



# Advancing Responsible AI in Public Sector Application

GPAI Edition

Edited by Abhishek Singh  
and Balaraman Ravindran



CRC Press  
Taylor & Francis Group



# Advancing Responsible AI in Public Sector Application

Responsible use of AI in public sector applications requires engagement with various technical and non-technical areas such as human rights, inclusion, diversity, innovation and economic growth. The book covers topics spanning the technological socio-economic spectrum, including the potential of AI/ML technologies to address social and political inequities, privacy-enhancing technologies for datasets, friction-less data sharing and data stewardship models, regional/geographical inequities in extraction and so forth.

Features:

- Focuses on technical aspects of responsible AI in the public sector.
- Covers a wide range of topics spanning the technological socio-economic spectrum.
- Presents viewpoints from public sector agencies as well as from practitioners.
- Discusses privacy-enhancing technologies for collecting, processing and storing datasets, and friction.
- Reviews frameworks to identify and address biased AI outcomes in the design, development and use of AI.

This book is aimed at professionals, researchers and students in artificial intelligence, computer science and engineering, policy-makers, social scientists, economists and lawyers.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

# Advancing Responsible AI in Public Sector Application

GPAI Edition

Edited by Balaraman Ravindran  
and Abhishek Singh



**CRC Press**

Taylor & Francis Group

Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group, an **informa** business

First edition published 2026  
by CRC Press  
2385 NW Executive Center Drive, Suite 320, Boca Raton FL 33431

and by CRC Press  
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

*CRC Press is an imprint of Taylor & Francis Group, LLC*

© 2026 selection and editorial matter, Abhishek Singh and Balaraman Ravindran;  
individual chapters, the contributors

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, access [www.copyright.com](http://www.copyright.com) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC please contact [mpkbookspermissions@tandf.co.uk](mailto:mpkbookspermissions@tandf.co.uk)

*Trademark notice:* Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

ISBN: 978-1-032-70393-0 (hbk)  
ISBN: 978-1-041-12201-2 (pbk)  
ISBN: 978-1-003-66357-7 (ebk)

DOI: 10.1201/9781003663577

Typeset in Times  
by Apex CoVantage, LLC

---

# Contents

Preface..... vii

About the Editors ..... ix

List of Contributors..... xi

**Chapter 1** Raising a Global Standard in the Procurement of Artificial Intelligence and Automated Decision Systems ..... 1

*Gisele Waters and Cari Miller*

**Chapter 2** Data Empowerment and Protection Architecture (DEPA) for Training Machine Learning (ML) Models ..... 35

*Shyam Sundaram, Kapil Vaswani, Gaurav Agarwal, Sunu Engineer, and AVS Sridhar*

**Chapter 3** Generative AI Governance: Technological Monoculture, Market Structure and the Risk of Correlated Failures..... 48

*Ramayya Krishnan, Prasanna Parasurama, Joao Sedoc, and Arun Sundararajan*

**Chapter 4** Empowering Citizens through Responsible AI Governance: Policy Recommendations for Public Algorithm Registers .....51

*Jens Meijen and Niharika Gujela*

**Chapter 5** Responsible Adoption of Cloud-Based Artificial Intelligence in Health Care: A Validation Case Study of Multiple Artificial Intelligence Algorithms for Diabetic Retinopathy Screening in Public Health Settings..... 62

*Mona Duggal, Anshul Chauhan, Ankita Kankaria, Preeti Syal, Vishali Gupta, Priyanka Verma, Vaibhav Miglani, Deepmala Budhija, and Luke Vale*

**Chapter 6** Participation in AI: Notes from the Trenches ..... 72

*Tarunima Prabhakar, Cheshta Arora, and Arnav Arora*

**Chapter 7** Risk Assessment Methodology for AI Regulation and Navigating Liability Determination in an AI-Driven World: A Policy Paper on Risk Assessment..... 87

*Aditya Mohan and Karthik Satishkumar*

<b>Chapter 8</b>	Harnessing the Potential of AI for Indian Agriculture: Using “Bhashini” as a Tool to Deploy Responsible AI and Increase the Uptake of AI Applications Among Farmers .....	119
	<i>Abhishek Raj, Harsh Singh, and Anshul Pachouri</i>	
<b>Chapter 9</b>	Regional Inequities in Extraction and Flow of Resources That Support and Power the Design, Development and Access to AI: Lessons from the Global South .....	129
	<i>Saikat Datta, Shachi Solanki, and Anand Venkatanaryanan</i>	
<b>Chapter 10</b>	Assessing the Trustworthiness of Generative AI Used in Higher Education.....	144
	<i>Adarsh Srivastava, Gokul Gawande, Divya Dwivedi, Manu Dev, Vinayak Kottawar, Ishwar Chavhan, and Roberto V. Zicari</i>	
<b>Chapter 11</b>	Actionable Ethics: From Philosophical Principles to Operational Initiatives for Responsible AI Projects in the Public Sector in the French Context.....	155
	<i>Anth��a Serafin, Lisa F��riol, and Bertrand Monthubert</i>	
<b>Chapter 12</b>	Supporting AI at Scale in the APEC Region through International Standards .....	167
	<i>Aurelie Jacquet, Karen Batt, and Jesse Riddell</i>	
<b>Chapter 13</b>	Suggested Framework for Improved Algorithmic Auditing in India .....	186
	<i>Harsh Lailer, Gadamsetti Srija, Aseem Saxena, and Agrima Lailer</i>	
<b>Chapter 14</b>	Artificial Intelligence, Government, and Challenges: Initial Insights from Rwanda’s Mbaza AI-Chatbot Project .....	200
	<i>Lea Gimpel and Keegan McBride</i>	
<b>Index</b> .....		211

---

# Preface

Artificial Intelligence (AI) is being deployed in solutions to various real-world problems across multiple domains globally. It has in the last two decades evolved into a vast interdisciplinary research area. In the Indian context, with its social impact going beyond the targeted national priority sectors aimed at improving governance and public service delivery, AI is steadily making its foray into various other sectors and will soon be ubiquitous. Based on the tremendous potential and ability of AI to improve outcomes in the fields of education, healthcare and agriculture, India's National Strategy for Artificial Intelligence lists AI as one of the most important factors in the reform agenda of the Government and has highlighted the need for robust, cutting-edge research and testing ecosystem for solving problems of the society using AI.

Apart from ensuring the safe, fair and secure functioning of deployed AI systems, there is also a growing need to manage AI systems responsibly. AI systems have caused harm by developing biases or prejudices while making predictions and decisions in the real world. Such instances indicate that AI systems could cause large-scale harm to marginalised sections of the society who often either go under-represented or over-represented in datasets used to train AI models. This highlights the urgent need for developing actionable guidance on safe and trusted AI to monitor and manage AI systems responsibly through interdisciplinary collaboration and contribution from other domains, such as Law, Social Sciences, Business/Management Studies and policy research. Interdisciplinary research in this area can be a boon to the world and can help overcome the standard concerns in the development and deployment of AI-based solutions such as fairness, ethics, privacy, security and interpretability.

Research in responsible AI has so far been dispersed, with individual institutes and researchers working on their own domains and problem statements. To promote collaboration and research in responsible AI, the Ministry of Electronics and Information Technology (MeitY), in collaboration with the Centre for Responsible AI (CeRAI), IIT Madras organised a Research Symposium as part of the Annual Global Partnership on AI (GPAI) Summit 2023 in New Delhi.

GPAI is an international and multi-stakeholder initiative to guide the responsible development and use of AI, inclusion, diversity, innovation and economic growth. As the council chair of GPAI, India hosted the Annual GPAI Summit on 12–14 December 2023 in Bharat Manadapam, New Delhi. The Summit was attended by delegates from 28 GPAI member countries and the European Union, and had more than 22,000 participants, including 15,000 virtual attendees. The attendees included AI experts, multilateral organisations and other relevant stakeholders.

The Symposium, under the theme Responsible AI in Public-Sector Applications, provided a platform for Indian and international academics and researchers to collaborate with other AI experts and present their actionable research on responsible AI in front of a global audience.

This volume contains the expanded chapter versions of the talks and abstracts presented at the Research Symposium. The conference abstracts were also published



by Taylor and Francis as a booklet and distributed during the event. The Symposium had two tracks: one with invited expert speakers, and the other, a regular conference shortlist track. Our invited speakers and shortlisted authors include exceptional scholars and practitioners from engineering and public policy fields.

The call for papers for the shortlist track was issued on 24 July 2023. The response wildly exceeded our expectations. We received more than 150 submissions from 38 countries on various topics, from responsible AI principles, algorithmic accountability, and explainability to responsible AI assessments in less than a month. We selected 11 submissions after a rigorous review process with support from a committee comprising members from academia, industry and government. Many good submissions had to be turned down to achieve the short final list that fit into the tight schedule of the symposium program. The shortlisted abstracts come from more than 10 countries and include academics, medical professionals, government officials, civil society actors and private enterprises. After presentation at the conference, the participants submitted full-chapter versions, which comprise this edited volume.

The selection engages with an interesting and wide range of topics along several dimensions:

- **Application domains:** Healthcare, social media, agriculture, education
- **AI technologies:** Generative AI, speech and language processing, medical imaging
- **Governance functions:** Ethical principles, regulation, standardisation, auditing, liability determination, data protection, community participation, procurement, intellectual property management, resource extraction and flow
- **Geographic regions:** India, East Africa, Western Europe, APAC

We gratefully acknowledge the support of the various individuals and institutions that have helped us along the way. Mr. Gagandeep Singh at Taylor and Francis has been generous and patient with us. The reviewers took precious time off their busy schedules to help us shortlist the abstracts. Our distinguished invited speakers kindly obliged to participate on short notice. All selected contributors worked on a tight deadline for the full-chapter submissions. We are grateful to all of them for making this book possible.

**Balaraman Ravindran**

*Professor & Head, Wadhvani School of Data Science and AI  
Head, Centre for Responsible AI (CeRAI)  
Indian Institute of Technology, Madras*

**Abhishek Singh, IAS**

*President & CEO  
National e-Governance Division (NeGD)  
Ministry of Electronics and Information Technology (MeitY)  
Government of India*

---

# About the Editors

**Abhishek Singh** is a distinguished officer of the 1995 batch of the Indian Administrative Service (IAS), and currently serves as Additional Secretary in the Ministry of Electronics and Information Technology (MeitY), Government of India. With nearly three decades of experience in governance, policy formulation and administration, he specializes in leveraging technology to enhance governance and public service delivery. As Additional Secretary, he holds responsibilities including Artificial Intelligence, Emerging Technologies, Cybersecurity, and Digital Skilling with additional charge of CEO of IndiaAI Mission.

An alumnus of IIT Kanpur, where he earned a B.Tech in Mechanical Engineering, Singh furthered his education with a Master's in Public Administration from the Harvard Kennedy School as a Mason Fellow. Over the years, he has held several key positions, including CEO of Karmayogi Bharat, MyGov and the National e-Governance Division (NeGD). His leadership has been instrumental in implementing landmark initiatives, such as CoWIN, DigiLocker and DIKSHA. At the state level, he was with the Government of Nagaland and Uttar Pradesh and was responsible for implementing development schemes at the grassroots level, ensuring law and order as also for collecting revenues for the States.

**Balaraman Ravindran** heads the Robert Bosch Centre for Data Science & Artificial Intelligence and the Centre for Responsible AI at IIT Madras, the leading interdisciplinary AI research centre in India. He is the Mindtree Faculty Fellow and Professor in the Department of Computer Science and Engineering at IIT Madras. He is the premier Deep Reinforcement Learning Expert and among the top three Machine Learning Experts in India. He has been elected as ACM Distinguished Member (2021) for his significant contributions to computing. He has been recognized, in 2020, as a senior member of the Association for Advancement of AI (AAAI) for his long-standing contributions to AI. He is also the co-director of the Prathap Subrahmanyam Centre for Digital Intelligence, Secure Hardware and Architecture (PSC-DISHA) and the Reconfigurable and Intelligent Systems Engineering (RISE) group at IIT Madras.

