐ Teach configurable privacy settings and default protections to limit unnecessary data collection.

☐ Provide practical examples for seniors to give or withdraw consent, ensuring choices in purchases, healthcare, and banking remain entirely within their control.

Educating older adults in Responsible AI is not a niche intervention but a foundational component of social justice. By empowering seniors with agency, privacy, and critical awareness, we ensure AI contributes to sustainable development rather than entrenching inequalities, safeguarding human dignity and shaping equitable technological futures.

DOI: 10.5281/zenodo.23666309

16.3 What Nine Years of STEM Outreach Data Can Teach Us About Designing AI Literacy

By Sara Jameel , Bahareh Goodarzi & Navid Sharifi , Concordia University

Ask a room of teenage girls to close their eyes and picture someone who works in artificial intelligence, and most will describe the same person: a man, at a desk, somewhere far away. It's not a guess. We've run this exercise at the start of an AI ethics workshop for two years running, inside GirlSET, a Concordia University STEM outreach program for girls aged 14 to 17. The answer barely changes. And it tells you something that most AI literacy programming in 2026 still gets wrong: the gap isn't about understanding how AI works. It's about who gets to imagine themselves inside it.

This matters right now because all levels of schools are scrambling. Generative AI arrived faster than policy could keep up, and the dominant response has been defensive: detection tools, usage bans (Duffy, 2026), disclaimers bolted onto assignments. Almost none of it asks a more basic question, one that outreach programs like ours have been quietly answering for years: how do you actually build a young person's confidence and critical judgment around a technical field they've been taught doesn't belong to them?

We have data (Goodarzi et al., 2025; Jameel et al., 2026) on this. Between 2017 and 2026, GirlSET has run ten cohorts and surveyed 111 alumni about how the program shaped their sense of engineering, technology, and their place in it. Those surveys were about engineering and technology broadly, not AI: the program is nearly a decade old, and its AI ethics workshop was added only two years ago, so the evidence here comes from outreach design that predates the AI content. The findings are consistent. Structured, hands-on, mentorship-driven exposure doesn’t just build interest, it builds a form of agency. Girls who entered the program with an existing curiosity about STEM left describing engineering as something they could reshape, not just study. Nearly 90% said programs like this one can meaningfully address gender disparity in the field.

But one finding stood out, and it's the one AI literacy efforts should be paying closest attention to. When we asked participants to name the barriers women face in engineering, the overwhelming majority pointed to interpersonal bias: being underestimated, dismissed, talked over, having to prove themselves twice. Far fewer named the structural mechanisms underneath it, hiring pipelines, pay gaps, who gets funded and who doesn't. It's not that participants didn't care about the bigger picture. It's that nobody had taught them to see it yet.There is a difference between recognizing bias as an interpersonal experience and understanding it as a structural one. Being underestimated or talked over is visible; the systems that shape who gets hired, promoted, funded, or represented are harder to see. Critical literacy means learning to connect the experience to the structures that produce it.

DOI: 10.5281/zenodo.23666309

MAIEI

This is precisely the gap that current AI literacy efforts risk repeating. Teach a student to "spot the bias" in a chatbot's output, and you've taught them to notice a symptom. Teach them to ask who built the system, what data trained it, and whose interests it optimizes for, and you've taught them to diagnose the disease. Our AI ethics workshop tackles the latter. We don't lecture on facial recognition bias; we walk participants through what happened to Robert Williams (American Civil Liberties Union [ACLU], 2024), wrongfully arrested in Detroit after a facial recognition system misidentified him, and then ask them to redesign the system themselves. We don't explain why Amazon's resume-screening tool (Dastin, 2018) penalized any application mentioning "women's," we let participants discover it, then hand them the pen.

The design principle underneath all of this is simple, and it's the one thing we'd want every educator building AI literacy curricula in 2026 to take away: representation has to come before restriction, and redesign has to come before detection. A student who has only ever been taught to flag what's wrong with an AI system has learned to be suspicious of it. A student who has been asked to imagine what a fairer version would look like has learned to build one.

None of this requires a new theory. It requires borrowing from a field (STEM outreach) that has already run the experiment, tracked the outcomes, and shown what works. As institutions rush to write new AI policy this year, the evidence base for how to teach young people to engage with technical systems critically, and with agency, already exists. It just hasn't been applied to AI yet.

Chapter 17: Civil Society – Austerity, Libraries, Coalitions, & Refusal

17.1 Optimized for Whom? AI, Austerity, and Ontario's Settlement Sector

By Annabelle Bernard, Community Researcher

17.2 When AI Comes to the Library: Public Libraries as Civil Society Infrastructure Under Pressure in the Age of AI – Lessons from a Québec Community Initiative

By Lesly Nzeusseu Kouamou, Eddy Damaris Nono Defo and Jade Maria Moisan, Université de Montréal & Anne Imouza, McGill University

17.3 Who Gets to Shape AI? Building Coalitions for Underrepresented Communities

By Archana Prasad, Isha Singh & Tom Simmons, Gooey.AI & Royal College of Art

17.4 Who Gets to Decide? Building Indigenous AI Governance Through Relationship, Authority and Refusal

By Keegan Leahy, Turtle Island Foundation & Michelle Okere, Indigenous Prosperity Foundation

DOI: 10.5281/zenodo.23666309

17.1 Optimized for Whom? AI, Austerity, and Ontario's Settlement Sector

By Annabelle Bernard , Community Researcher

Author's Note: The terms *community agencies* and *nonprofit* are used interchangeably, I have intentionally not made a distinction between those agencies operating with registered, charitable status and those operating without who would be considered a "non-qualified donee" by Canadian Revenue Agency.

Ontario's community agencies are facing their most urgent governance decision since the pandemic: should AI be adopted by their organizations? Agencies are having to reckon with the use of this new technology while mitigating funding cuts, volunteer shortages and increasingly hostile right-wing sentiment towards their missions (Ontario Non Profit Network, 2026).

During the pandemic, community agencies built up their digital infrastructure overnight, moving sensitive client files and casework online, adopting systems for virtual service delivery and for staff working remotely. Federal and provincial funding followed this transition with emergency pandemic and post-pandemic funds: the federal government forwarded about $883 million to Ontario nonprofits by 2021 (Ontario Non Profit Network, 2021), and the provincial government $150 million in 2023 (Government of Canada, 2024) for post-pandemic recovery.

But now, almost four years since generative AI went public, nonprofits are weighing the internal and external impacts of a technology that is widely understood to potentially undermine democratic institutions (Critical AI Stanford, 2026). Furthermore, nearly two-thirds of Canadian nonprofits using AI have no formal policy guiding its use (Lasby & Baker, 2026). Whether or not agencies can actually develop an AI policy now, given their struggle to find the money and expertise to implement it, they will increasingly work in a landscape where donors are expecting transparency on its use (Riley, 2026).

I had the opportunity in both my role as a manager at a GTA-focused nonprofit funder, and as a lead on a qualitative study of how GTA nonprofits were considering AI adoption from the standpoint of value alignment, to learn directly from community agencies about how they were contemplating the adoption of AI in their agencies. Many agencies were waiting for further guidelines and government funding to be able to consider how they could responsibly use AI for their specific work and mission. Absent were the government funding supports or guidelines that had helped shape pandemic policies. Ironically, agencies were noticing that federal departments had quietly begun adopting AI themselves, and even encouraging nonprofits to follow suit. They

DOI: 10.5281/zenodo.23666309

questioned how their work could be impacted both positively and negatively by a government wielding AI for the purposes of optimization and renewed austerity.

Immigration, Refugees and Citizenship Canada (IRCC) has experimented with AI since 2013 and fully adopted AI-based automation in 2017 (IRCC, 2026). This adoption has not seemingly come with an increased price tag; rather, it came ahead of a pretty massive funding cut led by significant changes in how Canada's immigration system (working visas and student visas among others) operated. The 2025 federal budget asked departments to cut $56.7 billion over four years (Government of Canada, 2025). Ontario will lose 17.3 percent of its settlement funding as part of a total $686 million four-year cut (Keung, 2025; CCPA, 2026). And that is shockingly not even the hardest hit province in Canada. The IRCC cuts come with a newly "optimized" settlement process but optimized for whom? IRCC is betting that higher-skilled immigrants will need less integration support and fewer community services, which doesn't reflect the evolving uncertainties of this new immigration system, such as lengthy security clearances for permanent residency (Searle, 2026).

As of January 2026, almost half of Toronto's settlement agencies are at risk of closing (Baxter, 2026). Settlement agencies which help newcomers with housing, employment, language classes and legal information, are also often the first responders to humanitarian crises. This was highlighted most recently in the spring of 2023, when asylum seekers from African countries arrived in Toronto in significant numbers (UWGT, 2025). In December 2022, the City of Toronto's shelter system had decided that it could no longer house asylum seekers without incurring a budget deficit (this cost was previously shared by IRCC). As refugees arrived in the cold temperatures without access to shelter, settlement agencies and faith groups scrambled to find sleeping spaces and resources for them, often at significant cost (Manucdoc & Draaisma, 2023). Casting this situation into an uncertain future, how will the response to a humanitarian crisis look under an "optimized" federal government and a weakened settlement sector? The City of Toronto itself has expressed alarm for the health of settlement agencies as IRCC funding cuts threaten the necessary scale of essential newcomer services, in Canada's "top immigrant-receiving destination" (City of Toronto, 2025).

The Ontario Nonprofit Network (ONN) named AI's rapid evolution a defining pressure on the sector in 2025 and 2026 with organizations feeling pushed to adopt it for efficiency (ONN, 2025, 2026). In my experience, staffing and funding cuts more than any certainty that AI serves newcomers better are what push settlement agencies toward AI adoption. That said, adoption should not be treated as inevitable. With the right time, investment and ethical frameworks, AI could be a useful tool, easing operational work like data entry, transcribing and grant writing. The trouble is that nonprofits often lack the in-house expertise to evaluate risk and safeguard data (Rotter & Bailkosk, 2025).

In late July 2026, the UN World Food Programme signed a five-year extension of its contract with Palantir, a US data-analytics company (Masurier, 2026). An internal audit leaked the year before had found inadequate risk management, no data privacy guidelines and no exit strategy. The renewal went ahead anyway. It is not hard to see how austerity could push agencies in Ontario and beyond towards similar private partners since AI may be the only support they can afford, and thus similar risks may ensue.

The pandemic showed that public investment can carry a sector through transformation, and that imperative applies to AI too. Funders and governments should pay for the implementation of that governance work, or else it might not happen in time to support the settlement agencies holding on for survival. How quickly Ontario's nonprofits adopt AI will not prove that it is worth adopting, but it will show what happens when austerity rather than planning sets the pace.

DOI: 10.5281/zenodo.23666309

17.2 When AI Comes to the Library: Public Libraries as Civil Society Infrastructure Under Pressure in the Age of AI – Lessons from a Québec Community Initiative

By Lesly Nzeusseu Kouamou, Eddy Damaris Nono Defo, and Jade Maria Moisan, Université de Montréal & Anne Imouza, McGill University

Across Québec, conversations about artificial intelligence (AI) are taking place in a perhaps surprising setting: public libraries. Through an AI literacy initiative developed by IVADO, a Québec-based interdisciplinary consortium for AI research, training, and knowledge mobilization, in collaboration with the Association des bibliothèques publiques du Québec (ABPQ), we have facilitated workshops where members of the public can discuss AI, question its promises and limitations, and explore its growing role in their lives. Yet these encounters raise a broader question: why are people turning to public libraries to make sense of AI in the first place?

Libraries as Infrastructure for a Digital Society

Beyond providing access to information and learning, public libraries foster civic engagement, social inclusion, and equity (Scott, 2011). As emerging technologies become part of everyday life, libraries increasingly help communities navigate technological change. They provide accessible, locally embedded spaces where people can ask questions, encounter different perspectives, and collectively reflect on AI's societal implications (Lo, 2025).