He has been closely collaborating with various industrial research labs, such as Google Research, Intel Research, Ericsson R&D, and many more, working on applications of AI techniques to solve difficult real-world problems. He received his PhD from the University of Massachusetts, Amherst and his Master's in research degree from the Indian Institute of Science, Bangalore. He has more than two decades of research experience in AI and ML, specifically, Reinforcement Learning. He has held visiting positions at the Indian Institute of Science, Bangalore, India; University of Technology, Sydney, Australia; and Google Research. Currently, his research interests are centred on learning from and through interactions and span the areas of geometric deep learning and reinforcement learning.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

---

# Contributors

**Gaurav Agarwal**

Core Volunteer  
ISPIRT Foundation  
Koramangala, Bangalore, India

**Arnav Arora**

Doctoral Student  
University of Copenhagen, Denmark

**Cheshta Arora**

Senior Researcher  
Western Norway Research Institute  
Sogndal, Norway

**Karen Batt**

Standards Australia  
Sydney, Australia

**Deepmala Budhija**

Data Scientist  
Post Graduate Institute of Medical  
Education & Research  
Chandigarh, India

**Anshul Chauhan**

Doctoral Student (PhD)  
Advanced Eye Centre  
Post Graduate Institute of Medical  
Education & Research  
Chandigarh, India

**Ishwar Chavhan**

Scientific Advisor  
Trustworthy AI Lab,  
Z-Inspection  
Pune, India

**Saikat Datta**

CEO  
DeepStrat  
New Delhi, India

**Manu Dev**

Scientific Advisor  
Trustworthy AI Lab, Z-Inspection  
Pune, India

**Mona Duggal**

Associate Professor  
Advanced Eye Centre  
Post Graduate Institute of Medical  
Education & Research  
Chandigarh, India

**Divya Dwivedi**

Legal Advisor  
Trustworthy AI Lab, Z-Inspection  
Pune, India

**Sunu Engineer**

Core Volunteer  
ISPIRT Foundation  
Koramangala, Bangalore, India

**Lisa Fériol**

PhD Student  
University of Toulouse, Inserm  
Toulouse, France and Ekitia

**Gokul Gawande**

Scientific Advisor  
Trustworthy AI Lab, Z-Inspection  
Pune, India

**Lea Gimpel**

Co-Lead  
GIZ  
Bonn, Germany

**Niharika Gujela**

Masters Student  
Hertie School  
Berlin, Germany

**Vishali Gupta**

Professor  
Advanced Eye Centre  
Post Graduate Institute of Medical  
Education & Research  
Chandigarh, India

**Aurelie Jacquet**

Consultant  
Australian Government  
Sydney, New South Wales,  
Australia

**Ankita Kankaria**

Assistant Professor  
Department of Family and Community  
Medicine  
All India Institute of Medical Science  
Bathinda, Punjab

**Vinayak Kottawar**

Head of Department  
Artificial Intelligence &  
Data Science  
D. Y. Patil College of Engineering  
Akurdi, Pune, India

**Ramayya Krishnan**

Dean  
Heinz College of Information Systems  
and Public Policy  
Block Center for Technology and Society  
Carnegie Mellon University  
Pittsburgh, PA

**Agrima Lailor**

Strategy Associate  
OncoCheck, New Delhi, India

**Harsh Lailor**

Lead  
Quality Council of India (QCI)  
New Delhi, India

**Keegan McBride**

Former MSc SSI Course Director  
University of Oxford, UK

**Jens Meijen**

Doctoral Researcher  
KU Leuven  
Belgium

**Vaibhav Miglani**

Statistician  
Post Graduate Institute of Medical  
Education & Research  
Chandigarh, India

**Cari Miller**

Founder and Lead Researcher  
Center for Inclusive Change and the  
Ethical AI Institute  
Wilmington, DE

**Aditya Mohan**

Senior Scientific Officer  
National Standards Authority of  
Ireland, Dublin

**Bertrand Monthubert**

Professor of Mathematics and President  
of Ekitia  
Associate Researcher  
University of Toulouse, Inserm  
Toulouse, France

**Anshul Pachouri**

Senior Manager – Government and  
Social Impact  
MicroSave Consulting (MSC)  
Lucknow, Uttar Pradesh, India

**Prasanna Parasurama**

Assistant Professor  
Goizueta Business School  
Emory University  
Atlanta, GA

**Tarunima Prabhakar**

Co-Founder  
Tattle Civic Technologies  
Delhi, India

**Abhishek Raj**

Technology Policy Analyst  
MicroSave Consulting (MSC)  
Lucknow, Uttar Pradesh, India

**Jesse Riddell**

Senior International Partnerships  
Manager  
Standards Australia  
Sydney, Australia

**Karthik Satishkumar**

Cyber Security Executive  
VP of Professional Services,  
APAC at Saviynt  
Melbourne, Victoria, Australia

**Aseem Saxena**

PhD Student – Oregon  
State University  
University of Oregon, Portland

**Joao Sedoc**

Assistant Professor  
Leonard N. Stern  
School of Business  
New York University, NY

**Anth  a Serafin**

Associate Researcher  
University of Toulouse, Inserm  
Toulouse, France

**Harsh Singh**

Associate  
MicroSave Consulting (MSC)  
Lucknow, Uttar Pradesh, India

**Shachi Solanki**

Deputy COO  
DeepStrat  
New Delhi, India

**AVS Sridhar**

Core Volunteer  
ISPIRT Foundation  
Koramangala, Bangalore, India

**Gadamsetti Srija**

Associate  
Modular Open Source  
Identity Platform  
Bangalore, India

**Adarsh Srivastava**

Scientific Advisor  
Trustworthy AI Lab, Z-Inspection  
Pune, India

**Shyam Sundaram**

Core Volunteer  
ISPIRT Foundation  
Koramangala, Bangalore, India

**Arun Sundararajan**

Professor  
Leonard N. Stern School of Business  
New York University

**Preeti Syal**

Element 5.1, Ex-Director  
National Institution for  
Transforming India (NITI),  
Aayog

**Luke Vale**

Professor  
Population Health Sciences Institute  
Newcastle University  
United Kingdom

**Kapil Vaswani**

Core Volunteer

ISPIRT Foundation

Koramangala, Bangalore, India

**Anand Venkatanaryanan**

CTO

DeepStrat

New Delhi, India

**Priyanka Verma**

Research Associate

Post Graduate Institute of Medical

Education &amp; Research

Chandigarh, India

**Gisele Waters**

Co-Founder and Lead Researcher

AI Procurement Lab

**Roberto V. Zicari**

Advisor

Trustworthy AI Lab, Z-Inspection

Pune, India

---

# 1 Raising a Global Standard in the Procurement of Artificial Intelligence and Automated Decision Systems

*Gisele Waters and Cari Miller*

## 1.1 INTRODUCTION

Governments and organizations continuously seek to capture the benefits from digital technology innovation. The goals of improving public-sector productivity and the provision of public services by governments are also aimed at stimulating their economies (Edquist & Hommen, 2000). However, government agency personnel procuring machine learning systems (Robbins, 2019; Yin et al., 2019), artificial intelligence (OECD AI Policy Observatory, 2023) and automated decisions systems (Richardson, 2022) most often have little or no knowledge about their design or functions or how well these align with public policy and societal needs (Mulligan & Bamberger, 2019). Research shows that state actors and government representatives rarely understand the AI and ADS systems they are procuring or deploying (Deloitte US, 2023; Hickok, 2022; Nagitta et al., 2022). Innovating technology policy and governance for AI as it intersects with society, therefore, requires that governments not only build internal capacity but also raise the standard in how teams *practice* and *administrate* the procurement of AI. Our definition of internal or institutional capacity is the ability for government representatives to keep up with the skills needed to meaningfully perform their jobs. We focus on the practices (administration of AI procurement) not capacity building by introducing a new category of standard in AI procurement with new process guidance and tools.

## 1.2 AI ADOPTION LANDSCAPE

It is estimated that governments worldwide spend up to USD 11 trillion contracting for goods and services for the public sector (World Economic Forum, 2021). About 12% is spent by countries in the Organization for Economic Cooperation and Development (OECD) on public procurement (OECD Data Explorer, 2016). Approximately



USD 600 billion worldwide is spent by governments on information technology (IT) (Statista, 2023) across segments. Statista (2023) also reports that this USD 600 billion is an 8.9% increase in IT spending from 2022. The expenditure signals are less clear about how much of government spending is attributed to artificial intelligence, specifically because governments worldwide are still learning to identify and classify AI systems and solutions. One forecast by the International Data Corporation (IDC) suggests that worldwide spending on AI, including software, hardware, and services for AI-centric systems, will reach USD 300 billion in 2026 (IDC, 2022). Another indicator of *public* investment in AI technologies in the United States, for example, is the level of spending on government AI contracts across the federal government. The data derived from a Bloomberg government-built model in the *AI Index Report on Policy and National Strategies* (OECD, 2021a) reveals that the amount is higher than ever, reaching USD 1.8 billion in 2020 as compared to USD 1.5 billion in fiscal year 2019, a 25% increase in one year alone. However, USD 1.8 billion is a sixfold increase from five years earlier in 2015, when spending on AI was only USD 300 million (OECD, 2021a). In contrast, the total central government expenditures on AI-focused companies were approximately USD 6 billion in China in 2018 (Colvin et al., 2020). In other words, AI-related spending in the USA and to a greater degree in China is significant, albeit occurring in two inherently distinct political extremes.

These statistics are meant to illustrate that having limited knowledge (or even none at all) about the design or functions of AI and ADS systems has not curtailed their procurement. Spending continues regardless of knowledge gaps or institutional capacity in government. In this chapter, we argue that, at least for the proportion of spending on *high-risk AI-enabled* technology (see Appendix 1.1 at the end of the chapter) serving the public interest, a voluntary higher standard of practice should be combined with existing jurisdictional requirements for acquisition. More specifically, certainly greater consideration should be given to adapting procurement processes when *AI-enabled solutions or ADS are making critical decisions for the populace* (e.g., access to educational assistance, housing, welfare, work opportunities, health services). Adding greater due diligence for these types of procurements could be considered part of an organization's fiduciary duty (Fontenot & Gaedt-Sheckter, 2020) – a duty inherent to responsibly managing public funds when purchasing from the private sector.

### 1.3 AUTOMATION WITHOUT CITIZEN REMEDY

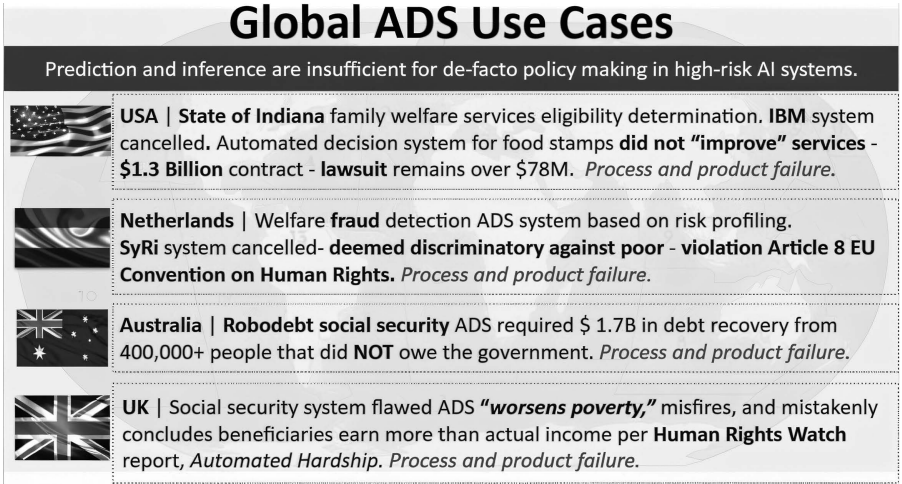
Government entities have been consumers and stewards of public use technology for more than half a century (Chen & Ahn, 2017; Edquist & Hommen, 2000). But today in government, AI and ADS are procured and used daily to support or replace human decisions and judgments that impact critical life opportunities, access, liberties, rights, and the safety of the citizenry (Eubanks, 2018; Fergusson, 2023). In fact, in the United States, the Electronic Privacy Information Center (EPIC) analyzed 621 contracts across states, and they reported that AI companies have taken over government decision making. Buying and adopting harmful high-risk AI systems that afford the public little to no protection without much human oversight (Fergusson, 2023) essentially guarantees potential harms to citizens. Lawsuits across the country

and indeed the world reflect the harms in outcomes legally contested (Acemoğlu, 2023; Kuziemski & Misuraca, 2020; Whittlestone et al., 2019). The multilayered complexity of AI systems (Weissinger, 2022) and processing of large datasets (for individuals and communities) in automated decisions will continue to present technical, socio-technical, and social challenges to institutional capacity unless either the knowledge gap, the practices, or both are purposefully addressed. P3119 is part of the needed innovation answering the call to learn a better way of practicing procurement adapted to AI. But these considerations also land in a landscape where the vast majority of traditional government procurement models worldwide have yet to *adapt* their acquisition regimes and procurement laws (Casovan & Shankar, 2022; Rimes, 2021; Sanchez-Graells, 2024b). “Accountability in a computerized society” (Nissenbaum, 1996, title) with *existing* technology is challenging enough because digitalization already pervades almost every aspect of public and private life (Nissenbaum, 1996). Unfortunately, while bringing the universally marketed benefits (e.g., faster analytics, greater efficiency, scale), the unique harms of high-risk AI systems (Acemoğlu, 2023; OECD Federal Ministry of Labour and Social Affairs, 2022, 2023; Xia et al., 2023) exacerbate the government accountability challenge.

What also makes the proliferation of high-risk AI and ADS used by government so dangerous is that uninformed procurement leads to adoption without meaningful opportunities for the public to dispute automated decisions (i.e., remedies, forms of redress are not considered when procurement contracts are signed, and life circumstances can often be adversely affected). Prediction and inference automation, for instance, can take priority over fundamental human rights, producing de-facto policy using appropriated funds (Rubenstein, 2021). An example of this is the U.S. Internal Revenue Service (IRS), which signed a USD 86 million contract with ID.me to provide biometric identity verification services in 2021. ID.me required taxpayers to submit their biometrics, but the risk is that if the service does not perform equally and equitably across different demographics, a taxpayer may be subject to identify theft at the highest level since the biometric identifiers are unchangeable (Buolamwini, 2022). Where is the responsibility of fiduciary duty (Benthall & Shekman, 2022/2023) when this happens? From the layperson’s perspective, when the value of automation and digital innovation takes precedence over reasoned policy administration, a *right to an effective remedy* (Article 47 in the European Union’s (EU) Charter for Fundamental Human Rights) can potentially be violated (EU Agency for Fundamental Rights, 2009). Worthy of mentioning is Article 41 in this same EU Charter, which provides complementary legal protection against harms by questionable government administration (digital or otherwise). Both of these Charter Articles give the EU supervisory authority similar to the right to be heard and to have decisions on one’s interests made fairly and impartially, embodied in the Due Process clauses of the U.S. Constitution and in a wide range of statutes, including the Administrative Procedure Act (APA) (Mashaw, 2007; Mulligan & Bamberger, 2019).

Moreover, AI vendors that sell to government agencies maintain their systems are proprietary, so teams are *forced* to depend on AI contractors without understanding their original design or function. The IRS had almost half a million cases in its Identity Theft Victims Assistance unit to work through by the end of 2023, with “unconscionable delays,” said a National Taxpayer Advocate Erin Collins (Alms, 2024).





**FIGURE 1.2** Global ADS use cases.

Source: Image by authors adapted by Microsoft Bing Designer.

for the United States all help to corroborate that many governments blindly rely on AI vendor marketing and their products to make critical government decisions about the public. To emphasize a worst-case scenario, Australia’s Royal Commission Report into the Robodebt Scheme also teaches us that governments reap what they sow when they proceed without caution (Commonwealth of Australia, 2023). In that scenario, it’s business as usual, as if ADS did not already add to shortcomings in large bureaucracies without adaptations to procurement or digitalization processes and standards (Autio et al., 2023; Sanchez-Graells, 2023). As Mulligan and Bamberger (2019) state more succinctly:

When the adoption of these systems is governed by procurement, the policies they embed receive little or no agency or outside expertise beyond that provided by the vendor. Design decisions are left to private third-party developers. There is no public participation, no reasoned deliberation, and no factual record, which [in essence] abdicates Government responsibility for policymaking.

(Mulligan & Bamberger, 2019, abstract)

Other precarious global ADS use cases highlighting legal precedents and harms to the public are shared in Figure 1.2.

In the face of these digital advances that are exceedingly different from information technology and past software applications (Fallon, 2023), the IEEE Standards Association authorized the development of a brand-new category of socio-technical standard in 2021, the P3119 Standard for the Procurement of AI and ADS. Inspired by the roundtable findings and authors of *AI and Procurement – A Primer* (Sloane et al., 2021), IEEE expanded their AI governance

standards (IEEE Standards Association, 2023a) that guide organizations on ethical AI design, deployment, and certifications into AI procurement. Voluntary consensus-based standards (Cihon, 2019; Rioux, 2020) can *strengthen and optimize* AI procurement approaches with due diligence processes and tools that are currently unavailable (OECD, 2021b). This type of standard will also support governments to use their procuring power as a *market-shaping mechanism* (MSM) that can create a clear demand for responsible AI solutions potentially solving market failures (Day One Project, 2022).

#### 1.4 PRESSURES FROM AI INFALLIBILITY: SELLING AI SPEED AT PROCUREMENT CONFERENCES

As part of the WG's research into the procurement landscape, we analyzed about twenty of the largest annual procurement conferences worldwide, none of them from 2020–2023 addressed the specific challenges aforementioned. The primary foci of these global procurement conferences, unsurprisingly, was the infallibility of AI solutions to offer unquestionable benefits to organizations. Other themes found were digital transformation, supply chain sustainability, expansion of AI tooling for increased operational efficiency, and robotic process automation among others – all related to AI vendor values of bigger, better, faster efficiency, automation, and productivity. Thus, while governments continue to address the need to adapt to the opportunities *and risks* in AI procurement and while the few innovate how AI-enabled technologies are procured for high-risk uses (United Kingdom (UK), Brazil, Chile, Bahrain, and United Arab Emirates) (WEF, 2020a), public and private sectors worldwide are still missing much-needed targeted AI strategies (Singh et al., 2023), practices, and tools. Tools that are much needed to enable adapted procurement practices in the public sector so that *responsible* AI are adopted and used (Autio et al., 2023; Kuziemski & Misuraca, 2020; Rubenstein, 2021).

General AI governance and ethical AI guidance frameworks are also proliferating globally in every jurisdiction at unprecedented rates. It seems every day locally, regionally, nationally, and transnationally new AI governance frameworks are published across public and private sectors. At least the OECD/G20 AI Principles have framed the global debate over AI policy (OECD, 2019). The OECD/G20 principles have significantly shaped the policies and practices of fifty governments who have formally endorsed them. But cohesive uniformity in regulation in this field with transnational collaboration will be a global Sisyphean task for many years to come. A positive signal in the ecosystem is an additional coordinating committee, a group of twenty (G20) member countries was proposed for the governance of artificial intelligence (CCGAI) to plan and coordinate on a *multilateral* level the mitigation of AI risks (Jelinek et al., 2021). We may be decades away from global harmonization between domains such as AI governance frameworks, laws, and international consensus-based standards, but global leaders are certainly awakening to the opportunity for AI policy coordination and a deeper understanding of AI risk. Europe is certainly leading in how laws and standards are harmonized against their EU AI Act and going further into operationalizing the same through conformity assessments where relevant (Edwards, 2022; Madiega, 2023; Veale & Zuiderveen Borgesius, 2021).

Despite historical AI vendor pressures to buy indiscriminately, vendor behavior is now under greater scrutiny than it was ten years prior. Regardless of geography, we vigorously agree with EPIC’s comments that agencies should not adopt AI solely for AI’s sake (EPIC, 2023). EPIC commented on the White House Executive Order, 14110 of October 30, 2023 (Executive Office of the President of the United States, 2023) and its implementation guidance offered by the Office of Management and Budget (Young, 2023). These national orders add credence to the notion that the United States is beginning to take the issuance of AI safety signals to agency administration and the private sector more seriously.

To date, grounded and practical guidance detailing processes on *how to* mitigate AI risks is exceptionally rare. IEEE-SA is the only international standards development organization (SDO) drafting the type of guidance that moves beyond AI policy, frameworks, principles, and best practices into a *detailed how-to process* guidance for the procurement of AI. P3119 also deliberately addresses the technical complexity of most AI models (especially for the high-risk domain), the supply chain accountability challenges, and the potential downside risks of the effects of scale that AI and ADS models and systems possess. Whereas other AI procurement approaches addressed by the U.S. Government Accountability Office’s AI Accountability Framework (US GAO, 2021), the World Economic Forum’s AI Procurement in a Box (WEF, 2020b), and the Ford Foundation’s Guiding Framework (Conti-Cook & Taraaz, 2023) offer *AI literacy*, *best practices*, *AI readiness principles*, and *red flags* for vetting technology vendors in the public sector (see Figure 1.2), P3119 offers comprehensible detailed “*how-to*” *process* guidance across the procurement life cycle. That life cycle starts in pre-procurement stages and moves all the way through to AI-specific contract monitoring (Miller & Waters, 2023) with expanded transparency requirements (Felzmann et al., 2019) tooled with new AI procurement protocols.

# AI Procurement Guidance Comparison





Responsible AI procurement requires detailed processes governance.  IEEE P3119 addresses this gap.		 <b>IEEE SA</b> STANDARDS ASSOCIATION  <b>P3119 AI Procurement</b>	 <b>GAO</b> U.S. GOVERNMENT ACCOUNTABILITY OFFICE  <b>AI Accountability Framework</b>	 <b>WORLD ECONOMIC FORUM</b>  <b>Procurement in a Box</b>	 <b>FORD FOUNDATION</b>  <b>Guiding Framework</b>
<b>Individual AI Literacy</b>	<ul style="list-style-type: none"><li>• Terminology</li><li>• Technical Education</li><li>• Lifecycle Risks</li></ul>	Not in scope	Limited Guidance	Best Practice Guidance	Best Practice Guidance
<b>Organizational AI Readiness</b>	<ul style="list-style-type: none"><li>• Policies</li><li>• Data Quality</li><li>• Process Discipline</li></ul>	Not in scope	Comprehensive, robust, detailed	Best Practice Guidance	Limited Guidance
<b>AI Procurement Processes &amp; Tools</b>	<ul style="list-style-type: none"><li>• Specific Steps</li><li>• Guides and Templates</li><li>• Tools and Rubrics</li></ul>	Comprehensive, weighted, robust, detailed	Well developed, robust questions, missing tools & rubrics	Theoretical Guidance	Theoretical Guidance

FIGURE 1.3 AI procurement guidance comparison.

Source: Image by authors.

For example, detailed activities and tasks in normative guidance are provided on how to analyze and define a perceived government problem (or business case/need) that may require an AI solution (or not), how to identify a government agency's risk appetite (Miller & Waters, 2024), how to assess and measure vendor answers to essential AI governance questions, and how to prompt vendors and evaluate their AI solutions across multiple criteria beyond proposed AI model cards (Trustible.ai, 2023).

Over two years, the P3119 Working Group volunteers developed consensus around a uniform set of definitions and also the process model approach that the standard would offer. For the most part, discussions in the WG centered around critical evaluations of AI risk identification, mitigation, and reduction within the following segments of a common procurement process model: (1) procurement need (public sector problem), (2) vendor AI governance, (3) solution solicitation/proposals, (4) contract negotiations, and (5) contract management to augment and support procurement modernization. These new P3119 processes are not meant to replace but rather to complement and optimize existing procurement requirements. The primary goal for this new standard is to offer government agencies and AI vendors the opportunity to adapt and innovate their procurement practices and solicited proposals in order to maximize the benefits of AI while minimizing the risks. An additional overview and further details will be provided in the P3119 section later in the chapter.

In summary, P3119 is meant to become a part of an organization's request for proposals (RFPs) or solicitations, integrated with solicitations in order to raise the standard in AI procurement administrative processes so that the public interest and their civil rights are *proactively* and *responsibly* addressed and protected. The vision is to help support team members in organizations and government agencies act in the best interests of the public they serve, particularly when public funds are used to procure high-risk AI solutions intended for the public's benefit. Acting in the best interest of others is in essence part of the definition of *fiduciary duty* (CLS, 2023). When fiduciary organizations like governments interact with users (citizens) and automate their operations and the public benefits adjudicated by the same, they have a legal duty to act with loyalty and care towards the public that trusts them (Benthall & Shekman, 2022/2023; IEEE-USA Position Statement, 2020).

## 1.5 TEAMS AND SANDBOX CHALLENGES: TRANSDISCIPLINARY COLLABORATION PREFERRED

AI procurement in general and high-risk AI procurement in particular are challenged by team and sandbox testing requirements. First, we address teams. The right teams to evaluate and test benchmarks in a sandbox would be optimal to learn from AI procurement failures and successes. We define sandboxes later. Here, we draw special attention to the form and function of teams as an optimal AI procurement strategy ambitiously working towards the ideal. We have already mentioned the knowledge gap in the public sector, and this points to the obvious need to upskill, educate, and build public sector capacity (Holmes et al., 2019). Many organizations and experts in the field of AI policy, AI work transformation, and AI governance have addressed basic training on AI fundamentals and the need for the same in AI procurement



**FIGURE 1.4** Transdisciplinary collaboration.

Source: Image by authors generated by Microsoft Bing Image Creator and adapted by Microsoft Bing Designer.

(Autio et al., 2023; Hickok, 2022; Rimes, 2021; Sloane et al., 2021; WEF, 2020a). We point to another aspect related to teams: the nature of the collaboration between members in the team to perform well in high-risk AI procurement.

AI procurement can benefit from the kind of transdisciplinary work (collaboration and research) the OECD recommends. The OECD defines this work as the integration of knowledge from different disciplines and non-academic stakeholder communities to address complex societal challenges (OECD, 2020). Transdisciplinary collaboration teams are ideal to begin exploration on the drivers, inhibitors, interests, and expectations of different actors in responsible AI procurement. It is no small feat to bring together an AI procurement team that is transdisciplinary *beyond* the interdisciplinary kind with the right skillsets into a high-performing transdisciplinary collaboration (between internal/external, public/private, technical/non-technical, academic/non-academic parties and even citizens as primary stakeholders) (Burris, 2022; Guimarães da Costa, 2021; Hocking et al., 2016). We recommend the bar be set high when raising team standards. Organizations and government agencies do not have to start there, but it is an honorable end state to strategically plan for. Teams will be better able to adapt RFPs/solicitations to adequately evaluate the AI tenders/vendors, for example (more on this later as described in the P3119 vendor and solution evaluation processes).

Types of diverse team expertise needed might be the following (including but not limited to): government leadership, data science, software engineering, procurement, program domain, user-experience development, human-centered service design, digital privacy, AI governance, and legal counsel. If the preferred team members are



not available in-house (it would be rare if they were), then proactive outreach to external parties with the right skills is absolutely needed for high-risk AI systems. Since buying AI (high-risk or not) will continue regardless of whether government personnel know or understand its material application, governments should strategically prepare for the purchases using a *caveat emptor* (Harris et al., 2008) philosophy that burns a fire under the organization's feet to search for the right teams far and wide. The diverse talent exists (Chakravorti et al., 2021; HAAS School of Business, 2019), but organizations must persist in their searches (Nihill, 2024). As Jennifer Pahlka, former U.S. deputy chief technology officer within the Office of Science and Technology Policy, said in an interview with Fed Scoop, "government needs to pair mandates with enablement" (Nihill, 2024). This statement was made after her testimony at the Senate Homeland Security and Governmental Affairs hearing regarding "Harnessing AI to Improve Government Services and Customer Experience" on January 10, 2024 (HSGAC, 2024). We could not agree more with Jennifer Pahlka's testimony and much of the testimony provided by the other witnesses. As we stated earlier, building diverse teams only *begins* to address the many additional challenges in AI procurement within the public domain.

## 1.6 WHY TRANSCEND

Public procurement is defined by the Organization for Economic Cooperation and Development (OECD) as "the purchase by governments and state-owned enterprises of goods, services, and works from other organizations" (OECD, 2023). When related to AI and ADS, these purchases are often found in the private sector because governments do not normally build AI or integrated ADS solutions. Also, as stated earlier, traditional procurement personnel have even less awareness of its impact on the communities served by it (Hickok, 2022; Mulligan & Bamberger, 2019; Sloane et al., 2021); hence, the need for special attention to the ways the private and public sectors collaborate to benefit the public. Complexity theory (Turner & Baker, 2019), Luhmann's theory of social systems (Luhmann, 1982), and the simulation of Robert Rosen's anticipatory system (Leydesdorff, 2005) inform our understanding of the complexity parameters impacting team member relationships. As such, our approach critically evaluates how successful human-computer interaction can be under the strains of a complex set of administrative processes (Demircioglu & Vivona, 2021) and the inherent conflict of interest between human buyers and technology sellers (Maksimainen, 2011).