Based on our experiences, we believe people turn to libraries because they offer opportunities to access information and expertise outside commercial environments from AI companies or paid AI training platforms. Unlike technology companies promoting their own products, public libraries have an educational mission centred on community needs. This suggests a demand for independent learning and collective reflection beyond technical instruction or technological adoption.

What Our Workshops Reveal

Launched in Montréal in early 2025, the IVADO-ABPQ initiative has gradually expanded across the Greater Montréal area and other Québec regions. By December 2026, 62 activities are expected to have taken place, reaching an estimated 930-1240 participants. The initiative has trained 13 student facilitators and offered two webinars for library personnel, attended by 160 and 90 people respectively, extending its reach beyond public workshops. Three workshop formats address AI's technical, creative, and societal dimensions: The ABCs of AI (i.e. AI 101),

DOI: 10.5281/zenodo.23666309

AI and Democracy, and Creation or Illusion: Does AI Have Talent? The latter was developed in response to concerns from the creative sector about generative AI, encouraging young people aged 12–17 to critically explore AI-generated art without generating images themselves.

What particularly struck us was the diversity of participants' questions. Many attendees were older adults concerned about employment, automation, and personal data. Younger participants often showed greater familiarity with AI tools, raising questions about creativity and their everyday uses, while parents wanted to better understand these technologies to guide their children. Across these encounters, participants were learning how to use AI and questioning its implications for their lives and communities. Public libraries thus provided opportunities for people to collectively explore what role they want AI to play in society.

When More Falls on the Library

Our workshops also revealed the important role of library staff. Beyond hosting activities, librarians helped participants access digital tools, navigate unfamiliar platforms, and continue learning after the sessions. They act as intermediaries between communities and specialized expertise, illustrating what Martel (2025) describes as libraries' role as civic infrastructures.

Yet these responsibilities come with considerable pressure. According to ABPQ data, Québec's public libraries meet only 31% of the human-resource quality target established by BiblioQUALITÉ, with substantial additional staffing needs. Meanwhile, ABPQ has reported interest in approximately 70 AI workshops for 2027. As more people turn to libraries for guidance on technological change, library staff are being asked to do more, despite already facing staffing and resource constraints. Supporting communities in this way takes time, expertise, and sustained institutional support. Without these resources, libraries risk becoming increasingly stretched as they take on new responsibilities. Without these resources, expanding libraries' responsibilities risks placing further pressure on an already stretched workforce.

Building AI Capacity With Libraries

AI literacy cannot simply mean making AI technologies available to everyone. It must also mean ensuring that people have spaces where they can question these technologies, understand their impacts, and decide whether and how they want to engage with them. Our experience in Québec suggests that public libraries can form an important part of this infrastructure. However, this requires moving beyond one-off workshops toward sustained partnerships among universities, governments, AI organizations, and library networks. The IVADO-ABPQ initiative offers one possible model, combining public engagement with training for library personnel and opportunities for continued regional expansion.

Public libraries should not be responsible for filling every gap created by the digitization of society. But the fact that people turn to these institutions to make sense of new technologies tells us something important about the kinds of infrastructure society needs in the age of AI.

17.3 Who Gets to Shape AI? Building Coalitions for Underrepresented Communities

By Archana Prasad, Isha Singh & Tom Simmons, Gooey.AI & Royal College of Art

While ghost workers in Manila, India, Nairobi help to moderate AI, their underserved communities remain misrepresented or erased by AI systems. The people whose labour props up AI have little say in how it is built. As AI reaches an unprecedented number of people and gets enmeshed in our daily lives, the question of who gets to shape AI becomes paramount. To mitigate the risks posed by AI systems, AI literacy programs have been a response. Efforts to teach people how AI works, and how to identify bias are important but they don't really alter the underlying power capillaries.

Our work at Gooey.AI with marginalized communities, underrepresented languages and frontline workers led us to ask a different question: what if communities that mostly watch or use AI were its co-designers?" What if they could shape the datasets, fine-tune models and suggest evaluation frameworks around it? This ties with UNESCO's latest recommendations on promoting a stronger role for civil society in AI.

We responded with Beyond Bias, a participatory project in collaboration with Goethe-Institut India, with over 24 partner organisations and 9 workshops. Artists, students, and cultural practitioners used our image Trainer and Video Generator tools to fine-tune AI models, and create culturally aware outputs based on their distinct artistic styles. We moved beyond generic AI images to culturally specific ones that ranged from women protesting on the streets of New Delhi in Mughal Miniature style, intricate embroidery patterns passed down inter-generationally, and bison and horses gallivanting across caves in the style of Lascaux paintings. Beyond Bias is described in Prasad, Singh & Simmons (2026).

This shifted our approach about AI literacy: a community member would be the best person to identify the culturally inauthentic outputs by AI models that conventional benchmarks may not discern. Tanveer Hasan, Executive Director of the Centre for Internet and Society, said the workshop moved attention from improving models to how bias and representation enter through the "[dataset] feeder mechanism itself."

Community participation would not single-handedly resolve the problem of meaningful representation and homogenisation of art as the overarching infrastructure remains controlled by big tech. We analysed our Beyond Bias outputs: our cultural bias evaluation framework scored workshop outputs at an average of 0.56/1.0 for their cultural specificity.[13] This indicates

[13] To assess the cultural specificity of these outputs after fine-tuning, a cultural bias evaluation framework was applied to a sample of 20 generated image outputs, using an LLM based evaluator. It had scores from 0 to 1.0, where 0 indicated a culturally generic or homogenised output and 1.0 indicated full cultural specificity and authenticity. 0.56 indicates a moderately culturally specific output with the presence of some distinctive cultural elements.

DOI: 10.5281/zenodo.23666309

that the outputs were moderately culturally specific and it also shows that fine-tuning did not eliminate cultural bias fully. These numbers do not show that participation failed, but they do act as evidence of the constraints of participatory processes. Even with noble intentions, a participatory approach may be limited in the changes it brings about.

This led us to think towards meaningful coalitions between communities, researchers, cultural institutions and other stakeholders, for more sustained action. We also pondered what would happen if participatory approaches moved between communities that occupied different spaces. We expanded our work further to include rural women journalists in India through the AI Journalism Fellowship (AIJF) 2026.

Using tailored AI agents, the journalists built comics and news videos that captured ecological destruction in their communities, local knowledge, displacement stories and other realities that would not make it to mainstream media. One journalist summarised this critical relationship with the AI system:

"*People around me used to say don't use AI, it gets misused, data gets hacked. But after joining this fellowship, I'm confident that if we use anything in a better way, it will prove positive for us*."

Through AIJF, the question of AI literacy pivoted to: What can communities teach us about what AI should do? As researchers examine the 'jagged frontier of AI', AIJF shows us that communities already possess knowledge that AI researchers can benefit from. Our fellows were able to critically engage with AI, and question its biases, producing representations of their communities that were specific and locally relevant.

Our most recent work LokaLibrary helps communities shape AI further through an AI literacy initiative with community librarians in Karnataka. This coalition includes a cultural organisation, an AI infrastructure layer, and the Government of Karnataka. This is a work in progress that aims to unveil another aspect: *how can community institutions become sites of AI literacy rather than AI adoption?*

Our work has spanned communities differing in identity and profession, and in their views on responsible AI adoption. This led us to think deeper about coalition building instead of community participation. As anxiety about AI rises along with optimism in its potential, creating partnerships between different groups helps to build collective capacity.

An artist may question whether an AI generated image can accurately depict their style. A journalist may ask if an AI claim is factual or hallucinated. And a librarian may ask if AI information is locally relevant. What those building these coalitions and participatory programs need to ask is: *who has the authority to decide what AI considers valid*?

The UN's 2026 Global Dialogue on AI Governance also called for meaningful participation by communities to develop and govern AI. Our trajectory from Beyond Bias to AIJF and LokaLibrary has shown us that while representation matters, it alone may not be sufficient.

Building coalitions helps different communities speak to power and hold collective prowess in the shaping of AI. One of the most important artefacts that came out of our Beyond Bias process was a manifesto for inclusive AI. Its recommendations have now informed our work on eco-labeling AI on our platform.

Coalitions may be fractured and uncertain, but coalition building is imperative for communities who don't have a stake yet in determining what AI becomes. As risks surrounding AI burgeon, cultural practitioners can shape it to reflect humanity's ethos rather than its biases.

DOI: 10.5281/zenodo.23666309

17.4 Who Gets to Decide? Building Indigenous AI Governance Through Relationship, Authority and Refusal

By Keegan Leahy, Turtle Island Foundation & Michelle Okere, Indigenous Prosperity Foundation

---

Editor's Note: *Renjie Butalid,* editor of this volume, is a participant in the Indigenous AI Roundtable described here. He is not a convenor of the Roundtable and did not write this piece. He reviewed it as he did every other contribution.

Artificial intelligence is being adopted faster than most governance systems can respond. By 2026, organizational AI adoption had reached 88 percent, while industry produced more than 90 percent of notable frontier models, concentrating power among a small group of institutions. Fundamental questions about data, consent, ownership and accountability are often decided before affected communities enter the room.

For Indigenous Peoples, these questions are not new. Generations of research were conducted on Indigenous communities, with language and oral histories held by external institutions, and knowledge collected for one purpose and later used for another, without consent. Recent First Nations-led research identifies this history of extraction without consent as a continuing barrier to trust and AI adoption. AI has created an immediate need to establish protocols at a speed rarely seen before.

In 2026, we established the Indigenous AI Roundtable to bring Indigenous leaders and practitioners together with people working across AI, data sovereignty, economic development, research and governance. Among those invited were Carol Anne Hilton, CEO of Indigenomics; Elder Wendy Phillips; Shani Gwin, founder of wâsikan kisewâtisiwin; and Taylor Sparklingeyes, a data sovereignty expert. Participants connected current debates to that longer history. We began not with a framework, but with questions.

This article records themes emerging from the first sessions; it does not present positions shared by every participant or substitute for the authority of Indigenous Nations.

Four questions kept returning:

1. Who defines the problem?
2. Who authorizes the use of knowledge?

DOI: 10.5281/zenodo.23666309

3. Who has the capacity to build, procure and govern the system?
4. Who is accountable when agreed conditions are not met?

Together, they move discussion from general principles toward the institutions and decisions through which power is exercised.

From AI Literacy to Authority

Leaders, boards and organizations need to understand the systems entering their workplaces, economies and communities. But literacy alone does not confer authority. A community may understand a technology yet have little control over the problem it addresses, the data it uses or the terms of deployment. The question is: Who gets to decide?

That authority includes the ability to say yes, set conditions or refuse. Refusal is not disengagement or rejection of innovation; it is an exercise of self-determination: the authority to decide whether technology should be developed, whose knowledge may be used and whether its use aligns with a Nation's priorities, laws and responsibilities.

Participants used **Data is Land** as a discussion frame, exploring how Indigenous data, like land, remains connected to relationships and responsibilities after collection. Knowledge may be held collectively, carry intergenerational obligations and depend on who may share it, when and for what purpose. Our discussions suggested that a one-time consent form may not exhaust responsibilities attached to later uses.

Language revitalization made this concrete. AI can support transcription, translation and learning. But does permission to record Elders decades ago include permission to use those recordings to train a model today? Who owns or commercializes it? Guidance developed in 2026 by Indigenous language practitioners, engineers and researchers similarly urges communities to examine data access, ownership and risk before engaging with AI providers. These are questions of authority and relationship before they are technical questions.