In acknowledgement of the tall order and the long tail of this high standard, we maintain filling the gaps in institutional capacity to advise on the technical and non-technical aspects is at least one healthy way to begin an AI procurement project if transdisciplinary collaboration is beyond the wire. The AI Procurement Lab (AIPL, introduced later) can support upskilling for the mastery of AI procurement transdisciplinary collaboration, but we will continue to stress that the average multidisciplinary (separately working towards a shared goal), cross-disciplinary (shared perspectives/functions working towards a shared goal), or even interdisciplinary (integrating contributions towards shared goal) teams will not suffice for high-risk AI procurement. Ideally, government agencies should train the recommended team members for high-functioning and high-performing versions of collaboration. Teams must

endeavor to build strong collaborative relationships. The transdisciplinary collaboration approach transcends disciplinary and domain boundaries to contribute a *unified* understanding of the problem and a *cohesive* method to work towards a shared goal (Farrell, 2011). This kind of transdisciplinary collaboration is also being done in public health contexts (Burris, 2022) and in efforts to advance knowledge about wicked problems such as sustainable development goals (Guimarães da Costa, 2021). The guidance and tools provided by P3119 can be the glue that helps bind collaborative relationships, because the level of specificity is focused on processes, activities, and tasks. With the support of IEEE, the standard could, in the future, be adapted into transdisciplinary AI procurement curricula guiding team members with bespoke modules. AI procurement risk management training modules are already available at the AI Procurement Lab in partnership with the Center for Inclusive Change (Miller & Waters, 2024).

## 1.7 CASE STUDY ILLUSTRATION

The Chief Procurement Officer for Nestlé, Patricia Stroup, stated that procurement is a *great connector* because it touches internal sections of business across functions, and it also connects to the greater external value chain (ProcurementMag, 2023). We agree but also remind our readers, procurement is not only a great connector for the operational act that is the *purchase* but also, inside government agencies, a set of very complex administrative processes (Demircioglu & Vivona, 2021) that can challenge even the best of intentioned teams, as illustrated by the New York City example that follows.

### 1.7.1 BEYOND NEW YORK CITY'S BEST-LAID PLANS

In late 2017, New York City became the first US jurisdiction to create a task force that made recommendations for government use of ADS (Richardson, 2019, p. 7). New York City (NYC) is a city of almost 9 million people (US Census Bureau, 2022), and the government intended to attempt a higher standard of responsible ADS adoption (as compared to large cities without the same efforts) in unexplored territory (called greenfields). Good administration is hard enough in known landscapes, let alone in unknown territory. Procurement greenfields (which also exist in many other industry sectors) refer to conceptual spaces where new value can be created because none has previously existed. In this unexplored greenfield, NYC wanted to proactively address how it could better protect the people it serves when adopting and using ADS. It possesses one of the largest municipal budgets in the world, and this project was thought to be an ideal laboratory to evaluate risks, opportunities, and benefits of ADS use in government. Governments worldwide are struggling with similar issues, so this short case study can be useful to better understand the team collaboration challenges with AI or ADS adoptions and procurements even when the law and positive intent support the mission.

The NYC Automated Decision Task Force was established by Local Law 49 (NYC, 2018) to write a set of recommendations that addressed various administrative innovations related to ADS. Their cross-disciplinary team members ranged from city officials such as the Mayor's Director, criminal justice counsel, human rights commissioner, and chief analytics officer, to an external information science researcher,

department of education general counsel, social services counsel, and assistant professor of computer science at New York University, to name a few. A diverse set of cross-disciplinary collaborators, for certain. Noteworthy of this NYC Task Force effort is that it not only included a group of city officials but also a set of both technical and socio-technical external experts to make recommendations on unexplored operational acts and administrative processes.

Per their web page, the set of recommendations the Task Force was charged with needed to address the following:

- I. a new procedure for impacted individuals to request information on decisions involving automated decision systems,
- II. a procedure for NYC to determine any disproportionate impact on protected categories of persons,
- III. a procedure for addressing any individual instances of “harm” if a system disproportionately impacts protected categories of persons,
- IV. a feasibility analysis of archiving agency systems and its associated data,
- V. a process for publicly disclosing information about agency systems, and
- VI. criteria for identifying which systems should be subject to one or more of the above.

(NYC, 2023, About pg)

These tasks were not trivial. Each one of them could be a lengthy project of discovery and reasoned deliberation among the team, requiring extensive collaboration throughout internal and external value chains. Indeed, in testimony submitted to the New York City Council Committee on Technology, two Task Force members wrote:

The intent of Local Law 49 of 2018 is to uphold two important principles in the use of ADS in City agencies: to enable greater government transparency and accountability, and to ensure fairness and equity. Yet the Task Force has failed to fully satisfy these principles. Despite numerous requests, Task Force members have not been given any information about ADSs used by the City. To date, the City has not identified even a single system. Task Force members need to know about relevant systems used by the City to provide meaningful recommendations. A report based on hypothetical examples, rather than on actual NYC systems, will remain abstract and inapplicable in practice. The Task Force cannot issue actionable and credible recommendations without some knowledge of the systems to which they are intended to apply. The need for examples has been raised by several of us on numerous occasions, but remained unaddressed until yesterday, just one day before this hearing, with the City suggesting that two examples might be forthcoming, at some unspecified future date. The City has cited concerns with privacy and security in response to our requests, but these cannot be used as blanket reasons to stand in the way of government transparency. Privacy and security considerations must be thoughtfully addressed as part of the process of formulating recommendations for transparency and accountability. However, we can only determine how to navigate these tensions if basic details about actual ADSs – and specific concerns that justifiably counsel against transparency – are shared with the Task Force. These cannot be negotiated in the abstract.

(Stoyanovich & Barocas, 2019, pp. 2–3)

This long quote was not redacted because of the breadth of value in insights that can be gained from the full text. Despite the ambitious intent of Local Law 49 to support the Task Force mission, in addition to the well-constructed cross-disciplinary team built for the task, the city itself raised concerns that created obstacles to transparency and obstacles to making progress on their own mission. By failing to provide the requested information, as the two members stated, the Task Force mission was a non-starter. This testimony reflects how difficult it is to operationalize new ways of achieving a shared mission and the nuances of challenging collaboration in the public sector. Despite best-laid plans, the team gathered, and the legal red carpet that paved the way, government protocols still curtailed forward progress in a timely manner. In the future, to cement accountability more firmly for trustworthy transparency (Felzmann et al., 2019), policy mechanisms need to clearly define the mission's objects of governance as well as comprehensible deliverables across government departments (Ada Lovelace Institute et al., 2021).

The AI Now Institute and Richardson (2019) wrote various recommendations based on the New York City experience as it relates to teams and advocacy coalition work that add weight to our emphatic calls for transdisciplinary collaboration teams. Among other recommendations, they advise that a multidisciplinary coalition is best. They also report that advocacy coalitions with members who come from a variety of disciplines, issue areas, practices, and skill sets must ensure that their collective strategies, knowledge, and interests *center around those most affected* by the ADS (Richardson, 2019, 2022, p. 50). A mission we support, for certain. But again, we push further. Multidisciplinary teams are a necessary beginning, although they are insufficient for optimal results in high-risk AI procurement. The good news is, in October 2023, the NYC mayor revealed the NYC AI Action Plan for “responsible municipal government use of AI” that is focused on procurement and workforce technology upskilling while also improving the city residents’ quality of life (Taele, 2023).

### 1.7.2 RELENTLESS GRIT

Even legal and administrative well-laid plans can be derailed by complication and complexity in social systems (Poli, 2013) when trying to keep the public safe, innovate public services, and stimulate economies with digital technology innovation. The lessons learned by the Australian Robodebt and New York City examples and countless others worldwide often lead to a similar conclusion, not unfamiliar to the private sector as well: individuals, organizations, and companies tend to reinforce the status quo (Fergusson, 2023; Marinotti, 2021; Pahlka, 2023; Richardson, 2019; Scott, 1997; Sieber & Brandusescu, 2021). This is one of the primary entrenchments with institutional innovations (Yang, 2016) that inform our rationale for emphasizing the need for high-performing teams using transdisciplinary collaboration for high-risk AI procurement. If the aim is to do right by citizens to guard their monies while improving public benefits and services that serve them, then protecting the status quo must be contradicted with grounded guidance and tools, excellence in talent, and relentless grit – all of which are needed for the brave few willing to test raised standards of practice for AI procurement in regulatory sandboxes.

1.8    **SANDBOX TESTING: BENCHMARKING  
RETROSPECTIVE OR FUTURE AI PROCUREMENT**

Sandboxes for children or adults are places where play (testing and exploring boundaries) can be safely accomplished. Regulatory sandboxes are controlled environments that can facilitate the development, testing, and validation (benchmarking) of innovation (products, standards, services, and systems) before their placement on the market or before putting into service a specific plan (Ivanova, 2021; Martin & Balestra, 2019; Pop & Adomavicius, 2021). In efforts to advocate for P3119 and seek a sandbox partner, we presented the standard’s development with more than eight different national or transnational government agencies (various in the USA, UK, Brazil, Belgium, India, Japan, Australia, and the EU). We found that finding a sandbox partner is far more challenging than expected, per the diverse feedback from representatives. After two-plus years of iterating with the WG on the processes in the maturing P3119 process model deliberating what a new AI procurement standard “should be” to support responsible AI adoption, it turns out that – regardless of the WG consensus gained or the rationale – historical regulatory regimes and law, institutional capacity and bandwidth, and negotiating the parameters (proof-of-concept agreement for a sandbox agreement between IEEE and X organization) are arduous challenges to partnering on the innovation of AI procurement.

In 2023, across eight months, Dr. Gisele Waters and Dr. Cari Miller, with the support of WG members, engaged government agencies in international and virtual conferences, and in countless email exchanges about partnering with IEEE in a regulatory sandbox (Martin & Balestra, 2019) (in other words, a proof-of-concept [POC] pilot meant to test the standard and benchmark AI use cases). These sandboxes would allow organizations, government agencies, and IEEE to better understand the P3119 value *pre-publication* with real-world feedback on the guidance and tools being developed, their procurement processes, and target use cases. Sandbox testing can afford the opportunity to benchmark existing AI use cases (retrospective or future), with legacy procurement administrative processes not yet adapted to match the complexity and risk landscapes of AI acquisitions.

The list of briefings in Table 1.1 reflects the global interest in our consensus-based working group knowledge of AI procurement developed across two-plus years.

**TABLE 1.1**  
**Global Interest: Local, National, Transnational Briefings on the IEEE P3119 AI Procurement Standard**

Country/ Region	Gov Agency, NGO, AI Vendor Conference	Focus	Supporting Organization	2023/2024	Policy Interest	Sandbox Interest
USA	MA Bay Transit Auth	AI Procurement	IEEE	December	YES	TBD
USA	Congressional AI Caucus	Procurement Standard & AI Gov	IEEE-USA	March	YES	N/A

**Table 1.1 (Continued)**  
**Global Interest: Local, National, Transnational Briefings on the IEEE P3119 AI Procurement Standard**

Country/ Region	Gov Agency, NGO, AI Vendor Conference	Focus	Supporting Organization	2023/2024	Policy Interest	Sandbox Interest
USA	DOD (OASD)	Procurement of AI	IEEE	August	YES	No
USA	NAIAC Subcommittee	Procurement of AI	IEEE	August	YES	N/A
USA	Federal Agency	Procurement of AI	Federal Gov	September	YES	N/A
USA	Homeland Security Government Affairs	Procurement of AI	IEEE	September	YES	Perhaps
USA- Texas	Applied Intelligence Live AI Vendor Conf	Procurement of AI	Informa Tech	September	N/A	No
Belgium	FARI AI for Good	Procurement of AI	CEIMIA (Canada)	September	N/A	N/A
Brazil	C4IR/Governo de Estado São Paulo	Procurement of AI	IEEE	September	Future interest	Perhaps
UK	Local Government Association	Procurement of AI	IEEE	October	YES	YES!
European Union	European Food Safety Authority	Procurement of AI	Intellera (Italy)	October	YES	YES!
India	Ministry of Electronics & IT	Procurement of AI, GPAI Summit	GPAI/CEIMIA (Canada)	December	Future interest	No
Japan	FRIS Symposium	Ethical design & procurement	IEEE	July 2023	N/A	N/A
Australia	Victorian State Gov	Procurement of AI	IEEE	2024	TBD	TBD
Chile	Chile Compra	Procurement of AI	IEEE	2024	TBD	TBD

The line items (rows) shaded in grey are agencies that have shown maximum interest in our process guidance and tools thus far. The conversations we have nurtured around the potential for sandbox testing and benchmarking with the EU and UK show the greatest promise.

**1.9 POTENTIAL SANDBOXES: EUROPEAN FOOD SAFETY  
AUTHORITY AND UK’S LOCAL GOVERNMENT  
ASSOCIATION**

In November 2023, EU and UK government entities confirmed their interest to explore with IEEE potential sandbox partnership engagements. These folks are brave explorers applying foresight to *sandbox experimentation* willing to be open to pilot

testing of the developing standard and learning about AI use cases and their administrative processes from the proof-of-concept testing and benchmarking. Specifically, the European Food Safety Authority (EFSA) and the UK Local Government Association confirmed their serious interest in a sandbox pilot testing of P3119 or one of its process components. EFSA is no stranger to public-private partnership explorations and evidence management in risk assessment (Bersani et al., 2022; EFSA, 2010). They recently published a report documenting AI use cases in specific verticals using their framework contracts (Cagnoni et al., 2023). Academia from various universities contributed to that report, in addition to public and private sector organizations. They explored the capability of AI tools used for public consultations, and their progress is available to the public.

The United Kingdom is also not a stranger to building or testing AI procurement innovations. The UK Guidelines for AI Procurement were published as part of the World Economic Forum's project (WEF, 2020b), the AI Procurement Toolbox. These UK guidelines are aimed at central government departments considering the suitability of AI solutions to improve existing and future services. The UK also used the WEF's Toolbox in two case studies in their Department of Business, Energy, and Industrial Strategy and the Food Standards Agency. Furthermore, the UK also recently approved through final Royal Assent a 2023 Procurement Act that will transform the way public procurement is regulated (United Kingdom, Procurement Act 2023 Chapter 54, 2023). Expectations are that it will come into effect by October 2024. Supplemental regulations have been consulted and will be published in the first quarter of 2024. Their recent interest (from a Local Government Association) in IEEE's P3119 standard could be interpreted as an attempt to lean further into more robust implementation and administrative guidance that has yet to address the grounded tactical activity and process details that are often needed for good administration. We will be following up with both government agencies to finalize a POC agreement between parties.

Our aspirations and those of the government agencies we met with are to better understand what *could be* as opposed to *what is* in the current AI procurement landscape. P3119 is a brand-new category of standard that requires tacit authorization to address legal, operational, and administrative constraints in how things are done today in any given geography. EFSA will have the EU AI Act to contend with, in addition to their own acquisition and procurement regulatory regimes. Likewise, the UK LGA brings unique requirements from their regulatory environment (GDPR, equality duties and the new bill). On top of all that, the UK's Information Commissioner's Office and the Equality and Human Rights Commission (regulators for data protection and the public sector equality duty) are tasked with developing questions and/or standards for local authorities to support their compliance with these statutory duties in the procurement of AI. Combined, this is an elaborate set of legal and administrative variables to wade through for any sandbox benchmarking.

Both of our potential sandbox partners have national and transnational considerations that require layers of approval and governmental administration deliberation, but they are both still willing to be open about exploring a sandbox opportunity. They consider benchmarking against a new standard seriously, and we continue nurturing



**FIGURE 1.5** Sandbox benchmarking.

**Source:** Image by authors generated by Microsoft Bing Image Creator and adapted by Microsoft Bing Designer.

discussions on a POC agreement. With real-world feedback on the standard and the nature of its usefulness to responsible AI procurement, we look forward to final publication of the standard by the end of 2024 (latest 2025) after ballot approval and any refinements on the draft development (IEEE Standards Association, 2024). For the benefit of our chapter readers, we have included short descriptions of existing published AI governance standards at IEEE-SA, all available online through the Get Program, in addition to a brief summary of the IEEE ethical AI certification program.

### **1.10 IEEE STANDARDS ASSOCIATION (IEEE-SA)**

The Institute of Electrical and Electronic Engineers Standards Association is an operating unit within the IEEE (seven offices worldwide) that develops global standards in a broad range of industries, including power, energy, AI, internet of things, consumer technology, consumer electronics, biomedical, and health care, learning information technology and robotics, telecommunication, automotive, transportation, home automation, nanotechnology, information assurance, emerging



technologies, and many more. Collaborative thought leaders and experts from more than 175 countries assist with developing volunteer consensus-based standards, conformity assessments, and a variety of certifications (IEEE Standards Association, 2023b).

### **1.10.1 IEEE STANDARDS IN AI GOVERNANCE**

Since 2016, IEEE-SA has been developing and supporting Ethically Aligned Design (IEEE, 2018) and global trustworthy AI realization through human-centric standards and AI ethics certification. In partnership with leading entities committed to the advancement of responsible AI systems, the following IEEE AI governance standards are provided through the IEEE GET Program. Their summaries are also found in the webpages linked for each standard in the primary GET Program index (IEEE Standards Association, 2023a). The AI Ethics and Governance Standards in the following sections support AI ethics, governance, and literacy and they inform human-centric design, age-appropriate design, and AI systems governance standardization, ethical considerations for AI design, and AI ethics certification.

### **1.10.2 IEEE STANDARD FOR AN AGE-APPROPRIATE DIGITAL SERVICES FRAMEWORK BASED ON THE FIVE RIGHTS PRINCIPLES FOR CHILDREN – 2021**

A set of processes by which organizations seek to make their services age appropriate is established in this standard. The growing desire of organizations to design digital products and services with children in mind and that reflects their existing rights under the United Nations Convention on the Rights of the Child (the Convention) is supported by this standard. While different jurisdictions may have different laws and regulations in place, the best practice for designing digital services that impact directly or indirectly on children is offered in this standard.

### **1.10.3 IEEE STANDARD MODEL PROCESS FOR ADDRESSING ETHICAL CONCERNS DURING SYSTEM DESIGN – 2021**

A set of processes by which organizations can include consideration of ethical values throughout the stages of concept exploration and development is established by this standard. Management and engineering in transparent communication with selected stakeholders for ethical values elicitation and prioritization is supported by this standard, involving traceability of ethical values through an operational concept, value propositions, and value dispositions in the system design. Processes that provide for traceability of ethical values in the concept of operations, ethical requirements, and ethical risk-based design are described in the standard.

### **1.10.4 IEEE STANDARD FOR TRANSPARENCY OF AUTONOMOUS SYSTEMS – 2021**

Measurable, testable levels of transparency, so that autonomous systems can be objectively assessed, and levels of compliance determined, are described in this standard.

### **1.10.5 IEEE STANDARD FOR DATA PRIVACY PROCESS – 2022**

The requirements for a systems/software engineering process for privacy-oriented considerations regarding products, services, and systems utilizing employee, customer, or other external users' personal data are defined by this standard. Organizations and projects that are developing and deploying products, systems, processes, and applications that involve personal information are candidate users of the IEEE Std 7002™ standard. Specific procedures, diagrams, and checklists are provided for users of the IEEE Std 7002 standard to perform conformity assessments on their specific privacy practices. Privacy impact assessments (PIAs) are described as a tool for both identifying where privacy controls and measures are needed and for confirming they are in place.

### **1.10.6 IEEE STANDARD FOR TRANSPARENT EMPLOYER DATA GOVERNANCE – 2021**

Specific methodologies to help employers in accessing, collecting, storing, utilizing, sharing, and destroying employee data are described in this standard. Specific metrics and conformance criteria regarding these types of uses from trusted global partners and how third parties and employers can meet them are provided in this standard. Certification processes, success criteria, and execution procedures are not within the scope of this standard.

### **1.10.7 IEEE ONTOLOGICAL STANDARD FOR ETHICALLY DRIVEN ROBOTICS AND AUTOMATION SYSTEMS – 2021**

A set of ontologies with different abstraction levels that contain concepts, definitions, axioms, and use cases that assist in the development of ethically driven methodologies for the design of robots and automation systems is established by this standard. It focuses on the robotics and automation domain without considering any particular applications and can be used in multiple ways, for instance, during the development of robotics and automation systems as a guideline or as a reference “taxonomy” to enable clear and precise communication among members from different communities that include robotics and automation, ethics, and correlated area.

### **1.10.8 IEEE RECOMMENDED PRACTICE FOR ASSESSING THE IMPACT OF AUTONOMOUS AND INTELLIGENT SYSTEMS ON HUMAN WELL-BEING – 2020**

The impact of artificial intelligence or autonomous and intelligent systems (A/IS) on humans is measured by this standard. The positive outcome of A/IS on human well-being is the overall intent of this standard. Scientifically valid well-being indices currently in use and based on a stakeholder engagement process ground this standard. Product development guidance, identification of areas for improvement, risk management, performance assessment, and

the identification of intended and unintended users, uses and impacts on human well-being of A/IS are the intents of this standard.

1.10.9 IEEE CERTIFAIED™

IEEE CertifAIED™ is a certification program for assessing ethics of Autonomous Intelligent Systems (AIS) to help protect, differentiate, and grow product adoption.

1.11 INTRODUCING IEEE P3119, THE STANDARD FOR THE PROCUREMENT OF AI AND ADS

The IEEE P3119 standard is designed to provide normative and informative guidance and tools, in addition to reliable and repeatable processes that can help government agencies leverage the benefits of AI while mitigating the risks when seeking to procure systems the operate in high-risk domains. In Figure 1.6, a comparison to other AI governance guidance helps to distinguish its value in relation to others and also contextualize the standard’s components described in the next section. Among the various comparison elements in the left column are, for example, whether P3119 will require an audit, controls, model evaluation guidance, or model transparency. In comparison, P3119 checks all the boxes for these much-needed elements in AI guidance and tools.

To guide with standard tools that mitigate downside risk while leveraging strengths, the P3119 WG developed an international consensus-based process model devised to enhance and strengthen current and customary procurement practices. The P3119 standard includes five processes that help users identify, map, treat, and monitor risks commonly associated with high-risk AI systems. Each of the five

Comparing IEEE P3119 to Other AI Governance Sources							IEEE SA Consortium
Guidance	Voluntary	Voluntary	Voluntary	Voluntary	Enforceable	Enforceable	Voluntary
Publishing Entity	IEEE P3119	ISO 42001	ISO 23894	NIST AI RMF	EU AI Act	UK Procurement Act	UK Guidelines for Procurement
Requires an AI Audit	Yes, Process 5	Yes	No	No	Partial	No	No
Requires Organisational AI Policy	Yes, Processes 1 and 2	Yes	Yes	Yes	Partial	No	Partial
Provides AI Model Evaluation Guidance	Yes, Processes 2 and 3	No	No	Yes	Partial	No	No
Recommends AI Controls	Yes, Full Standard	Yes	No	No	Partial	No	Partial
Requires AI Risk Assessment	Yes, Processes 1-4	Yes	Yes	Yes	Yes	No	Partial
Requires AI Model Transparency	Yes, Processes 2 and 2	No	No	No	Yes	No	No
Requires AI Impact Assessment	Yes, Process 1	Yes	Yes	No	Partial	No	Yes
Requires AI Incident Reporting	Yes, Process 5	Yes	Yes	Yes	Yes	No	No
Provides Specific AI Procurement Tools	Yes, Full Standard	No	No	No	No	No	No

For updates on IEEE P3119: <https://standards.ieee.org/ieee/3119/10729/>    **Yes** = meets criteria, **Partial** = partially meets criteria, **No** = Does not meet criteria

FIGURE 1.6 Comparison of P3119 to other AI governance guidance.  
Source: Image by authors.

process steps within the standard includes a systematic pattern of information to help the reader easily understand how to navigate each process and achieve successful outcomes. This process component structure within each of the five processes includes the purpose of the process, the scope of the process, expected outcomes, the necessary activities and tasks, along with specific inputs and outputs to conduct the process. The purpose and intent of each component element are described, along with additional documentation that commonly aligns with certain components. Once we have established the structural foundation upon which each process is built, we will elaborate on the five processes that are included in the P3119 standard.

### 1.11.1 PROCESS ELEMENTS: STRUCTURE

**Purpose.** Each of the five processes in the standard starts with a purpose statement. Establishing the reason to conduct each process step provides the necessary focus for the user and is the beginning of boundary setting for the process.

**Scope.** The next component in each process step includes a statement about the scope of the process. The essential information found in this statement focuses specifically on the boundaries of the process and often provides explicit statements about what is excluded from the process.

**Outcomes.** In order to manage expectations, a brief statement on the outcomes is written in bullet point form listing the outcomes using action verbs. It is important to distinguish outcomes from outputs, which is also an informational element of each process. An outcome is an accomplishment that may or may not produce a more tangible output element (see Outputs).

**Activities.** This informational element is the instructional portion of each process. This is where all the required activities to successfully complete the process are carefully outlined.

**Tasks.** The tasks describe in detail how to accomplish each activity in the process.

**Appendices.** While the tasks should provide sufficient details explaining how to fulfill the process requirements, occasionally an appendix is also supplied with further information. Appendices are the tools that carry great utility and accessibility in form and function. They can include additional detailed guidance, external research, or resources, use cases for further clarity, scorecards, rubrics, and other types of templates that help facilitate the work at hand.

**Inputs.** Each process will require inputs to operationalize the activities and tasks. Inputs may include outputs from previous processes or other data, resources, files, or knowledge commonly available to the organization (e.g., a standard solicitation template, a contract template).

**Outputs.** The outputs defined for each process include a list of deliverables that will be produced by completing all the steps (activities and tasks). Resulting deliverables from the process may be used as inputs to subsequent process steps (e.g., risks mitigation *outputs* identified during the solution evaluation process will be used as *inputs* to the contract negotiations process).

1.11.2 IEEE P3119 FIVE KEY PROCESSES

Together, the purpose, scope, outcomes, activities, tasks, inputs, and outputs represent one process within the standard. As mentioned, the P3119 Standard includes five processes across four stages in total. The standard is meant to address the *additional risks and advantages* that AI presents because of the uniqueness and complexity within AI. However, it should be noted that the working group has taken extra care to make sure that the processes within the standard naturally align with *existing* procurement or tender processes. As aforementioned, this standard is not meant to replace any existing procurement processes but rather to enhance and augment well-established procurement stages and practices to address the new interrogation and assessment needs introduced by the emergence of AI-enabled technologies. Second, the team of experts drafting P3119 have also gone to great lengths to identify, and in many cases, develop useful guides, templates, and rubrics to ensure high-quality results are achieved. This library of tools means that anyone wishing to adopt the standard will have access to AI and ADS-oriented solicitation questions, along with evaluation guides and rubrics to help them interpret, and in some cases quantitatively score, vendors’ responses, for example. The five processes in the standard are depicted at an overview in Figure 1.7.

The P3119 working group conducted extensive global research and determined that there were five milestones within the procurement lifecycle that are critically important opportunities for leaders, program managers, and procurement teams to assess, identify, and manage AI risks effectively. These five points in the lifecycle

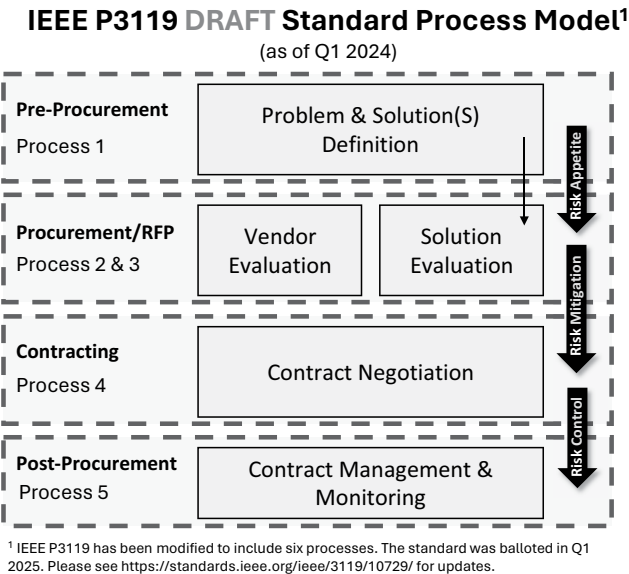


FIGURE 1.7 IEEE P3119 AI procurement draft standard process model.