Consent Requires Accountability

Participants returned to cases where consent was never sought. Public accessibility is not permission to use Indigenous stories, artwork, histories or cultural knowledge. Refusal is distinct from withdrawal: withdrawal concerns permission once granted; refusal is the right to withhold it in the first place, including for uses that are technically possible or publicly accessible.

While the Roundtable has not agreed on a single remedy, participants identified disclosure, community notification, information to assess harm, meaningful withdrawal where technically feasible and benefit-sharing when Indigenous knowledge contributes commercial value. Remedies should be shaped with affected communities, not defined solely by what technology providers can deliver.

Participation is Also Part of Sovereignty

Participants also discussed Indigenous-directed language tools, Nation-controlled data environments, Indigenous procurement and partnerships where communities negotiate development terms. They connected AI to physical infrastructure: data centres depend on land, water, electricity, cooling systems and mineral supply chains.

For some Nations, development may create ownership, employment or revenue opportunities; others may determine that the environmental and cultural costs are unacceptable. Refusal, in this context, is not a failure to participate in innovation but an expression of sovereignty and the authority to determine what happens on, to and with their lands, resources and knowledge. The common issue is whether Nations have the information, capacity and recognized authority not only to set the conditions of participation, but also to decline it.

Governance Cannot Be Extracted

This work does not begin from a blank page. For First Nations, OCAP® provides a foundational standard for ownership, control, access and possession. The CARE Principles and the United Nations Declaration on the Rights of Indigenous Peoples provide further grounding for collective benefit, authority, responsibility, ethics, self-determination and participation in decisions affecting Indigenous Peoples. The Roundtable is exploring how these frameworks apply to training data, automated inference, model ownership and infrastructure, including the authority to refuse participation, restrict access to knowledge and establish boundaries that cannot be negotiated away in the pursuit of technological advancement. These frameworks emerge from distinct communities and political struggles and cannot be reduced to an AI checklist.

Indigenous participation cannot be measured by access to technology alone, but by authority over knowledge, benefits and the conditions of participation. Those conditions may lead to partnership, stewardship or refusal. The ability to say no is as fundamental to self-determination as the ability to participate, innovate and benefit. Indigenous AI governance must recognize Nations' authority to decide not only how technology enters their communities, but whether it enters at all.

Chapter 18: AI in Government – What It Buys and Who Writes the Rules

18.1 When Government Gets Copilot, What Exactly Is It Buying?

By Anca Ciulpan, Independent Researcher

18.2 Accountable To Whom? Let the Public Write the Rules for Public Sector AI

By Tariq Khan, London Borough of Camden

18.3 AI for All, Decided by Whom? Civil Society Power Beyond Consultation

By Norman Valdez, BrainTrainr & Alex Tveit, Sustainable Impact Foundation

DOI: 10.5281/zenodo.23666309

MAIEI

18.1 When Government Gets Copilot, What Exactly Is It Buying?

By Anca Ciulpan, Independent Researcher

In 2026, U.S. federal agencies faced an unusually easy offer. Under the General Services Administration's OneGov agreement with Microsoft, agencies using Microsoft 365 G5 could add Copilot at no additional cost until their next enrollment anniversary. The offer remained available through September 2026, alongside discounts covering Microsoft 365, Azure and other services. Microsoft raised prices for several government suites on July 1; OneGov discounts continued, but new purchases used the revised list prices.

It is tempting to read this as a pricing story. It is more important as a governance story.

A free first year lowers the threshold for adoption. That is useful if the technology works. It is also consequential when the supplier already provides the email, documents, identity systems, collaboration tools and cloud infrastructure through which government employees work. Copilot does not arrive as an isolated application. It arrives inside an existing administrative environment.

The government is not simply buying software. It is deciding where AI may operate in public work: drafting, summarising, searching, analysing and preparing briefings. These are not peripheral tasks. They shape how departments form recommendations, how advice reaches decision-makers and how decisions are recorded and defended. AI may influence that chain even when a human formally makes the decision.

Federal policy recognises the risk. OMB Memorandum M-25-22 asks agencies to consider sourcing, portability and interoperability to avoid costly dependencies on a single vendor. It addresses data rights, testing, privacy and contractual protections against vendor lock-in.

There is an uncomfortable tension in the same guidance. OMB identifies exclusive discounts, tying and other arrangements that encourage spending consolidation as practices agencies should consider when evaluating vendors. Yet the same policy direction encourages faster, cost-effective AI adoption, and OneGov offers precisely that through government-wide discounts. Does an agency adopting OneGov fulfil the guidance or risk undermining it? Both interpretations are possible: the agreement can deliver immediate savings while deepening reliance on a single supplier. M-25-22 does not prohibit such agreements, but it leaves agencies to reconcile competing objectives. Without a clear assessment of long-term dependency, guidance intended to constrain vendor lock-in risks becoming a justification for the concentration it seeks to prevent.

Tokens make this governance problem more concrete.

DOI: 10.5281/zenodo.23666309

MAIEI

Microsoft's Copilot ecosystem does not rely solely on fixed per-user licences. Copilot Studio uses Copilot Credits for usage-based AI services. Some capabilities are metered through model consumption, while certain licensed employee-facing uses incur no separate credit charges. Other agents and automated workflows may consume credits or use pay-as-you-go billing.

This matters because a token is not merely a technical unit. It can become a unit of public expenditure. The cost of an AI-assisted process may depend on how much text is processed, which model is used, whether an agent retrieves organizational data, how many actions it performs and how often it runs. A procurement team can therefore know the license price without knowing the eventual operational cost of the AI-enabled workflow.

That creates new governance questions. Who can activate metered services? Who sets consumption limits? Can an agency independently verify how credits are calculated? What happens if a vendor changes the model, the credit schedule or the amount of computation needed for the same task? A system that is cheap to enter can still create an expanding, difficult-to-forecast dependency once workflows are designed around consumption.

The Government Accountability Office identified related institutional weaknesses in April 2026. Across five federal agencies, officials reported shortages of technical specialists able to define AI requirements and assess vendor proposals. GAO also identified challenges involving data rights, contract terms, testing and costs, while agencies were not systematically sharing procurement lessons.

That is where the Copilot case becomes significant. The concern is not that discounting is improper. It is that adoption can move faster than the state's capacity to understand and govern what it has adopted.

The lock-in is also organizational. Once staff learn to draft, retrieve and analyze through one assistant, workflows change, permissions are configured and managers start expecting increased efficiency. Leaving becomes more than cancelling licenses; it becomes an institutional problem.

A UK Department for Work and Pensions Copilot trial published in January 2026 illustrates both sides. Involving 3,549 staff, it estimated average time savings of 19 minutes daily and reported improvements in job satisfaction and perceived work quality. Yet users also encountered problems with search quality and consistency, and the evaluation emphasised human editing and judgement.

Good governance therefore must cover both decisions and consumption. Agencies should define which tasks AI may influence, preserve evidence for consequential uses, test systems independently, audit sensitive-data access, and retain staff capable of challenging outputs. They should also require transparent metering, budget caps, usage logs, model-change controls and an exit plan that accounts for data, workflows and skills, not just licenses.

The government should use technology that improves public administration. But “free for a year” answers only the easiest question. The harder one is whether the government can still understand, contest, budget for and replace the system after it becomes part of everyday work. If it cannot, the most important cost was never on the invoice.

DOI: 10.5281/zenodo.23666309

18.2 Accountable To Whom? Let the Public Write the Rules for Public Sector AI

By Tariq Khan, London Borough of Camden

In April 2026, the New Statesman published an investigation into how AI entered the British state. Its author, Will Dunn, spent a year interviewing ministers, officials and advisers. He describes AI systems that now help to draft legislation and shape spending decisions. His account contains no singular moment when the decision to integrate AI was made. AI arrived through many small choices, and the public was not asked about any of them.

Two other things happened in 2026. In May, the Ada Lovelace Institute published *Measuring Up*, which found little evidence that AI governance measures achieve what they set out to do. In July, the House of Commons passed the Public Office (Accountability) Bill, usually called the Hillsborough Law, and it completed committee stage in the House of Lords on 21 October, with report stage and third reading still to come.

The bill creates a duty of candour for public bodies and makes it an offence to mislead the public. Its background includes the Post Office Horizon scandal, in which an institution spent years defending its software against the people that software had harmed.

All three point to the same problem. Public bodies have AI governance, but the public has no part in it.

Who the Governance is For

In most public bodies, AI governance works like this: an officer fills in an impact assessment, a board reviews it, and an ethics framework supplies the principles. The institution writes each of these documents for its own use. The public appears in them as the subject of the assessment and nowhere else.

This matters most under pressure. Suppose a supplier offers a free pilot, or a budget gap makes automation look attractive. The officer who objects can point only to an internal policy, meaning the business case usually prevails.

Two Juries

In 2024, the Irish patient organisation IPPOSI set up a citizens' jury on AI in healthcare. An independent process picked 25 people who broadly matched the population. They sat for ten weeks, mostly online, heard from expert witnesses, and wrote recommendations for the government. One of their topics was accountability. When a clinical decision rests on AI, who answers for it?

DOI: 10.5281/zenodo.23666309

MAIEI

The second example comes from Brussels, where researchers asked a jury of residents to deliberate on an AI system that would ration electricity during shortages. Their 2026 paper describes mainstream AI ethics as top-down and expert-led, and says that it tends to exclude the people the systems affect. In the jury, residents defined the problem for themselves before they discussed any solution.

Some public bodies have gone further and made what residents wrote operational. Camden Council in London, where I work, adopted a Data Charter in 2022 that a residents' panel drafted. The Charter now sits inside the council's approval process: every data-enabled project's impact assessment is checked against its principles, and the Charter underpins the AI ethics principles that all AI proposals must meet. That check has already led to some services dropping or revising a proposed AI tool. Camden has started to carry the Charter towards supplier contracts, but the approval process, not the contract, is still where it does its work.

However, none of these efforts are binding. A jury makes recommendations, and a government can ignore them. A group of 25 cannot speak for a whole country. The Brussels authors also warn that participation can turn into a performance, where the institution holds the event and then changes nothing. That outcome costs public trust. A body that convenes a panel therefore needs to adopt the results formally and report on them afterwards.

What You Can Do In 2027

If you work in a public body, start before procurement does. Run a deliberative panel before you talk to suppliers, and ask it the questions that a business case leaves out. Which decisions should never be automated? Which ones must always come with an explanation? A panel costs money, but much less than a deployment that fails in public.

Take the panel's answers to a formal decision and publish them, with a named senior officer as the owner. Then put them into your contracts. A supplier can treat an ethics framework as your internal paperwork. A condition that the public wrote, and that the institution adopted in an open meeting, is much harder to dismiss.

Add one commitment with a date on it. Promise to tell the public when technology changes a frontline service, and bring a panel back after a year to check. No commencement date has been set for the Hillsborough Law, but when its duty of candour takes effect, that record is your evidence.

The officer who has to say no gains the most from all this. An objection that rests on personal concern is easy to overrule. An objection that rests on a rule the public wrote and the institution adopted is part of the job. In 2027, more public servants will need that backing.

18.3 AI for All, Decided by Whom? Civil Society Power Beyond Consultation

By Norman Valdez , BrainTrainr & Alex Tveit, Sustainable Impact Foundation

On October 2, Prime Minister Mark Carney announced Canada's National Council on Artificial Intelligence, mandated to advise on an 'AI for All' strategy reaching into public services, sovereignty and democracy. Yet civil society and affected communities have no designated place in its announced composition. We helped organize AI by All / L'IA par tous, launched October 13, and write as campaign participants, not neutral observers. While other contributions in this chapter examine governance from within institutions, ours asks how people outside them can exercise influence.

Who Built the Room?

The announcement identifies accomplished researchers, executives, investors and public-sector figures, but no public nomination process or selection criteria. Reporting identifies 13 members; the announcement names no chair and makes no promise to publish advice. Our 16-page membership analysis maps expertise concentrated in finance, academia, government, AI research, technology and consulting, without designated seats for civil society, labour, disability advocates, newcomers or youth. This challenges institutional design, not members' qualifications. Canada already uses sector-based representation in its separate AI and labour advisory arrangements. First Nations, Inuit and Métis governments also require distinct rights-respecting relationships, not another stakeholder seat, consistent with the UN Declaration Act.

These omissions matter because AI governance involves choices about rights and power, not only technical performance. Housing providers, unions, disability advocates and environmental groups witness different consequences of the same system. The Bots at the Gate report warned in 2018 that immigration automation raised risks to human rights and procedural fairness. Frontline knowledge belongs upstream, before advice becomes policy.

What the Campaign Has Shown

The demand predates this Council. In 2025, civil-society leaders, including us, proposed a Civil Society and Communities Council alongside the AI Strategy Task Force. The government's 30-day AI consultation received more than 11,300 responses; the People's Consultation on AI gathered further public evidence. Neither created a lasting place for civil society in national advice. Consultation gathers voices; it does not necessarily share power.

Our open letter makes four requests: broader representation with separate Indigenous participation; a standing civil-society body; published agendas, evidence, conflicts and

DOI: 10.5281/zenodo.23666309

government responses; and public assessments of major recommendations against rights, equity, labour and environmental impacts. By November 2, the campaign recorded 250 signatures. Transformations Community amplified it, and the Montreal AI Ethics Institute endorsed its core demands. Coalition partners need not agree on every AI policy. Individuals can respond quickly, while organizational endorsement requires discussion and authorization. That asymmetry is part of organizing. But endorsement, circulation and signatures do not themselves change who decides. We have identified no public federal commitment to these reforms; that says nothing definitive about private discussions, but leaves accountability invisible.

Authority Beyond Consultation

A Civil Society and Communities Council must be an institution, not a listening session. We propose open nominations through community networks, fixed terms, paid and accessible participation, and an independent secretariat. Indigenous Nations should determine their own participation processes. The body should see draft National Council advice early, place items on its agenda, commission impact assessments, and publish dissent alongside official recommendations. It must have a direct channel to responsible ministers, rather than depending on discretionary invitations.

Three safeguards would give that role force: a public mandate secured through a formal instrument and ideally legislation; a requirement for timely, reasoned government responses, identifying what was adopted, deferred or rejected; and the ability to request independent review of high-impact procurements or deployments and refer unresolved concerns to parliamentary scrutiny. These are proposed procedural powers, not a veto: elected ministers remain accountable for decisions.

What Would Change?

Before a government-wide AI contract is signed, affected communities could scrutinize privacy, accessibility, labour impacts, environmental costs and redress. Disagreement would remain visible, and governments would have to explain their choices. That would strengthen, not derail, Canada's AI ambitions. Instead of writing another open letter whenever a new advisory body appears, civil society would have an enduring route to challenge decisions before they harden.

The government's strategy is ‘AI for All.’ Our campaign is ‘AI by All’ because governing *for* people and governing *with* them are different. By 2027, the test should not be how many people were consulted, but whether communities can change what the government considers and what ministers must answer for. A seat matters; a say with institutional weight matters more.

CITATIONS & CONTRIBUTOR BIOS

INTRODUCTION

1.1 Power, Fracture, Resistance



Renjie Butalid co-founded the Montreal AI Ethics Institute (MAIEI) in 2018 with the late Abhishek Gupta, and serves as its Executive Director. He is editor of the State of AI Ethics Report (SAIER) and The AI Ethics Brief. Renjie is also Vice President Business Development at Metrika and previously led McGill University's Dobson Centre for Entrepreneurship.

Cite this Article: Butalid, R. (2026). Power, Fracture, Resistance. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.12-15. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

1.2 What Holds: Building a Shared Evidence Base



Rebecca Finlay is CEO of Partnership on AI, a global nonprofit bringing together more than 150 partners to advance responsible AI. Previously, she founded the AI & Society program at CIFAR, pioneering international collaboration on AI's societal impact. A recognized leader in AI governance, Rebecca was appointed a thematic co-lead for the UN Global Dialogue on AI Governance in 2026. An elected Fellow of the American Association for the Advancement of Science, she holds degrees from McGill University and the University of Cambridge.

Cite this Article: Finlay, R. (2026). What Holds: Building a Shared Evidence Base. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.16-18. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

1.3 Can AI Governance Keep Pace? Five Challenges and the Road Ahead



Dr. Daniel S. Schiff is an Associate Professor of Technology Policy at Purdue University and founding Co-Director of GRAIL, the Governance and Responsible AI Lab. Before academia, he served as the founding Responsible AI Lead at JP Morgan Chase & Co. and Secretary of the IEEE 7010 standard, the first industry AI ethics standard. His research on AI governance, deepfakes, public attitudes, and education has been published in venues including the *American Political Science Review, Management Science, Policy Studies Journal,* and *AAAI/ACM AIES,* and featured in outlets like *The New York Times, Washington Post, CNN, The Atlantic,* and *MIT Technology Review.*

Cite this Article: Schiff, D. S. (2026). Can AI Governance Keep Pace? Five Challenges and the Road Ahead. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.19-22. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

1.4 - What Resistance Looks Like Now



Dr. Renée Sieber researches the intersection of public engagement and computation. Some of that research is more computational (e.g., in retrieval augmented generation and natural language processing), and some focuses on the role of the public in that computation. Increasingly, her focus is on the conflicts between citizen involvement in AI governance and ethics as citizens demand a larger voice and, in some cases, fight against AI. Her latest project is on citizen resistance to and refusal of AI.

Cite this Article: Sieber, R. (2026). What Resistance Looks Like Now. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.23-24. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

PART I: PARTICIPATION, SOVEREIGNTY, & PRACTICE

Chapter 2: Participation – Clauses, Assemblies, & Assessments

2.1 The Vacuum Inside the Filler: Human Oversight in Government AI Guidance



Marcelo Haponiuk Rocha is an attorney (OAB-PR, since 1994) and PhD candidate in Bioethics at PUCPR, Curitiba. His research maps how AI governance instruments — legislation, standards, taxonomies, guidance — handle human autonomy, tracing the paradox of autonomy delegation. He holds a master's in Bioethics and an MBA in AI, completed doctoral coursework and qualifying exams in Education (PUCPR), and wrote Ética, Bioética e Segurança do Paciente (FAEL, 2022). He mentors startups on AI ethics in Vale do Pinhão, Curitiba's innovation program.



Luciana Soares Percegona, MD, is a nephrologist and the Clinical Coordinator and Technical Lead of the Renal Transplant Program at Hospital São Vicente, Curitiba, Brazil, where she has practiced since 2006. Trained in internal medicine and nephrology, with fellowships in pancreas-kidney transplantation (Hospital Beneficência Portuguesa, São Paulo) and pancreatic islet transplantation (Diabetes Research Institute, Miami), she holds a master's in Health Sciences (PUCPR). Her clinical practice centres on consent, refusal, and patient autonomy in transplantation.

Cite this Article: Rocha, M. H., & Percegona, L. S. (2026). The Vacuum Inside the Filler: Human Oversight in Government AI Guidance. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.27-29. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

2.2 - Towards a Citizens' Track on AI Governance



Tim Davies is coordinator of the Citizens' Track on AI: an initiative to give communities a powerful voice in shaping the future of artificial intelligence. He draws on a background in civic participation, open government, facilitation and field-building including work at Connected by Data, the World Wide Web Foundation, Open Data Services Co-operative and the Global Data Barometer.

Cite this Article: Davies, T. (2026). Towards a Citizens' Track on AI Governance. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.30-31. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

2.3 - Making Participation Count: A Method for Fundamental Rights Impact Assessment



Chiara Ullstein is a Research Associate at the Department of Computer Science at the Technical University of Munich. Her research explores public perception of AI and public participation in AI governance. She co-led the Global AI Dialogues, a multi-country dialogue series. She was a Friedrich Schiedel Fellow at the TUM Think Tank, and held visiting researcher positions at Nokia Bell Labs, The Tokyo University, KEIO University, Kyoto Institute of Technology, and Carleton University.

Cite this Article: Ullstein, C. (2026). Making Participation Count: A Method for Fundamental Rights Impact Assessment. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.32-34. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 3: AI Sovereignty – Compute, Clouds, & Who Gets a Seat

3.1 - AI Sovereignty in an Age of Economic Shock



Trisha Ray is an associate director and resident fellow at the Atlantic Council's GeoTech Center. She leads research projects on the intersection of geopolitics and AI, with her current work focused on sovereign AI, AI talent and capacity-building. Ray brings a global lens, rooted in work across India, the US, and emerging economies to help decisionmakers in government and industry navigate policy landscapes that are increasingly strategic, political, and interconnected.

Cite this Article: Ray, T. (2026). AI Sovereignty in an Age of Economic Shock. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.36-37. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

3.2 - The Empty Sovereignty Promises of Canada's 2026 National AI Strategy



Véronique Sioufi, PhD, (she/her) is Senior Racial Equity Researcher and Policy Analyst at BC Policy Solutions, where she leads a community-driven research desk dedicated to applying an intersectional lens to socio-economic policy. This work is guided by advocates from community organizations, unions and academia. Véronique earned a PhD in Geography and an MA in Communication from Simon Fraser University where her SSHRC-funded research focused on the ways inequality is reproduced through digital platforms, data-driven technologies and artificial intelligence systems.

Cite this Article: Sioufi, V. (2026). The Empty Sovereignty Promises of Canada's 2026 National AI Strategy. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.38-40. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

3.3 - No Seat, No Data: The Pacific and the Limits of AI Sovereignty Talk



Aydon Moala-Edwards is an Australian Tongan research, policy and technology practitioner. He founded Moala Global, a Pacific-focused research and decision-infrastructure practice working on AI governance, climate and human security, and is a Research Affiliate at the Monash Global Peace and Security Centre. He currently consults within the Australian development sector and is a member of UNSW Intelligent Systems for State and Societal Resilience.

Cite this Article: Moala-Edwards, A. (2026). No Seat, No Data: The Pacific and the Limits of AI Sovereignty Talk. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.41-42. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 4: Governance Frameworks & the Practitioners Building Them

4.1 - Ethical AI Needs a New Opening and the Sovereignty Push Might Just Be It



Claudia Del Pozo is the Founder and Executive Director of Eon Institute, a Mexican think tank working at the intersection of emerging technology, public policy, and social impact. Her work focuses on AI governance, the enabling conditions for AI transformation, the future of work and education, and the implications of emerging technologies for democracy and society. She brings a Latin American perspective to global technology debates, connecting ideas, institutions, and decisionmakers across the region and internationally to advance responsible and inclusive technological transformation.

Cite this Article: Del Pozo, C. (2026). Ethical AI Needs a New Opening and the Sovereignty Push Might Just Be It. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.44-45. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

4.2 - Practitioner-Led AI Governance Across the Global Majority



Chinasa T. Okolo, Ph.D., is the Founder & Scientific Director of Technecultura and a recent Computer Science Ph.D. graduate from Cornell University. Her research focuses on AI governance and safety for the Global Majority, datafication and algorithmic marginalization, and the geopolitical impacts of AI. Dr. Okolo has been recognized as one of the world's most influential people in AI by TIME, honored in the inaugural Forbes 30 Under 30 AI list, and advises numerous multilateral institutions, philanthropic initiatives, academic institutions, and nonprofit organizations. She has previously worked as the inaugural Policy Lead for Black in AI, a World Bank Group Africa Fellow, a Policy Specialist at the United Nations Office for Digital and Emerging Technologies (ODET), a Fellow at the Brookings Institution, and in research-based roles at Apple and Microsoft.