Source: Image by authors.

include: (1) defining the problem and setting the solution requirements during pre-procurement, (2) evaluating vendors and (3) assessing solutions during the normal course of procurement processing, (4) negotiating with vendor(s) during the contracting phase, and (5) monitoring the AI solutions during the contract monitoring phase.

### 1.11.3 STAGE 1: PRE-PROCUREMENT (PRE-SOLICITATION/RFP BEFORE THE ISSUANCE OF BUDGET OR CALLING FOR PROPOSALS/BIDS)

**Problem Definition.** The working group discovered that problem definition was a commonly “skipped” step in the procurement lifecycle (Guszcza et al., 2020; Kuziemski & Misuraca, 2020; Sloane et al., 2021; Sloane & Chowdhury, 2021; WEF, 2020a). Research pointed to vendor interventions identifying solutions in search of problems, which often led to procurement actions taken without the benefit of establishing a fully defined and legitimate “business need” (Conti-Cook & Taraaz, 2023). Identifying the business need or problem – along with the source, depth, scope, and frequency of occurrence – is essential to ensure that AI specifically is even needed (Ada Lovelace Institute et al., 2021; Pahlka, 2023). If an AI or ADS solution were a relevant option that could truly improve the situation, then this stage will also require engaging an array of stakeholders to identify potential risks and harms and establish a risk appetite to guide the risk mitigation strategies and controls. As written earlier in the chapter, this is an additional reason why transdisciplinary collaboration would be optimal.

**Solution Requirements.** An additional step to be taken during the pre-procurement process is to establish a set of requirements for the vendors. It is critical for the solicitation requirements to align with responsible system design principles and practices (Miller & Waters, 2024). The activities and tasks in this process step also require an impact assessment as a final measure to evaluate the requirements against any unintended risks or harms. As noted earlier, several appendices are provided to facilitate the work at hand.

### 1.11.4 STAGE 2: PROCUREMENT/RFP/SOLICITATION

Stage 2 encompasses two processes: vendor and solution evaluation processes.

**Vendor Evaluation.** When publishing a procurement and soliciting vendors for an AI solution, specific questions about a vendor’s organizational AI governance practices should be included. The P3119 standard provides an appendix with specific questions, evaluation guidance for responses, and a scoring rubric. The activities and tasks in this process explain how to use the appendix to effectively evaluate the vendors and determine each vendor’s organizational AI governance maturity level. This enables the evaluators to easily identify any maturity gaps so the gaps can be mitigated with specific and time-bound mitigation tactics (e.g., Vendor A does not have a whistleblower policy; as a mitigation tactic, Vendor A must implement a whistleblower policy no later than 30 days after the contract start date).

**Solution Evaluation.** Evaluating an AI solution also requires a series of specific customs that are unique to AI risks and opportunities. Like other processes in the standard, the vendor evaluation process also includes an extensive appendix to

support the activities and tasks. The appendix includes an array of questions along with assessment guidance, exemplary response criteria, unsatisfactory response criteria, and additional references for further evaluation. This level of guidance enables a rigorous assessment process to help identify sound AI solutions, along with risks and potential mitigation tactics related to ethical choices, model designs, and other elements of the AI lifecycle. It should be noted that the working group has also aligned the questions with the NIST AI RMF 1.0 and other laws, regulations, and standards to help procurement teams easily see how and where compliance measures compare. Similar to Figure 1.6, the Comparison of P3119 to Other AI Governance Guidance, in the future the working group will be comparing select processes and process components to other related categories of guidance where relevant.

### 1.11.5 STAGE 3: CONTRACTING

**Vendor Negotiations.** Similar to defining the business problem prior to determining the need for an AI solution, the P3119 working group and complementary research supports that the strengths and weaknesses of AI contracting remain contested and requires additional attention (Fergusson, 2023; Miller & Waters, 2023; Sanchez-Graells, 2023, 2024a). Subsequently, working group members from the software and systems engineering fields, in addition to our procurement colleagues and legal experts, corroborated that traditional IT and software contract templates are commonly used but are not adequately addressing the nuances that AI systems present to government entities (Bertelsmann Stiftung, 2020; Mulligan & Bamberger, 2019).

As a result, this process within the standard provides a discussion of basic contract clauses that should be considered and used as reference points for negotiation with vendors to address common AI system risks. Undoubtedly, government agencies, organizations, and AI vendors will have their own sets of contracts requirements. However, the WG built consensus around the notion that business as usual in contracting has not adequately protected the public in the past. We understand that germane regulation (addressing AI-specific risks) and implementation guidance may take a while to catch up to protect the public from power asymmetries (Sanchez-Graells, 2024b). This lends credence to having reference contract language as negotiation first principles. Even though charters like the one from the EU on Fundamental Human Rights and others that are similar exist, AI continues to challenge good administration. The process of negotiating with the most promising vendor(s) is a critical opportunity to control the reference language, if at all possible, and subsequently the unique risks posed by any gaps in the vendor's (or vendors') organizational AI governance practices and/or proposed AI solution(s) (Miller & Waters, 2024). Our guidance provides a best practice guide to incorporating risk mitigation language into the negotiated contract(s).

### 1.11.6 STAGE 4: POST-PROCUREMENT

**Monitoring.** Once the contract (or contracts) has been established, the AI solution must be monitored. This is an imperative aspect of deploying an AI solution. The activities and tasks in this process identify key elements of contract monitoring

designed to control and manage known risks and continually assess the system for new and emerging risks. The process provides a best practice guide that outlines the need for key performance indicators, metrics, and parameters that define the risk tolerances that were agreed to within the negotiated contract, as well as any actions that are required if the upper bounds of a tolerance (appetite) metric is exceeded (Carmichael, 2022).

### **1.11.7 ALIGNMENT WITH RISK MANAGEMENT PRACTICES**

Although AI has many advantages, the P3119 standard leans toward a risk management approach similar to the standard of management of information security, cybersecurity, and privacy risks (ISO/IEC, 2022). Referring to Figure 1.7, the standard incorporates several risk management principles woven through the process model. These are noted in the figure on the right side as arrows moving down through each step. The first process in the standard incorporates the development of a risk appetite, which is designed to set the bar for risk identification and corresponding mitigation tactics. The second risk management element involves risk assessment. Assessing risks occurs during the vendor and solicitation evaluation processes. Through those evaluations, risks are identified and mitigation tactics are mapped. The third risk management element requires risk control for the purpose of achieving an acceptable risk tolerance. Controlling the risks through the P3119 standard primarily occurs within the contract negotiation process. More specifically, contract terms and conditions are used to codify the mitigation tactics. In addition, the contract negotiation process also incorporates guidance related to monitoring the risk mitigation tactics once deployed. Any risk mitigation metrics that exceed the risk tolerance can then be appropriately redressed so that the risk tolerance is brought back into an acceptable range.

P3119 is only a first step towards raising the standard in the procurement of AI and ADS to protect the public interest. To serve its promise it can be used as an instrument that advocates for the public through its normative guidance and tools. But transdisciplinary collaboration teams, sandboxes, and upskilling of procurement capacity are all still needed. For future success in reducing risks and leveraging the benefits of AI and ADS with “good administration” that fulfills the fiduciary duty described herein, basic research in the complications and complexity of the social systems is required. For organizations, government agencies, and vendors to adapt to AI specific requirements effectively, more greenfields of standards implementation, education, training, policy deliberation, and practice need to be explored in the future. Building a ready, willing, and able community of practice in AI procurement is a sound next step complementing this new category of international consensus-based standard such as P3119. These gaps and missing value are why we founded the AI Procurement Lab.

### **1.12 FOUNDING THE AI PROCUREMENT LAB**

Recognizing the need for responsible procurement practices when sourcing high-risk AI systems, Waters and Miller have recently launched the AI Procurement Lab (AIPL). For all the reasons mentioned in this chapter, the founders acknowledged that procuring high-risk AI is materially and critically different from buying traditional



IT solutions, and a community of practice of likeminded stakeholders that care about the impact to the public needed to be formed. Although high-risk AI offers new opportunities and efficiencies, these solutions also pose novel risks and liabilities that continue to emerge at a rapid pace.

To practice responsible AI procurement, novel better approaches are required to address this evolving risk landscape. Through collaboration, discovery, and upskilling, the founders aim to ensure that AI procurement procedures appropriately (1) identify, (2) evaluate, (3) mitigate, and (4) monitor AI/ADS opportunities *and* risks. The core mission of the AIPL is to bring practitioners together to advance responsible procurement practices that not only take advantage of the benefits of AI but also protect society from its potential harms. The founders believe that this can only occur by drawing upon transdisciplinary collaboration and research (OECD, 2020), human-centered design practices (Nagitta et al., 2022; Naudé & Dimitri, 2021) and applying responsible AI principles across the entire procurement lifecycle (Autio et al., 2023; Demircioglu & Vivona, 2021; Hickok, 2022) – from needs assessments (pre-solicitation/pre-procurement) to system decommissioning – because every step in the life cycle matters.

The AIPL researches and develops best practices, guidance, and tools that foster responsible procurement of high-risk artificial intelligence and automated decision systems. We offer procurement professionals and organizations opportunities to develop skills, benchmark, and adopt research-based resources that ensure responsible procurement of high-risk artificial intelligence solutions.

### 1.13 ACKNOWLEDGMENT: RESOLUTE VOLUNTEERS AND STANDARD WRITERS

Raising a new category of standard in AI procurement is a transdisciplinary marathon combining imagination, art, and science. We start with many questions that lead to even more questions. International consensus building on the project authorization and the standard's process model (Waters, 2021) began with forty-eight working group members across twelve countries. Many of these people contributed their time during monthly meetings in the first year of development, discussing various iterations of processes in the process model itself. Still others offered initial drafting and built consensus around the early versions of the standard's uniform set of definitions. Today, consensus building continues with dedicated writers (across six countries) diligently drafting each process component and building out a set of related practical administrative tools, including weighted rubrics and scoring guides for how to evaluate AI vendors and their solutions.

Making it possible to offer our standard's guidance and tools *in advance* of publication to our potential sandbox partners are a select few resolute volunteers that should be acknowledged for their collaboration efforts. At every meeting, these individuals in a highly collaborative group continue to push on assumptions and expand group thinking on the subject matter. Without their dedication and energy, development would be impossible. Their relentless grit and participation in this greenfield category also help to make the marathon immensely rewarding. The following members continue to draft, read, review, and refine the current versions of the standard's normative guidance and tools:

Dr. Gisele Waters, USA, Dr. Cari Miller, USA, Andrew Gamino-Cheong, USA, Grant Fergusson, USA, Roya Pakzad, USA, Iran, Richard Moreno, USA, Sara Soubelet, Argentina, Clara Clemente Langevin, Brazil, Cristina Muresan, UK, Mana Sadeghipour, UK, Maria Paz Hermosilla, Chile

Special acknowledgement goes to two generous parties: Ruth Lewis, the Chair of the IEEE Society for Implications of Technology Standards Committee, who graciously shared the P3119 value in Japan and other global conferences, and to Intellera Consulting SrL in Italy, who introduced us to the EFSA personnel and are guiding us in relationship management. We would like to express our sincerest gratitude to all mentioned who have opened the doors to conversation, learning, and supporting the development of this nascent field that is responsible AI procurement.

## REFERENCES

- Acemoğlu, D. (2023). Harms of AI. In J. B. Bullock and others (Eds.), *The Oxford handbook of AI governance* (pp. C65P1–C65N5). <https://doi.org/10.1093/oxfordhdb/9780197579329.013.65>
- Ada Lovelace Institute, AI Now, & Open Government Partnership. (2021). *Algorithmic accountability for the public sector*. [Opengovpartnership.Org](https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/). <https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/>
- Alms, N. (2024). *IRS has ‘unconscionable delays’ in helping identity theft victims, taxpayer advocate says*. NEXTGOV FCW. <https://www.nextgov.com/digital-government/2024/01/irs-has-unconscionable-delays-helping-identity-theft-victims-taxpayer-advocate-says/393294/?oref=ng-home-top-story>
- Autio, C., Cummings, K., Elliott, B. S., & Noveck, B. S. (2023, June). *A snapshot of AI procurement challenges: Diagnosing perceived and actual risks impeding responsible AI acquisition in government*. TheGovLab.
- Benthall, S., & Shekman, D. (2022/2023). Designing fiduciary artificial intelligence. *ACM International Conference Proceeding Series*. <https://doi.org/10.1145/3617694.3623230>
- Bersani, C., Codagnone, J., David, L., Foiniotis, A., Galasso, G., Mancini, S., . . . Pellegrino, M. (2022). Roadmap for actions on artificial intelligence for evidence management in risk assessment. *EFSA Supporting Publications*, 19(5). <https://doi.org/10.2903/sp.efsa.2022.en-7339>
- Bertelsmann Stiftung. (2020). *Automating society report 2020 Switzerland*. AlgorithmWatch GmbH.
- Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R., Weitkamp, E., & Winfield, A. (2020). The ethics of artificial intelligence: Issues and initiatives, panel for the future of science and technology, scientific foresight unit (STOA). *European Parliamentary Research Service*, PE, 634–452. <https://doi.org/10.2861/6644>
- Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data and Society*, 8(1). <https://doi.org/10.1177/2053951720983865>
- Buolamwini, J. (2022). The IRS should stop using facial recognition. *The Atlantic*. <https://www.theatlantic.com/ideas/archive/2022/01/irs-should-stop-using-facial-recognition/621386/>
- Burris, S. (2022). *Transdisciplinary integration: The only way forward for public health*. Petrie-Flom Center Harvard Law School: Bill of Health. <https://blog.petrieflom.law.harvard.edu/2022/03/31/transdisciplinary-integration-public-health/>
- Cagnoni, S., Emiliani, V., Lombardo, G., Alkema, W., Hooijmans, C., Alkema, S., Novotny, T., Hair, K., Nic, M., Macleod, M., Bannach-Brown, A., van Beuningen, N., Hijlkema, N., & Wever, K. (2023). Implementing AI vertical use cases - scenario 1. *EFSA Supporting Publications*, 20(8), 8223E. <https://doi.org/10.2903/sp.efsa.2023.en-8223>