Cite this Article: Okolo, C. T. (2026). Practitioner-Led AI Governance Across the Global Majority. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.46-48. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

4.3 - The Mirror We Hold



Image credit:
CC BY-NC-SA 4.0
Kaii Chiang

Audrey Tang is Taiwan's cyber ambassador, first digital minister, architect of the Taiwan Model for digital diplomacy, 2025 Right Livelihood laureate and 2026-2027 Carnegie distinguished fellow. Audrey assisted in developing participatory democracy platforms such as vTaiwan and Join, as well as bringing civic innovation into Taiwan's public sector. Other achievements include shaping the country's acclaimed COVID-19 response and safeguarding its 2024 presidential and legislative elections from cyber interference. Currently focused on developing the Civic AI and 6-Pack of Care framework with colleagues at the Institute for Ethics in AI at the University of Oxford, Audrey is a changemaker, humanist and leader who invites all like-minded souls to join in waging peace and striving to free the future — together.

Cite this Article: Tang, A. (2026). The Mirror We Hold. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.49-50. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

4.4 - A Framework With No Scars Has Never Been Run



Aida Gregorian is Head of Legal Platform Research and Automation at LawVo, where she leads legal architecture and AI governance for a multi-jurisdictional legal AI platform. A practising Ontario lawyer and former General Counsel and Chief Compliance and Privacy Officer, she came to AI from insurance defence and regulated practice. She completed AI ethics training at MILA, including the TRAIL framework, and holds a post-graduate AI certification from Oxford. Her work focuses on what governance requires once it has to survive deployment across incompatible legal systems.

Cite this Article: Gregorian, A. (2026). A Framework With No Scars Has Never Been Run. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.51-53. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

PART II: SURVEILLANCE, WARFARE, & OVERSIGHT

Chapter 5: The Expanding Surveillance State

5.1 - The Warrant Nobody Could Explain



Dr. Maria Lungu is an Assistant Professor in the School of Government & International Affairs at Kennesaw State University. Her research examines the intersection of law, public administration, and artificial intelligence, with a particular focus on emerging technologies in policing and the courts. She recently completed a postdoctoral fellowship at the University of Virginia's Digital Technology for Democracy Lab. Her work has been published in international journals.

Cite this Article: Lungu, M. (2026). The Warrant Nobody Could Explain. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.56-57. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

5.2 - Guilty by Default: How Rushed Surveillance Law is Reversing the Burden of Proof



Isabelle Castro is a tech journalist and podcaster, leading the Utopia in Beta newsletter. Winner of the State Street Journalism award 2024 for her reporting on Bitcoin ETFs, her research spans blockchain technologies, privacy tech and AI, exploring emerging technologies and the digital societies shaping our world.

Cite this Article: Castro, I. (2026). Guilty by Default: How Rushed Surveillance Law is Reversing the Burden of Proof. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.58-60. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

5.3 - Disability Surveillance in the UK: AI, Care, and Power



Dr. Amy Gaeta is a Research Associate at the Leverhulme Center for the Future of Intelligence at the University of Cambridge. Gaeta's research cuts across Drone Studies, Disability Studies, Feminist STS, Visual Cultures, and Media Studies. Gaeta is interested in stories and storytelling about the affective, sensorial, and psychological dimensions of human-technology relations under surveillance and state power. They are writing Drone Life: Disability, Gender, and the Human for Minnesota UP, which explores the politics of ability and gender in civilian drone cultures.

Cite this Article: Gaeta, A. (2026). Disability Surveillance in the UK: AI, Care, and Power. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.61-63. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

5.4 - Strategic Litigation as a Transformative Tool for Advancing AI Activism in the Global South: Lessons from the Commonwealth Caribbean



Amanda Janell DeAmor Quest is a Commonwealth Caribbean lawyer and emerging AI ethics and AI governance thought leader. Amanda's work on AI ethics, human rights, and the rule of law has been featured by the AI Law - International Review of Artificial Intelligence Law, among other transnational platforms. Through her public intellectual work — which has been published by leading newspapers in eleven Caribbean jurisdictions and featured internationally — Amanda contributes to discussions on AI ethics, AI governance, and human rights.

Cite this Article: Quest, A. J. D. (2026). Strategic Litigation as a Transformative Tool for Advancing AI Activism in the Global South: Lessons from the Commonwealth Caribbean. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.64-66. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 6: Military AI Takes Centre Stage

6.1 - The Off Switch: How National Security Took Control of a Public AI Model



Asim Javed is the author of the forthcoming Future Without Consent: How AI, Power, and Profit Are Deciding What Comes Next, which examines how the AI industry's race for supremacy overrides human agency and consent. He holds an M.Sc. in Artificial Intelligence from the University of Oxford and works as an AI Methodologist at Scaled Agile, Inc. He maintains a Timeline of the AI Era at futurewithoutconsent.com and publishes a weekly newsletter: The Consent Report at theconsentreport.substack.com.

Cite this Article: Javed, A. (2026). The Off Switch: How National Security Took Control of a Public AI Model. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.68-69. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

6.2 - The Many Hands Problem: Securing Accountability in the Age of AI Warfare



Ayaz Ul-Huq Syed is a policy analyst working at the intersection of emerging technology, democracy, and international law. He holds an MSc in International Development & Humanitarian Emergencies from LSE, where he researched bottlenecks hindering regulation pertaining to the development and deployment of AI targeting modalities for lethal autonomous weapons systems.

Cite this Article: Syed, A. U. (2026). The Many Hands Problem: Securing Accountability in the Age of AI Warfare. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.70-72. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

6.3 - Hide and Seek: The Paradox of Representation in Military AI



Shimona Mohan is a Researcher with the Integrated Approaches and Security & Technology Programmes at UNIDIR, where she develops research projects and multi-stakeholder capacity-building at the intersections between gender, security and emerging technologies (particularly AI and cybersecurity). She was named one of the 100 Brilliant Women in AI Ethics for 2024 and holds a Master in International Affairs from the Geneva Graduate Institute.



Katherine Chandler is an Associate Professor and Director of Culture and Politics in the School of Foreign Service at Georgetown University. She has a Ph.D from the University of California-Berkeley and works on media and technology, including drone aircraft, artificial intelligence, data centres and PowerPoint, to understand how violence and militarism scale from the human body to global politics.

Cite this Article: Mohan, S., & Chandler, K. (2026). Hide and Seek: The Paradox of Representation in Military AI. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.73-75. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

6.4 - Towards a "Smart Mix"?: Governing AI for International Peace and Security



Branka Marijan is a Senior Researcher at Project Ploughshares, Senior Fellow at CIGI, and Lecturer at the University of Toronto's Munk School of Global Affairs and Public Policy. Her work focuses on artificial intelligence, emerging technologies, national security, and international governance, with particular expertise in military AI and autonomous weapons. She regularly advises governments and contributes to international policy discussions on responsible AI, arms control, and security. She holds a PhD from the Balsillie School of International Affairs.

Cite this Article: Marijan, B. (2026). Towards a "Smart Mix"?: Governing AI for International Peace and Security. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.76-78. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 7: Oversight Built for Another Era

7.1 - The Stadium as Testing Ground: How the World Cup Risks Exporting the Authoritarian Playbook



Taras Kovalchuk is an AI governance policy researcher with expertise in EU regulatory practice, tech policy, international law, and anti-corruption. He has previously worked with De.Fi, the International Chamber of Commerce, Transparency International, and Asters. Taras holds law degrees from institutions in Ukraine, the EU (Université Paris 1 Panthéon-Sorbonne and Leiden University), and the US (Chicago-Kent College of Law).

Cite this Article: Kovalchuk, T. (2026). The Stadium as Testing Ground: How the World Cup Risks Exporting the Authoritarian Playbook. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.80-82. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

7.2 - Canada's Lobbying Act Was Built for a World That No Longer Exists



Renee Black is the founder of GoodBot and is passionate about understanding the complex and dynamic evolution of technology and society, with the goal of advancing digital ecosystems that uphold human thriving. Renee is also the founding Executive Director of PeaceGeeks, a Canadian nonprofit working at the intersection of technology, conflict, and humanitarian response. She holds a Commerce degree from Dalhousie University, an MA in International Affairs from the University of Ottawa, and a certificate in conflict resolution from Cornell. Connect with Renee on LinkedIn.



Kaitlin Bustos is an AI Researcher and Tech Policy Researcher at the intersection of AI safety and governance. She currently serves as a Researcher at GoodBot, as well as an AI Safety Research Mentor at Sentient Futures. Her work focuses on identifying pathways toward stable, cooperative outcomes in autonomous systems whilst developing responsible regulatory frameworks. Kaitlin holds a B.Sc. in Psychology with a minor in Statistics from the University of Victoria. Connect with Kaitlin on LinkedIn.

Cite this Article: Black, R., & Bustos, K. (2026). Canada's Lobbying Act Was Built for a World That No Longer Exists. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.83-85. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

7.3 - The Algorithmic Cost of Living: How Your Personal Information Will Be Weaponized Against You



Emily Osborne is a Policy Researcher at the Canadian Shield Institute, a new think tank developing modern policy solutions for Canada, where her work focuses on digital sovereignty and the impacts of technology. She is also a fellow at the Canadian Anti-Monopoly Project. Her commentary has appeared in numerous outlets, including the Globe and Mail, Tech Policy Press and the Walrus.

Cite this Article: Osborne, E. (2026). The Algorithmic Cost of Living: How Your Personal Information Will Be Weaponized Against You. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.86-88. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

PART III: LABOUR & INFRASTRUCTURE

Chapter 8: AI at Work and the Workers Bargaining Over It

8.1 - Contract Bargaining in the Age of AI



Crystal Braunwarth, Business Representative of IATSE 891 represents motion picture technicians across British Columbia. She co-leads bargaining for the BC Council of Film Unions where she also focuses on IATSE's human-centric worker protections against AI displacement, application and contract language. As Vice-President of the BC Federation of Labour and Co-Chair of the BCFED AI Committee, she fights to ensure workers shape how AI enters their industries. She is a powerful lobbying voice in Victoria and Ottawa for BC's screen workers.



Hillary Bergshoeff is an Assistant Business Representative at IATSE 891. She supports the local's AI strategy, education, and engagement programs, advocating for a human-centric approach to AI. Through her work with the BCFED Standing Committee on AI and New Technology, Hillary is helping shape resolutions and policy recommendations to advocate for worker justice, strengthening worker power, developing bargaining language, and advocating for government regulation and oversight of AI development, implementation, and worker protections.

Cite this Article: Braunwarth, C., & Bergshoeff, H. (2026). Contract Bargaining in the Age of AI. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.91-93. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

8.2 - The Clause and the Algorithm



Shi Kang'ethe is a communications specialist and independent journalist based in Nairobi, Kenya, with over a decade in international development and digital advocacy. She writes and speaks on AI, labour and workplace culture, and serves as Regional Partner Representative for AI-Powered Women and as Cohere Labs Regional Africa Co-Lead.

Cite this Article: Kang'ethe, S. (2026). The Clause and the Algorithm. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.94-95. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

8.3 - When Workers Bring AI to Work (From Private Experiment to Organisational Capability)



Robert Schouten is a philosopher and AI governance practitioner, founder of Atlas Research and co-founder of dBSC in Sweden. He combines research on AI ethics and the changing organisation of work with hands-on organisational consulting. His background also spans healthcare, process improvement and management. He focuses on professional judgement, tacit knowledge, employee-led AI adoption and the translation of governance principles into workable processes that preserve the human capabilities needed for responsible AI use.

Cite this Article: Schouten, R. (2026). When Workers Bring AI to Work (From Private Experiment to Organisational Capability). In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.96-98. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

8.4 - When a Profession Is Automated: Translators and What Comes Next



Debbie Folaron is Full Professor of Translation Studies and Jean Monnet Chair (2020-2023) in the Département d'études françaises at Concordia University, Montréal. She is currently Principal Investigator of the Canadian Social Sciences & Humanities Research Council (SSHRC) Insight grant "Translating Romani Voices". In 2012, she co-founded the journal Translation Spaces | A multidisciplinary, multimedia, and multilingual journal of translation. Her research has focused on translation and localisation practices in the digital/AI age and in multilingual minoritized Romani contexts.



Lynne Bowker is Full Professor and Canada Research Chair in Translation, Technologies and Society at Université Laval and Adjunct Professor at the School of Continuing Studies at McGill University. She is a certified translator with the Ordre des traducteurs, terminologues et interprètes agréés du Québec (OTTIAQ) and editor of the journal *Digital Translation*. Her research focuses on the uses and impacts of translation technologies within and beyond the language industries.



Joss Moorkens is an Associate Professor at the School of Applied Language and Intercultural Studies in Dublin City University (DCU), Science Lead at the ADAPT Centre, Principal Investigator at RINN AI, Secretary of DCU's Institute for Ethics, member of the Centre for Translation and Textual Studies, and board member at the European Masters in Translation Network. He writes about translation technology interaction and translation ethics, and coedits the journal *Translation Spaces*.

Cite this Article: Folaron, D., Bowker, L., & Moorkens, J. (2026). When a Profession Is Automated: Translators and What Comes Next. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.99-101. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 9: Data Centres and Communities

9.1 - Energy, Regional Politics, & the AI Build-Out



Anne Pasek is the Canada Research Chair (tier II) of Media, Culture and the Environment, based out of both the Trent School of the Environment and the Department of Culture, Arts, and Media. Her work explores the political economy and cultural politics of the tech sector, with a focus on energy and climate issues.



Eliot Tretter is Associate Professor in the Department of Geography at the University of Calgary. He is author of Shadows of a Sunbelt City (2016), University of Georgia Press. His latest book project, titled Petrocity, explores the complex and contradictory effects of Canada's hydrocarbon extraction on Calgary's urbanization. He holds a SSHRC grant, Digitizing Carbon Capitalism, which examines how the digital economy is radically transforming work and labour in the extractive industries, especially in Alberta.

Cite this Article: Pasek, A., & Tretter, E. (2026). Energy, Regional Politics, & the AI Build-Out. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.103-104. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

9.2 - Data Centred: What Canada's Data Centre Expansion Means for Host Communities



Dr. Alexander Carlo is a Postdoctoral Researcher at the Brookfield Centre in Real Estate & Infrastructure at York University's Schulich School of Business. His research examines institutional investment in real estate, infrastructure, and other real assets. His recent work studies the expansion of Canada's data centre industry and the local employment effects of data centres in the United States. He holds a PhD in Real Asset Finance from Maastricht University.



Dr. Lyndsey Rolheiser is an Assistant Professor of Urban and Real Estate Economics at York University's Schulich School of Business, in the Brookfield Centre in Real Estate and Infrastructure. Her research examines how environmental conditions such as air pollution, extreme heat, and wildfire smoke shape the value and use of real assets with recent work focusing on the growth of North American data centres and their local economic effects. She holds a PhD in Urban Economics from the Massachusetts Institute of Technology.

Cite this Article: Carlo, A., & Rolheiser, L. (2026). Data Centred: What Canada's Data Centre Expansion Means for Host Communities. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.105-107. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

9.3 - How Communities Across Canada are Pushing Back Against Hyperscale Data Centre Development



Image credit:
Nicole Marie Burton

Janna Frenzel is a researcher and communication strategist based in Tio'tia:ke/Montréal. Her work investigates how grassroots and non-profit actors are attempting to decarbonize digital infrastructures and to reimagine communication systems through principles of climate justice and respect for planetary boundaries.

Anne Pasek is the Canada Research Chair (tier II) of Media, Culture and the Environment, based out of both the Trent School of the Environment and the Department of Culture, Arts, and Media. Her work explores the political economy and cultural politics of the tech sector, with a focus on energy and climate issues.

Sarah-Louise Ruder is a Research Associate at the University of the Fraser Valley and an Adjunct Professor at Simon Fraser University. Her forthcoming book, AI at the Table: What Automation Means for Our Food System (Ruder & Bronson, 2027), examines the labour, justice, and environmental impacts of artificial intelligence used across the food system.

Cite this Article: Frenzel, J., Pasek, A., & Ruder, S. (2026). How Communities Across Canada are Pushing Back Against Hyperscale Data Centre Development. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.108-110. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

9.4 - The $15-Billion Text Message: Holding AI Infrastructure to Account



Priscila Chaves Martínez is an AI strategist with nearly 20 years of experience across industry, research and public policy in all seven continents. She has held AI strategy and R&D roles at global corporations and advised governments on their national AI strategies. She holds an MBA from NYU Stern and a Master of Studies in AI Ethics and Society from the University of Cambridge, and is completing Postgraduate studies in Artificial Intelligence for Business at Oxford's Saïd Business School.

Cite this Article: Chaves Martínez, P. (2026). The $15-Billion Text Message: Holding AI Infrastructure to Account. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.111-113. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

PART IV: CARE, CLASSROOMS, & CREATIVE WORK

Chapter 10: Simulated Care – Triage, Companions, & Robots

10.1 - Invisible by Design: Disease, Illness, and Sickness in Algorithmic Triage



Rosa E. Martín-Peña leads the Ethics of Medical AI group as a postdoc at CELLS and the Lower Saxony Centre for AI and Causal Methods in Medicine, Leibniz University Hannover. In interdisciplinary collaboration with clinicians, bioinformaticians and data scientists, among others, she works on data ethics frameworks to detect epistemic injustice in clinical AI. Her interest in causal inference comes from what is at stake in medicine: saving lives. She researches what constitutes a medical digital twin and its ethical implications.

Cite this Article: Martín-Peña, R. E. (2026). Invisible by Design: Disease, Illness, and Sickness in Algorithmic Triage. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.116-117. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

10.2 - A Case for Human Connection as a Key Design Principle for Social AI Companions



Connor Wright is Director of Partnerships at the Montreal AI Ethics Institute, holding a first-class masters in AI Ethics & Society from the University of Cambridge, and a first-class undergraduate degree in Philosophy with advanced proficiency in Spanish from the University of Exeter. He has published works on trustworthy AI, AI hype, and griefbots, and has consulted on various AI governance projects. His main passion lies in the form of social AI, as well as the cross-section between the Sub-Saharan African philosophy of Ubuntu and AI, stemming from his upbringing in South Africa.

Cite this Article: Wright, C. (2026). A Case for Human Connection as a Key Design Principle for Social AI Companions. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.118-120. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

10.3 - Embodiment and Affective Harm: Social Robotics in the Era of Foundation Models



Alva Markelius is a PhD candidate at the Affective Intelligence and Robotics Laboratory, Department of Computer Science & Technology, University of Cambridge, and recipient of the Top 100 Brilliant Women in AI Ethics™ award 2024. Her research explores the ethics of social AI and robotics — particularly at the intersection of embodied robotics and foundation models. She has a background in Cognitive Science and holds an MSt in AI Ethics & Society from the Leverhulme Centre for the Future of Intelligence, University of Cambridge.

Cite this Article: Markelius, A. (2026). Embodiment and Affective Harm: Social Robotics in the Era of Foundation Models. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.121-123. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 11: AI in Education – Detection, Fatigue, & Curiosity

11.1 - More than 'AI-Ready': What Students Want from AI Literacy in Higher Education



Katie Ledingham is a Senior Lecturer in Responsible and Transformative Innovation at the University of Exeter Business School, where she is also Education Lead at the Centre for Responsible Innovation. She co-convenes the final-year module BEM3090, Rethinking Innovation and Entrepreneurship. Her work spans research, teaching and executive education, with a focus on responsible AI, AI literacy and transformative innovation. She explores how organisations, and innovation processes more broadly, can help shape more responsible and sustainable futures.



Rob Pearce is a Lecturer in Innovation and Entrepreneurship at the University of Exeter Business School. He is Co-Director of the With Proficiency in Innovation and Entrepreneurship pathway and co-convenes BEM3090, Rethinking Innovation and Entrepreneurship. His research and teaching focus on pedagogical innovation, transformative learning and futures-oriented education. He is particularly interested in how education can help students navigate technological change with a clear sense of human purpose.

Cite this Article: Ledingham, K., & Pearce, R. (2026). More than 'AI-Ready': What Students Want from AI Literacy in Higher Education. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.125-127. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

11.2 - Fatigue and Overwhelm: Institutional and Instructor Challenges in an Age of AI



Heather von Stackelberg (M.A.) is an experienced writer, instructor and curriculum developer, currently teaching Ethics (among other things) in the Machine Learning Analyst program at NorQuest College Edmonton. Before this role she was an educator at Amii (the Alberta Machine Intelligence Institute) where she was a vital contributor to their educational content on ethics and other social impacts of AI. Heather was also a member of the inaugural Learning Cohort at MAIEI.



Working to weave together industry, community, and student experience in meaningful ways, **Caylee Kreller (M.A., TESL, M.Ed.)** is a seasoned educator whose work is grounded in a commitment to equity and community connection — values that shape everything from her classroom practice to her research. She is a member of the University of Alberta Research Ethics Board, is published in the British Journal of Guidance and Counselling, and co-leads the Generative AI SoTL Research team — bringing a thoughtful, human-centered lens to how AI is integrated in higher education.

Cite this Article: von Stackelberg, H., & Kreller, C. (2026). Fatigue and Overwhelm: Institutional and Instructor Challenges in an Age of AI. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.128-130. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

11.3 - Starting from Curiosity: Encouraging Youth AI Literacy through Self-Directed Research



Miranda Pate is an MA student in the Digital Humanities program at McGill University. She is the Director of Kodera's AI Youth Research Fellowship, a program that promotes AI literacy in high school and CEGEP students by allowing them to research an AI topic of their choice with a graduate student mentor.



Sophie Wu is a PhD student in Computer Science at Columbia University. She previously graduated from the Digital Humanities MA at McGill University, and during this time founded the Kodera AI Youth Research Fellowship.



Chelsea Sun is a third-year undergraduate student in sociology and economics at McGill University. She is the current Co-Director of Publications at Kodera and is working to uplift different perspectives within AI justice.

Cite this Article: Pate, M., Wu, S., & Sun, C. (2026). Starting from Curiosity: Encouraging Youth AI Literacy through Self-Directed Research. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.131-132. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 12: AI and the Arts – Authorship, Credit, & Voice

12.1 - Creative Agency in the Age of Generative AI



Charles Migos is the founder of Intangible.ai, a platform pioneering the use of spatial intelligence and camera-based workflows as a control layer for generative AI, to accelerate human creative decisions across film, television, games, advertising, and live experiences. A designer, technologist, and inventor with more than 180 patents in technology and systems design, Charles brings three decades of experience, including leadership roles at Apple, Sony, Microsoft, and Unity.

Cite this Article:

Migos, C. (2026). Creative Agency in the Age of Generative AI. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.134-135. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

12.2 - Unearned and Undisclosed: The Absence of AI Verification Standards



Kei Baritugo is a narrative and movement builder. She is the Director of Global Marketing Communications at the Montreal AI Ethics Institute. She is also the founder of BoldLove, a purpose-driven PR agency. With over 20 years of experience in strategic global campaigns and experiential event production, she has worked with award-winning changemakers and held executive leadership positions in arts and entertainment, technology and social impact.