- Carmichael, M. (2022). *Risk appetite vs risk tolerance: What is the difference*. Information Systems and Control Audit Association ISACA. <https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2022/risk-appetite-vs-risk-tolerance-what-is-the-difference>
- Casovan, A., & Shankar, V. (2022). A risk-based approach to AI procurement. *The Regulatory Review: A Publication of the Penn Program on Regulation*. <https://scholarship.law.upenn.edu/regreview-opinion/39/>
- Chakravorti, B., Bhalla, A., Chaturvedi, R. S., & Filipovic, C. (2021). 50 global hubs for top AI talent. *Harvard Business Review*. <https://hbr.org/2021/12/50-global-hubs-for-top-ai-talent>
- Chen, Y. C., & Ahn, M. J. (2017). Routledge handbook on information technology in government. In *Routledge handbook on information technology in government*. <https://doi.org/10.4324/9781315683645>
- Cihon, P. (2019, April). *Standards for AI governance: International standards to enable global coordination in AI research & development*. Future of Humanity Institute, University of Oxford.
- CLS. (2023). *Fiduciary duty*. Cornell Law School Legal Information Institute. [https://www.law.cornell.edu/wex/fiduciary\\_duty](https://www.law.cornell.edu/wex/fiduciary_duty)
- Colvin, T. J., Liu, I., Babou, T. F., & Wong, G. J. (2020). *A brief examination of Chinese government expenditures on artificial intelligence R&D*. IDA Science & Technology Policy Institute. <https://www.ida.org/-/media/feature/publications/a/ab/a-brief-examination-of-chinese-government-expenditures-on-artificial-intelligence-r-and-d/d-12068.ashx>
- Commonwealth of Australia. (2023). *Report to the royal commission into the Robodebt scheme*. [robodebt.royalcommission.gov.au/](https://robodebt.royalcommission.gov.au/)
- Commonwealth Ombudsman. (2019). *Automated decision making better practice guide*. Ombudsman.Gov.Au. [https://www.ombudsman.gov.au/\\_\\_data/assets/pdf\\_file/0030/109596/OMB1188-Automated-Decision-Making-Report\\_Final-A1898885.pdf](https://www.ombudsman.gov.au/__data/assets/pdf_file/0030/109596/OMB1188-Automated-Decision-Making-Report_Final-A1898885.pdf)
- Conti-Cook, C., & Taraaz. (2023). *A guiding framework to vetting public sector technology vendors*. Ford Foundation. <https://www.fordfoundation.org/work/learning/research-reports/a-guiding-framework-to-vetting-public-sector-technology-vendors/>
- Day One Project. (2022). *Innovating with procurement: Solving market failures & creating industries*. Day One Project. <https://uploads.dayoneproject.org/2022/02/14125252/Market-Shaping-Primer.pdf>
- Deloitte US. (2023). *2023 Global chief procurement officer (CPO) survey: Orchestrators of value*. Deloitte Insights. <https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-2023-global-chief-procurement-officer-survey.pdf>
- Demircioglu, M. A., & Vivona, R. (2021). Positioning public procurement as a procedural tool for innovation: An empirical study. *Policy and Society*, 40(3), 379–396. <https://doi.org/10.1080/14494035.2021.1955465>
- Edquist, C., & Hommen, L. (2000). Public technology procurement and innovation theory. In *Economics of science, technology, and innovation (ESTI)* (V16 ed., pp. 5–70). Boston, MA: Springer. [https://doi.org/10.1007/978-1-4615-4611-5\\_2](https://doi.org/10.1007/978-1-4615-4611-5_2)
- Edwards, L. (2022). *The EU AI Act: A summary of its significance and scope*. Ada Lovelace Institute. <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf>
- EFSA. (2010). Application of systematic review methodology to food and feed safety assessments to support decision making. *EFSA Journal*, 8(6). <https://doi.org/10.2903/j.efsa.2010.1637>
- EPIC. (2023). *Comments of the electronic privacy information center to the office of management and budget*. Electronic Privacy Information Center. [https://downloads.regulations.gov/OMB-2023-0020-0140/attachment\\_1.pdf](https://downloads.regulations.gov/OMB-2023-0020-0140/attachment_1.pdf)
- EU Agency for Fundamental Rights. (2009). *European charter for fundamental rights 2017*. FRA. Europa.EU. <https://fra.europa.eu/en/eu-charter/article/47-right-effective-remedy-and-fair-trial>

- Eubanks, V. (2018). *Automating inequality how high-tech tools profile, police, and punish the poor*. Picador/St. Martin's Press.
- European Law Institute. (2022). *Model rules on impact assessment of algorithmic decision-making systems used by public administration*. Universitat Wein. [https://www.europeanlawinstitute.eu/fileadmin/user\\_upload/p\\_eli/Publications/ELI\\_Model\\_Rules\\_on\\_Impact\\_Assessment\\_of\\_ADMSs\\_Used\\_by\\_Public\\_Administration.pdf](https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Model_Rules_on_Impact_Assessment_of_ADMSs_Used_by_Public_Administration.pdf)
- Executive Office of the President of the United States. (2023). Federal register: Safe, secure, and trustworthy development and use of artificial intelligence. In *National archives federal register the daily journal of the USG*. National Archives.
- Fallon, A. (2023). *What's the difference between AI and IT automation?* Tech Target IT Operations. <https://doi.org/10.21608/jstc.2023.291249>
- Farrell, K. N. (2011). Tackling wicked problems through the transdisciplinary imagination. *Journal of Environmental Policy & Planning*, 13(1), 75–77. <https://doi.org/10.1080/1523908x.2011.557901>
- Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larriex, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data and Society*, 6(1), 1–14. <https://doi.org/10.1177/2053951719860542>
- Fergusson, G. (2023). *Outsourced and automated: How AI companies have taken over government decision-making*. Electronic Privacy Information Center. <https://epic.org/wp-content/uploads/2023/09/FINAL-EPIC-Outsourced-Automated-Report-w-Appendix-Updated-9.26.23.pdf>
- Fontenot, L., & Gaedt-Sheckter, C. (2020). *Fiduciary duty considerations for boards of companies using AI*. GibsonDunn Law360. <https://www.gibsondunn.com/wp-content/uploads/2020/01/Fontenot-Gaedt-Sheckter-Fiduciary-Duty-Considerations-For-Boards-Of-Cos.-Using-AI-Law360-1-3-2020.pdf>
- Guimarães da Costa, N. (2021). Transdisciplinary collaborations for achieving the SDGs. In W. Leal Filho, A. Marisa Azul, L. Brandli, A. Lange Salvia, & T. Wall (Eds.), *Partnership for the goals. Encyclopedia of the UN sustainable development goals* (pp. 1291–1306). Cham: Springer. [https://doi.org/10.1007/978-3-319-95963-4\\_138](https://doi.org/10.1007/978-3-319-95963-4_138)
- Guszcza, J., Lee, M., Ammanath, B., & Kuder, D. (2020). Human values in the loop: Design principles for ethical AI (Deloitte Insights). *Deloitte Review: Technology and Ethics*, (26). [www2.deloitte.com/us/en/insights/focus/cognitive-technologies/design-principles-ethical-artificial-intelligence.html](http://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/design-principles-ethical-artificial-intelligence.html)
- HAAS School of Business. (2019). *Play 1. Enable diverse and multi-disciplinary teams working on algorithms and AI systems*. HAAS.Berkeley.Edu. [https://haas.berkeley.edu/wp-content/uploads/EGAL\\_Playbook\\_Play1\\_Teams.pdf](https://haas.berkeley.edu/wp-content/uploads/EGAL_Playbook_Play1_Teams.pdf)
- Harris, I., Mainelli, M., & Jones, H. (2008). Caveat Emptor, Caveat Venditor: Buyers & sellers beware the tender trap. *Journal of Strategic Change*, 17, 1–9. <https://doi.org/10.1201/9781420056655-12>
- Hickok, M. (2022). Public procurement of artificial intelligence systems: New risks and future proofing. *AI and Society*. <https://doi.org/10.1007/s00146-022-01572-2>
- Hocking, V. T., Brown, V. A., & Harris, J. A. (2016). Tackling wicked problems through collective design. *Intelligent Buildings International*, 8(1), 24–36. <https://doi.org/10.1080/17508975.2015.1058743>
- Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. INDEPENDENTLY PUBLISHED.
- Hooker, J. (2018). Ethics of artificial intelligence. *Taking Ethics Seriously*, 211–219. <https://doi.org/10.4324/9781315097961-14>
- HSGAC. (2024). *Harnessing AI to improve government services and customer experience – committee on homeland security & governmental affairs*. HSGAC.Senate.Gov.

- IDC. (2022). *Worldwide spending on AI-centric systems will pass \$300 billion by 2026, according to IDC*. International Data Corporation. <https://www.idc.com/getdoc.jsp?containerId=prUS49670322>
- IEEE. (2018). IEEE standard review – Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)* (pp. 197–201). IEEE. <https://ieeexplore.ieee.org/document/9398613>
- IEEE. (n.d.). *IEEE P3119 standard for the procurement of artificial intelligence and automated decision systems*. Standards Association.
- IEEE Standards Association. (2023a). *GET program for AI ethics and governance standards*. IEEE. <https://ieeexplore.ieee.org/browse/standards/get-program/page/series?id=93>
- IEEE Standards Association. (2023b). *IEEE standards association*. Standards.IEEE.Org. <https://standards.ieee.org/>
- IEEE Standards Association. (2024). *IEEE standards association, developing standards, an introduction*. Standards.IEEE.Org. <https://standards.ieee.org/develop/>
- IEEE-USA Position Statement. (2020, July). *Artificial intelligence: Accelerating inclusive innovation by building trust*. <https://ieeusa.org/assets/public-policy/committees/aipcl/AITrust0720.pdf>
- ISO/IEC. (2022). *ISO/IEC 27005:2022(en) information security, cybersecurity and privacy protection – guidance on managing information security risks*. International Organization for Standardization. <https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27005:ed-4:v1:en>
- Ivanova, Y. (2021). *Regulatory sandboxes in the artificial intelligence act rationale for regulatory sandboxes in the AI proposal*. European Commission CNECT A2.
- Jelinek, T., Wallach, W., & Kerimi, D. (2021). Policy brief: The creation of a G20 coordinating committee for the governance of artificial intelligence. *AI and Ethics*, 1(2), 141–150. <https://doi.org/10.1007/s43681-020-00019-y>
- Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44(6). <https://doi.org/10.1016/J.TELPOL.2020.101976>
- Law Commission of Ontario. (2021, April). Regulating AI: Critical issues and choices. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3853249>
- Leydesdorff, L. (2005). Anticipatory systems and the processing of meaning: A simulation study inspired by Luhmann’s theory of social systems. *Journal of Artificial Societies and Social Simulation*, 8(2).
- Luhmann, N. (1982). The world society as a social system. *International Journal of General Systems*, 8(3), 131–138. <https://doi.org/10.1080/03081078208547442>
- Madiega, T. (2023, June). *BRIEFING – EU legislation in progress – artificial intelligence act*. EPRS | European Parliamentary Research Service.
- Maksimainen, J. (2011). *Aspects of values in human-technology interaction design: A content-based view to values*. Jyväskylä Studies in Computing, 144.
- Marinotti, J. (2021, Spring). Article 3 part of the intellectual property law commons | tangibility as technology. *Georgia State University Law Review*, 37, 671–738.
- Martin, A., & Balestra, G. (2019). Using regulatory sandboxes to support responsible innovation in the humanitarian sector. *Global Policy*, 10(4), 733–736. <https://doi.org/10.1111/1758-5899.12729>
- Mashaw, J. L. (2007). Reasoned administration: The European Union, the United States, and the project of democratic governance. *George Washington Law Review*, 76(1), 99–124.
- Matsuo, Y. (2017). *The Japanese society for artificial intelligence ethical guidelines*. Japanese Society for Artificial Intelligence (JSAI).
- Miller, C., & Waters, G. (2023). *AI procurement: Essential considerations in contracting*. Inclusive Change. <https://www.inclusivechange.org/ai-governance-solutions/ai-contract-clauses>