Cite this Article: Baritugo, K. (2026). Unearned and Undisclosed: The Absence of AI Verification Standards. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.136-138. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

12.3 - The Causes of the Causes



Amanda Silvera is a Canadian voice actor and the founder and CEO of the Society for Original Biometric Identity Rights (SOBIR). Her work focuses on consent, creative rights and biometric identity protection in AI. In 2026 SOBIR published A Standard of Consent for the Use of Biometric Identity and its companion, A Framework of Practice for the Use of Biometric Identity.

Cite this Article: Silvera, A. (2026). The Causes of the Causes. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.139-140. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

PART V: AGENTS, BODIES, & OPEN MODELS

Chapter 13: Agentic AI – Accountability at Speed

13.1 - What My Own Audit Trail Got Wrong



Roshan Ghadamian is a systems architect working on governance infrastructure for agentic AI and philanthropic capital. He is Principal Researcher at IRSA and a Visiting Fellow at Ormond College, University of Melbourne. His current work instruments pre-action authorisation gates in production agent deployments, publishing the resulting decision traces as open datasets. He writes on institutional design for emerging sectors and has published widely on capital architecture, governance, and the limits of oversight mechanisms.

Cite this Article: Ghadamian, R. (2026). What My Own Audit Trail Got Wrong. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.143-144. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

13.2 - Agentic AI: Self-Regulation Is the Regulation – The Semantic and Reasoning-Layer Governance Gap Across Institutional Boundaries



Maureen Doyle-Spare is an enterprise governance architect and AI governance researcher specializing in semantic layer governance, runtime governance of the reasoning layer, and cybersecurity for autonomous and multi-agent systems. She is the originator of the Agentic 3 C's Framework and Semantic Control Plane runtime architecture, and a pioneer in reasoning-layer governance. Her research examines how institutions preserve authorized meaning as agents form Operational Interpretations that governance controls evaluate before execution authority is emitted across enterprise operating environments at scale.

Cite this Article: Doyle-Spare, M. (2026). Agentic AI: Self-Regulation Is the Regulation – The Semantic and Reasoning-Layer Governance Gap Across Institutional Boundaries. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.145-148. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

13.3 - Automating the Grounds, Not the Decision: What Care Records Should Have to Say



Jae-Seong Lee writes safety requirements for AI systems as regulatory subgroup lead of IEEE P2933.1, and studies humanoid robot governance at ETRI, Korea's national ICT research institute. He works at both ends of the standards process, from technical committees to citizen forums. He sits on the Legal Committee of the AI Ethics and Integrity International Association and the Early Career Review Board of the Journal of Information, Communication and Ethics in Society, and wants limits settled before more machines arrive.

Cite this Article: Lee, J. (2026). Automating the Grounds, Not the Decision: What Care Records Should Have to Say. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.149-150. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

13.4 - Automation and Bureaucracy: Principals, Agents, and Accountability in Public Administration



Dr. Jennifer Raso is a socio-legal-technical scholar who investigates relations between discretion, algorithmic systems, and administrative law. She is particularly intrigued by how institutional decisions are co-produced and the implications for procedural fairness and substantive justice. Dr. Raso explores these issues primarily in the context of social benefits and immigration. Her work is funded by SSHRC, the Fonds de Recherche du Québec, and is the subject of numerous articles and a book with Routledge's New Trajectories in Law series.

Cite this Article: Raso, J. (2026). Automation and Bureaucracy: Principals, Agents, and Accountability in Public Administration. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.151-153. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 14: AI and the Body – Biometrics, Health Data, & Bodily Autonomy

14.1 - Confessions in the Cloud: How 140 Million Therapy Exchanges Became Training Data



Annie Gilbertson leads Proof News' projects and investigations. Annie was part of a pair of reporters that uncovered the major tech companies using material from unwitting YouTube creators to train AI models. At The Markup, she was part of a team that won a national Murrow Award for exposing racial bias in predictive policing software. Her work has appeared in WIRED, The Washington Post, The Wall Street Journal and NPR, among other outlets. She teaches journalism at UC Berkeley.

Cite this Article: Gilbertson, A. (2026). Confessions in the Cloud: How 140 Million Therapy Exchanges Became Training Data. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.155-157. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

14.2 - From Gender Affirming Care to Reproductive Rights: The Surveillance Risk of Bodily Data



Jess Reia (they/them) is an Assistant Professor of Data Science & Public Policy and a 2025-2027 Andrew Carnegie Fellow at the University of Virginia, where they co-lead the Digital Technology for Democracy Lab. Reia works primarily on technology policy, human rights, and urban governance. A policymaker by training, they have collaborated with governments for over a decade and conducted research that has been published in four languages

Cite this Article: Reia, J. (2026). From Gender Affirming Care to Reproductive Rights: The Surveillance Risk of Bodily Data. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.158-159. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

14.3 - The Biometric Dragnet



Emile Dirks, PhD, is a senior research associate at the Citizen Lab at the Munk School of Global Affairs & Public Policy, University of Toronto, and chair of the Digital Threats Committee at PEN Canada.



Moses Karanja, PhD, is a post-doctoral fellow at the Citizen Lab. Together, they research the rights implications of biometric and digital identification systems around the world.

Cite this Article: Dirks, E., & Karanja, M. (2026). The Biometric Dragnet. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.160-162. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 15: Open Models & the Communities Building With Them

15.1 - The People's AI Stack: Mapping the Social and Political Choices Behind Artificial Intelligence



Stephanie Dick is an Assistant Professor in the School of Communication at Simon Fraser University and a historian of computing, artificial intelligence, and mathematics. Her research examines how AI systems reshape knowledge, work, institutions, and power. Her research examines the historical and contemporary relationships among artificial intelligence, labour, knowledge, and governance, with particular attention to how computational systems redistribute authority, reshape institutions, and transform the organization of physical and cognitive labour.



Fergus Linley-Mota is Founder and Director of the Dialogue on Technology Project at Simon Fraser University's Morris J. Wosk Centre for Dialogue. His work focuses on expanding the range of voices shaping decisions about AI and related technologies, and on understanding how computational technologies intersect with democratic life. He was recently a Senior Fellow at McGill's Centre for Media, Technology and Democracy, and a Fellow with Action Canada and the International Strategy Forum.



Hillary Bergshoeff is an Assistant Business Representative at IATSE 891. She supports the local's AI strategy, education, and engagement programs; advocating for a human-centric approach to AI. Through her work with the BCFED Standing Committee on AI and New Technology, Hillary is helping shape resolutions and policy recommendations to advocate for worker justice, strengthening worker power, developing bargaining language, and advocating for government regulation and oversight of AI development, implementation, and worker protections.

Cite this Article: Dick, S., Linley-Mota, F., & Bergshoeff, H. (2026). The People's AI Stack: Mapping the Social and Political Choices Behind Artificial Intelligence. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.164-166. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

15.2 - Re-Remix: Open Source Models for the World



Prof. Vukosi Marivate is Director of the African Institute of Data Science and Artificial Intelligence (AfriDSAI), ABSA UP Chair of Data Science and Professor of Computer Science at the University of Pretoria, where he leads the Data Science for Social Impact group. His research focuses on Machine Learning (ML), Artificial Intelligence (AI), and Natural Language Processing (NLP), particularly for African and other low-resource languages. He co-founded Lelapa AI, the Masakhane Research Foundation, and the Deep Learning Indaba. His work spans social challenges in science, energy, public safety, and utilities, aiming to create AI for Africans by Africans. Vukosi is a member of the Independent Scientific Panel for AI for the UN and a member of the African AI Council at Smart Africa.



Emma Ruttkamp-Bloem is the UNESCO Chair in AI for Community Resilience, Head of Philosophy, and a faculty fellow at the African Institute for Data Science and AI at the University of Pretoria. She leads the AI ethics group at the Centre for AI Research (CAIR). She is the Chairperson of the UNESCO World Commission on the Ethics of Science and Technology (COMEST), and a member of UNESCO's AI Ethics Experts without Borders Network and the WEF's Global Future Council on Autonomous Systems.



Chijioke Okorie is Associate Professor of IP and innovation law in the Department of Private Law at the University of Pretoria, where she leads Data Science Law Lab, a research group that deploys research in law and produces evidence and policy advice to support the growth of data science research across Africa. She is a Senior Fellow of the African Institute for Data Science and AI, Associate Editor of South African Intellectual Property Law Journal and the author of several articles on IP and information justice issues in Africa. More information about Chijioke can be found at www.chijiokeo.com

Cite this Article: Marivate, V., Ruttkamp-Bloem, E., & Okorie, C. (2026). Re-Remix: Open Source Models for the World. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.167-168. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

15.3 - ParcoursLab: A Course Recommender Made by Students, for Students



Riccardo Di Sipio is a Senior Applied AI Research Scientist at ServiceNow in Toronto. A former experimental particle physicist (CERN LHC, top quark research), he moved into natural language processing and spent years building machine learning systems for HR tech, most recently at Dayforce. He also serves as Strategic Projects Lead at the Human Feedback Foundation, where he mentors student teams on applied AI projects. He has also published academic research on applied quantum computing.



Elena Yunusov works to keep humans in the loop as AI reshapes society. She is Executive Director of the Human Feedback Foundation (HFF), a Canadian AI nonprofit and Linux Foundation AI & Data member. HFF's flagship initiative, RAISE (Responsible AI for Social Impact), is Canada's first national nonprofit AI adoption program, delivered with Creative Destruction Lab and The Dais. HFF also leads AI Tinkerers Toronto, Canada's largest AI/ML practitioners community, and the open-source AI in Production mentorship program.



Aaryan Shroff is a Software Engineer at Ramp. He previously interned at Stripe, NVIDIA, and Cohere, and holds a Bachelor's degree in Computer Science from the University of Waterloo. At Waterloo, he served as Design Lead and Workshop Lead for the UW Data Science Club, where he led the ParcoursLab team of developers.



Claudel Rheault is a researcher focused on data labelling, participatory AI and human-AI interaction. Currently completing an MSc in Data, Inequality and Society, she brings a hybrid background in communication studies, interaction design and sociology. Over eight years, she has helped build ML-powered products for cybersecurity and insurance, and scaled human-centred approaches to AI development. She most recently worked in data labelling and model validation, co-creating tools with data workers and exploring new approaches to data work.

Cite this Article: Di Sipio, R., Yunusov, E., Shroff, A., & Rheault, C. (2026). ParcoursLab: A Course Recommender Made by Students, for Students. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.169-171. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

PART VI: LITERACY, COALITIONS, & GOVERNMENT

Chapter 16: AI Literacy Across the Lifespan

16.1 - AI Literacy in K-12: To Whose End?



Mia Travers-Hayward is a researcher and policy analyst at the Canadian Teachers' Federation. She leads a national AI strategy to protect human-centred education and hold tech companies accountable for harms in public schools.

Cite this Article: Travers-Hayward, M. (2026). AI Literacy in K-12: To Whose End? In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.174-175. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

16.2 - Prioritizing Responsible AI Education for Older Adults



Tina Crouse works in AI Ethics & Strategy. She is a ResponsibleAI Practitioner and Management Consultant supporting nonprofits and small business owners. She has been a long standing DEI advocate and is the AI Practice Coach supporting seniors in AI literacy. Tina is the creator of the Grant Gauge™, Canada's 1st unbiased grant-making software. She is a member of Women Defining AI; Global Council for Responsible AI; Women in AI Canada; The AI Collective; Ottawa ResponsibleAI Hub.

Cite this Article: Crouse, T. (2026). Prioritizing Responsible AI Education for Older Adults. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.176-178. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

16.3 - What Nine Years of STEM Outreach Data Can Teach Us About Designing AI Literacy



Sara Jameel holds a degree in Educational Technology and an undergraduate degree in Engineering. She has designed and delivered an AI ethics and literacy workshop within a university gender-inclusive STEM outreach program for two years, and is co-author of published research papers at ASEE and CEEA-ACÉG conferences evaluating that program's outcomes across eight cohorts.



Dr. Bahareh Goodarzi is a Senior Lecturer in the Department of Electrical and Computer Engineering at Concordia University, Canada. She holds a Ph.D. in Computer Science, and her interests include artificial intelligence, engineering education, and inclusive STEM learning. She is a principal organizer of GirlSET, a STEM summer camp that encourages girls aged 14–17 to pursue engineering through hands-on activities. She has co-authored peer-reviewed publications on GirlSET presented at the ASEE and CEEA-ACÉG conferences, focusing on broadening participation and promoting equity, diversity, and inclusion in engineering education.



Dr. Navid Sharifi is a lecturer in Mechanical, Industrial and Aerospace Engineering at Concordia University. His work focuses on engineering education, inclusive teaching, and student engagement. He co-leads GirlSET and has contributed to research on its design and outcomes. He has received the Gina Cody School awards for Sustained Excellence in Teaching and for Contribution to Equity, Diversity, and Inclusion.

Cite this Article: Jameel, S., Goodarzi, B., & Sharifi, N. (2026). What Nine Years of STEM Outreach Data Can Teach Us About Designing AI Literacy. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.179-180. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 17: Civil Society – Austerity, Libraries, Coalitions, & Refusal

17.1 - Optimized for Whom? AI, Austerity, and Ontario's Settlement Sector



Annabelle Bernard is a community researcher and Master of Community and Regional Planning student at the University of British Columbia. She brings 15 years of experience in the nonprofit sector between Toronto and Vancouver, Canada. Her current qualitative research examines the ethical tensions facing equity-based nonprofits as they consider and adopt generative AI. She is based in Vancouver.

Cite this Article: Bernard, A. (2026). Optimized for Whom? AI, Austerity, and Ontario's Settlement Sector. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.182-184. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

17.2 - When AI Comes to the Library: Public Libraries as Civil Society Infrastructure Under Pressure in the Age of AI – Lessons from a Québec Community Initiative



Lesly Nzeusseu Kouamou is a doctoral candidate in Organizational Psychology at the Université de Montréal. Her work explores the responsible adoption of AI in organizations and society, with a particular focus on AI ethics, human-centered AI, workplace well-being, and public engagement. Lesly is an IVADO Scholar affiliated with Research Regroupement 5 (R5 - Ethics, EDI, and Indigenous Engagement), which focuses on studying the ethical and societal dimensions and implications of AI.



Jade Maria Moisan is a SSHRC scholar and completing a Master's degree in library and information science at the Université de Montréal's École de bibliothéconomie et des sciences de l'information (EBSI). She is also a member of the UNESCO Chair on Open Science and the Quebec research chair on the discoverability of French scientific content, contributing to research and communications. Her research focuses on inequalities in science, particularly accessibility, gender parity, and language. As an AI mediator with IVADO, she leads the deployment of its program in the Eastern townships.



Eddy Damaris Nono Defo holds a Master's degree in Environmental Science and Management from the University of Liège. He is currently pursuing a Master's degree in Environment and Sustainable Development at the Université de Montréal. He also works as a research assistant at the UNESCO Chair in Urban Landscape. As part of his master's research, he is interested in citizen participation in AI integration policies for mobility. More broadly, his research interests focus on the societal impacts of AI. Eddy is also a scholarship recipient of IVADO's Regroupement 5.



Anne Imouza is a PhD Candidate at McGill University. She is a member of the Center for the Study of Democratic Citizenship and Mila. Her research lies at the intersection of public opinion and artificial intelligence, where she examines trust and support for AI in public services among the Canadian population and across diverse professional domains, such as in journalism.

Cite this Article: Nzeusseu Kouamou, L., Nono Defo, E. D., Moisan, J. M., & Imouza, A. (2026). When AI Comes to the Library: Public Libraries as Civil Society Infrastructure Under Pressure in the Age of AI – Lessons from a Québec Community Initiative. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.185-186. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

17.3 - Who Gets to Shape AI? Building Coalitions for Underrepresented Communities



Archana Prasad is an artist, technologist, and cultural entrepreneur, and Founder & Chief Creative Officer of Gooey.AI. Her practice sits at the intersection of participatory design, responsible AI, culture, and community-led innovation. A PhD researcher at the Royal College of Art, she explores how creative and cultural communities can shape more equitable AI systems. She previously founded Jaaga and BeFantastic and worked as a design researcher at Microsoft Research India.



Dr. Isha Singh is a Communications Manager at Gooey.AI, working across research, communications, and participatory AI initiatives. She holds a doctorate in English Literature from Jawaharlal Nehru University, New Delhi. She has also authored a book of short stories with Harper Collins India.



Professor Tom Simmons is a sound artist, researcher, and Professor of Digital Arts at the Royal College of Art, where he is Head of Programme for Digital Direction. His research explores the politics and aesthetics of sound and storytelling in relation to communities, ecologies, and emerging technologies. He has led and contributed to interdisciplinary research spanning AI narratives, ecological citizenship, craft, digital storytelling, and collective learning.

Cite this Article: Prasad, A., Singh, I., & Simmons, T. (2026). Who Gets to Shape AI? Building Coalitions for Underrepresented Communities. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.187-189. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

17.4 - Who Gets to Decide? Building Indigenous AI Governance Through Relationship, Authority and Refusal



Keegan Leahy is Founder and CEO of the Turtle Island Foundation and an Anishinaabe finance executive who has spent three decades working across global Indigenous economic development, technology, energy, and capital markets. A U.S. Air Force combat veteran, his work focuses on building Indigenous-led institutions and advancing Indigenous sovereignty in emerging systems, including artificial intelligence, data governance, and finance. He holds an MBA and MSc in Disruptive Innovation from Hult International Business School and has helped launch more than 300 companies and projects over his tenure.



Michelle Okere is a Métis-Cree executive, entrepreneur, and advisor working at the intersection of Indigenous economic prosperity, technology, and governance. She is Executive Director of the Indigenous Prosperity Foundation and co-founder of the REET Institute. She co-leads a national Indigenous AI Roundtable bringing together leaders working across AI, data sovereignty, governance, research, and economic development. Michelle holds an Executive MBA in Indigenous Business Leadership from Simon Fraser University and brings a systems-focused lens to advancing Indigenous participation in emerging technologies.

Cite this Article: Leahy, K., & Okere, M. (2026). Who Gets to Decide? Building Indigenous AI Governance Through Relationship, Authority and Refusal. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.190-192. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

Chapter 18: AI in Government – What It Buys and Who Writes the Rules

18.1 - When Government Gets Copilot, What Exactly Is It Buying?



Anca Ciulpan is a senior governance and transformation professional with over 12 years' experience delivering complex regulatory, risk and organisational change across financial services, government and technology. Her work focuses on turning regulatory and governance requirements into practical frameworks, controls and measurable outcomes. She is particularly interested in responsible AI, regulatory accountability and how organisations can translate emerging AI requirements into effective governance and decision-making.

Cite this Article: Ciulpan, A. (2026). When Government Gets Copilot, What Exactly Is It Buying? In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.194-196. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

18.2 - Accountable To Whom? Let the Public Write the Rules for Public Sector AI



Tariq Khan is Chief Information Officer at the London Borough of Camden, where he leads technology, data and AI strategy. He writes here in a personal capacity. He holds a Master's in AI and ethics from the University of Cambridge.

Cite this Article: Khan, T. (2026). Accountable To Whom? Let the Public Write the Rules for Public Sector AI. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.197-198. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

18.3 - AI for All, Decided by Whom? Civil Society Power Beyond Consultation



Norman Valdez is a Peruvian-Canadian AI strategist, educator, and advocate advancing responsible innovation for the public good. As CEO and Co-Founder of BrainTrainr and Co-Founder and Co-Lead of Alianza Latina, he builds AI literacy, leadership, and capacity in historically excluded communities. An Emerging Technology Fellow at Community Foundations of Canada and member of the Center for AI and Digital Policy's AI Policy Group, Norman brings his Indigenous Aymara perspective to technology governance, championing equity, democratic participation, and meaningful collective action.



Alex Tveit (CA) is a systems-focused leader advancing community empowerment and inclusive innovation. He is Co-Founder and CEO of the Sustainable Impact Foundation and an Emerging Technology Fellow at Community Foundations of Canada. He leads the AI Affinity Group at Philanthropic Foundations Canada, co-stewards the Canadian Network for Complex Systems, and serves on boards spanning climate, democratic deliberation, and healthcare. His work pairs systems thinking with community-centered practice to shift power and make technology serve society.

Cite this Article: Valdez, N., & Tveit, A. (2026). AI for All, Decided by Whom? Civil Society Power Beyond Consultation. In R. Butalid & C. Wright (Eds.), *The State of AI Ethics Report (Volume 8) Power, Fracture, Resistance.* pp.199-200. Montreal AI Ethics Institute (MAIEI). DOI: 10.5281/zenodo.23666309. Available at: https://montrealethics.ai/state.

MONTREAL AI ETHICS INSTITUTE (MAIEI) TEAM



Renjie Butalid is Executive Director and Co-founder of the Montreal AI Ethics Institute (MAIEI), an international non-profit dedicated to democratizing AI ethics literacy, which he established in 2018 with the late Abhishek Gupta. MAIEI publishes The AI Ethics Brief and the State of AI Ethics Report (SAIER), and is a member of the AI Alliance, the LF AI & Data Foundation, and Partnership on AI. In 2024, MAIEI was a founding member of the U.S. AI Safety Institute Consortium (AISIC) at NIST, now renamed to the NIST Artificial Intelligence Consortium.

His work on the SAIER is as much curation as editing. Each volume brings together researchers, practitioners, organizers, public servants and workers who would rarely appear in the same publication, drawn from Africa, Asia, Europe, the Pacific and the Americas. The most consequential questions about AI are not technical ones, and answering them takes more than one discipline. Published annually under a Creative Commons licence, each volume is also a record of its year: what was decided, what it cost, and who was in the room when it happened.

Renjie is also Vice President, Business Development at Metrika, a digital asset risk management platform, where he leads partnerships with regulated financial institutions including global systemically important banks, asset managers, and rating agencies. He previously led the expansion of McGill University's Dobson Centre for Entrepreneurship. He holds a BA in Economics from the University of Waterloo and an MA in International Relations from Corvinus University of Budapest.



Connor Wright is Director of Partnerships at the Montreal AI Ethics Institute, holding a first-class masters in AI Ethics & Society from the University of Cambridge, and a first-class undergraduate degree in Philosophy with advanced proficiency in Spanish from the University of Exeter. He has published works on trustworthy AI, AI hype, and griefbots, and has consulted on various AI governance projects. His main passion lies in the form of social AI, as well as the cross-section between the Sub-Saharan African philosophy of Ubuntu and AI, stemming from his upbringing in South Africa.



Kei Baritugo is a narrative and movement builder. She is the Director of Global Marketing Communications at the Montreal AI Ethics Institute. She is also the founder of BoldLove, a purpose-driven PR agency. With over 20 years of experience in strategic global campaigns and experiential event production, she has worked with award-winning changemakers and held executive leadership positions in arts and entertainment, technology and social impact.



Keilah Abello is the Program and Operations Coordinator at the Montreal AI Ethics Institute. She is also a multidisciplinary AI literacy specialist, holding undergraduate degrees in computer science and sociology from the University of British Columbia. Previously, she has helped develop AI literacy educational resources for academic institutions, supported AI model training, and has worked with youth-focused non-profits. Passionate about the social impact of emerging technologies, she is particularly interested in making technology accessible, inclusive, and human-centered.