- Miller, C., & Waters, G. (2024). *Risk management framework for the procurement of AI systems (RMF PAIS 1.0)*. The Center for Inclusive Change.
- Mökander, J., & Axente, M. (2023). Ethics-based auditing of automated decision-making systems: Intervention points and policy implications. *AI and Society*, 38(1), 153–171. <https://doi.org/10.1007/s00146-021-01286-x>
- Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics*, 27(4), 1–30. <https://doi.org/10.1007/s11948-021-00319-4>
- Mulligan, D. K., & Bamberger, K. A. (2019). Procurement as policy: Administrative process for machine learning. *Berkeley Technology Law Journal*, 34. <https://doi.org/10.2139/ssrn.3464203>
- Nagitta, P. O., Mugurusi, G., Obicci, P. A., & Awuor, E. (2022, April). Human-centered artificial intelligence for the public sector: The gate keeping role of the public procurement professional. *Procedia Computer Science*, 200, 1084–1092. <https://doi.org/10.1016/j.procs.2022.01.308>
- Naudé, W., & Dimitri, N. (2021). Public procurement and innovation for human-centered artificial intelligence. *SSRN Electronic Journal*, 14021. <https://doi.org/10.2139/ssrn.3762891>
- Nihill, C. (2024). *Ex-White House official says congress, federal agencies should do more for AI talent search*. Fedcoop.Com. <https://fedcoop.com/ex-white-house-official-says-congress-federal-agencies-should-do-more-for-ai-talent-search/>
- Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, 2(1), 25–42. <https://doi.org/10.1007/BF02639315>
- Noveck, B. S. (2021). Solving public problems: A practical guide to fix our government and change our world. *Solving Public Problems: A Practical Guide to Fix Our Government and Change Our World*, 1–449.
- NYC. (2018). *City of New York local law 49*. NYC Legista. <https://nyc.legistar1.com/nyc/attachments/f97d9c19-cffa-4411-8773-14f5f5730be1.pdf>
- NYC. (2023). *NCY automated decision systems task force*. NYC.Gov. <https://www.nyc.gov/site/adstaskforce/members/members.page>
- OECD. (2019). *OECD/G20 AI principles*. OECD AI Policy Observatory. <https://oecd.ai/en/work/documents/g20-ai-principles>
- OECD. (2020). *Addressing societal challenges using transdisciplinary research*. STI Policy Papers. <https://www.oecd-ilibrary.org/deliver/0ca0ca45-en.pdf?itemId=/content/paper/0ca0ca45-en&mimeType=pdf>
- OECD. (2021a). *AI index report: Policy and national strategies chapter 7*. OECD AI Policy Observatory. [https://wp.oecd.ai/app/uploads/2021/03/2021-AI-Index-Report\\_Chapter-7.pdf](https://wp.oecd.ai/app/uploads/2021/03/2021-AI-Index-Report_Chapter-7.pdf)
- OECD. (2021b). *Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems*. OECD Publishing Digital Economy Papers. [https://www.oecd-ilibrary.org/science-and-technology/tools-for-trustworthy-ai\\_008232ec-en](https://www.oecd-ilibrary.org/science-and-technology/tools-for-trustworthy-ai_008232ec-en)
- OECD. (2023). *OECD definition of public procurement*. Directorate for Public Governance.
- OECD AI Policy Observatory. (2023). *Updates to the OECD's definition of an AI system explained, intergovernmental*. The AI Wonk. <https://oecd.ai/en/work/ai-system-definition-update>
- OECD Data Explorer. (2016). *Access the data: About the data on public procurement*. OECD Public Procurement. [https://qdd.oecd.org/subject.aspx?Subject=GOV\\_PUBPRO\\_2016](https://qdd.oecd.org/subject.aspx?Subject=GOV_PUBPRO_2016)
- OECD Federal Ministry of Labour and Social Affairs. (2022). *Framework for the classification of AI systems*. OECD Publishing Digital Economy Papers. <https://doi.org/10.1787/cb6d9eca-en>
- OECD Federal Ministry of Labour and Social Affairs. (2023). *Common guideposts to promote interoperability in AI risk management*. OECD Publishing AI Papers.

- <https://www.oecd-ilibrary.org/deliver/ba602d18-en.pdf?itemId=/content/paper/ba602d18-en&mimeType=pdf>
- Pahlka, J. (2023). *Recoding America*. Metropolitan Books.
- Poli, R. (2013). A note on the difference between complicated and complex social systems. *CADMUS Journal*, 2(1), 142–147.
- Pop, F., & Adomavicius, L. (2021). *Sandboxes for responsible artificial intelligence*. EIPA Briefing. <https://search.ebscohost.com/login.aspx?direct=t>
- ProcurementMag. (2023). *Procurement: Top 100 CPO's*. ISSUU. <https://procurementmag.com/magazine/top-100-cpos-2023>
- Richardson, R. (2019, December). *A shadow report of the New York City automated decision system task force*. AI Now Institute.
- Richardson, R. (2022). Defining and demystifying automated decision systems. *Maryland Law Review*. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3811708](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3811708)
- Rimes, T. (2021). *Trends for the future: Public procurement professionals adapt to the changing and challenging times ahead*. American City and Country. <https://www.americancityandcountry.com/2021/12/21/trends-for-the-future-public-procurement-professionals-adapt-to-the-changing-and-challenging-times-ahead/>
- Rioux, N. (2020). *Twenty-third annual report on federal agency use of voluntary consensus standards and conformity assessment activities*. NISTIR 8329.
- Robbins, S. (2019). AI and the path to envelopment: Knowledge as a first step towards the responsible regulation and use of AI-powered machines. *AI & Society*. <https://doi.org/10.1007/s00146-019-00891-1>
- Rubenstein, D. S. (2021, forthcoming). Acquiring ethical AI. *Florida Law Review*, 73, 1–61.
- Sanchez-Graells, A. (2023b). *Can the government just go and 'confidently and responsibly' buy artificial intelligence?* University of Bristol Law School Blogs. <https://legalresearch.blogs.bris.ac.uk/2023/05/can-the-government-just-go-and-confidently-and-responsibly-buy-artificial-intelligence/>
- Sanchez-Graells, A. (2024a). *How to crack a nut, A blog on EU economic law*. Howtocrackanut.Com. <https://www.howtocrackanut.com/>
- Sanchez-Graells, A. (2024b). Resh(AD)ping good administration: Addressing the mass effects of public sector digitalisation. *Laws*, 13(1), 9.
- Scott, P. G. (1997). Assessing determinants of bureaucratic discretion: An experiment in street-level decision making. *Journal of Public Administration Research and Theory*, 7(1), 35–57. <https://doi.org/10.1093/oxfordjournals.jpart.a024341>
- Sieber, R., & Brandusescu, A. (2021). *Final report. Civic empowerment in the development and deployment of AI systems. Critiquing and rethinking accountability, fairness, and transparency workshop at ACM's FAccT. FAccT CRAFT*.
- Singh, J. P., Shehu, A., Wesson, C., & Dua, M. (2023). *Global AI infranstructures report*. George Mason University AI Strategies Team & Stimson Center.
- Sloane, M., & Chowdhury, R. (2021). *Procuring & embedding AI systems in the public sector*. Carnegie Council AI & Equality Initiative. <https://www.carnegiecouncil.org/media/series/aiei/20211006-procuring-embedding-ai-systems-public-sector-mona-sloane-rumman-chowdhury>
- Sloane, M., Chowdhury, R., Havens, J. C., Lazovich, T., & Alba, L. C. R. (2021). *AI and procurement: A primer*. New York Univeristy. <https://doi.org/10.17609/bxzf-df18>
- Statista. (2023). *Government IT spending by segment 2019–2023*. Technology & Telecommunications: Software. <https://www.statista.com/statistics/1154210/worldwide-government-it-spending-forecast-by-segment/>
- Stoyanovich, J., & Barocas, S. (2019). *Testimony regarding update on local law 49 of 2018 in relation to automated decision systems (ADS) used by agencies before NYC council communication on technology*. NCY Council & Committee on Technology.
- Taele, C. (2023). *AI-ready New York City focuses on workforce, procurement*. Route 50 Connecting State and Local Government Readers. <https://www.route-fifty.com/emerging-tech/2023/10/ai-ready-new-york-city-focuses-workforce-procurement/391472/>

- Trustible.ai. (2023). *Towards a standard for model cards*. Trustible.Ai/Post. <https://www.trustible.ai/post/towards-a-standard-for-model-cards>
- Turner, J. R., & Baker, R. M. (2019). Complexity theory: An overview with potential applications for the social sciences. *Systems*, 7(1), 7823–7830. <https://doi.org/10.3390/systems7010004>
- United Kingdom, Procurement Act 2023 Chapter 54, Parliamentary Bills. (2023). <https://www.legislation.gov.uk/ukpga/2023/54/contents>
- US Census Bureau. (2022). *Quick facts: New York City*. Census.Gov. <https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045222>
- US GAO. (2021, June). *Framework for federal agencies and other entities*. <https://www.gao.gov/products/gao-21-519sp>
- Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU artificial intelligence act – analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4), 97–112. <https://doi.org/10.9785/crl-2021-220402>
- Wan, S., & Sieber, R. (2023). Artificial intelligence (AI) adoption in Canadian municipalities: In-house development versus outsourcing. *18th International Conference on Computational Urban Planning and Urban Management CUPUM*, 1–16. <https://doi.org/10.17605/OSF.IO/6YR5V>
- Waters, G. A. (2021). *IEEE SA – process model and requirements aimed at AI procurement in a new IEEE standard*. IEEE SA Beyond Standards. <https://standards.ieee.org/beyond-standards/process-model-and-requirements-aimed-at-ai-procurement-in-a-new-ieee-standard/>
- WEF. (2020a). *Unlocking public sector AI AI procurement in a box: Challenges and opportunities during implementation*. World Economic Forum. <https://www.weforum.org/publications/ai-procurement-in-a-box/challenges-and-opportunities-during-implementation/>
- WEF. (2020b). *Unlocking public sector AI procurement in a box: AI government procurement guidelines*. World Economic Forum. <https://www.weforum.org/publications/ai-procurement-in-a-box/ai-government-procurement-guidelines/#report-nav>
- Weissinger, L. B. (2022). AI, complexity, and regulation. In J. B. Bullock and others (Eds.), *The Oxford handbook of AI governance*. <https://doi.org/10.1093/oxfordhb/9780197579329.013.66>
- Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019, January). The role and limits of principles in AI ethics: Towards a focus on tensions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195–200). <https://dl.acm.org/doi/abs/10.1145/3306618.3314289>
- World Economic Forum. (2021). *Unlocking public sector artificial intelligence WEF*. AI and Machine Learning. <https://www.weforum.org/projects/unlocking-public-sector-artificial-intelligence/>
- Xia, B., Lu, Q., Zhu, L., Lee, S. U., Liu, Y., & Xing, Z. (2023). *From principles to practice: An accountability metrics catalogue for managing AI risks*. CoRR.
- Yang, K. (2016). Creating public value and institutional innovations across boundaries: An integrative process of participation, legitimation, and implementation. *Public Administration Review*, 76(6), 873–885. <https://doi.org/10.1111/puar.12561>
- Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Conference on human factors in computing systems – proceedings* (pp. 1–12). <https://doi.org/10.1145/3290605.3300509>
- Young, S. D. (2023). *Proposed memorandum for the heads of executive departments and agencies*. Executive Office of the President, OMB. <https://ai.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-Public-Comment.pdf>



---

# Appendix 1.1 Example of High-Risk AI Applications and Systems

Category	High-Risk AI Applications and Systems
Education	Targeting advertisements, determining access, predicting achievement, evaluate learning outcomes, autonomous test proctoring, AI-driven curriculum delivery, AI-augmented classrooms, AI-recommended learning paths, AI-driven assessments, emotional state detection
Employment	Recruitment, hiring, candidate scoring/ranking, targeted job advertising, skills scraping/assessment, AI-driven interviewing, AI-driven assessments, task allocation, quota setting, automated scheduling, performance monitoring, behavior assessment/monitoring, promotion determination, pay determinations, career path recommendations, succession planning, discipline determination, termination, nudges, emotional state detection
Healthcare	Medication, hospitals, doctors, diagnostics, drug discovery and distribution, family planning, patient care, preventative services, wearables, mental health chatbots
Financial Services	Access to credit, credit scores, background checks, insurance, loans, mortgages, interest, and policy rate fairness/equity
Housing	Background checks, eligibility, affordability, rent controls
Government Benefits	Benefits eligibility (grant, reduce, revoke, or reclaim), e.g., welfare, healthcare, social security, HeadStart
Public Services	Dispatching of emergency first response services, density/placement/availability of emergency and other public services
Critical Infrastructure	Transportation, communications, emergency services, healthcare, safe food
Essential Utilities	Electric, water, gas, communications
Law Enforcement	Polygraphs, deep fake detection, crime analytics (identifying unknown patterns, hidden relationships, fact interpretation), emotional state detection
Justice and Legal	Recidivism scoring, sentencing determinations, probation risk assessments
Immigration	Risk assessment (security, irregular immigration, health), travel document and supporting document verification, application verification (asylum, visa, residence permits), eligibility checking (asylum, visa, residence permits), emotional state detection
Biometric Identification	Security access points, facial recognition, voice and language processing, speech to text, retina scan, fingerprint scan, DNA swabs, emotional state detection
Safety Components	Autonomous vehicles, autonomous drones, HOV lane monitoring, supply of water/gas/electricity monitoring, AI-driven surgery components

Source: Adapted from <https://www.euaiact.com/annex/3>.

---

---

# 2 Data Empowerment and Protection Architecture (DEPA) for Training Machine Learning (ML) Models

*Shyam Sundaram, Kapil Vaswani, Gaurav Agarwal, Sunu Engineer, and AVS Sridhar*

## 2.1 INTRODUCTION

One of the most significant benefits of AI models and large language models (LLMs) is their ability to help solve societal problems. The potential applications of LLMs are vast and varied. They can be used for text completion, machine translation, text summarization, question answering, and creating chatbots that can hold conversations with humans. LLMs can also be trained on diverse types of data, including code, images, audio, video, and more. LLMs can be used to analyze social media data to identify and track the spread of misinformation and hate speech. They can also be used to analyze medical records to identify patterns and trends that can help improve patient outcomes.

The use of large language models (LLM) has revolutionized the field of overall data sciences and in specific natural language processing (NLP) technologies. LLMs are statistical language models that are trained on massive amounts of data and can be used to generate and translate text and other content, as well as perform other NLP tasks. LLMs are typically based on deep learning architectures, their foundation derived from the famous Google paper (Vaswani et al., 2017).

With the powerful capabilities for relevant cases, there are lots of risks also being created, such as misinformation, fake news and privacy, and many more. This urgent requirement for addressing this situation is globally understood. We cover this topic with a relevant and fictitious case study to drive this technical framework.

## 2.2 CORE FOUNDATION FOR DEPA TRAINING

Data empowerment and protection architecture (DEPA) for training is founded on three core concepts: (1) digital contracts, (2) confidential clean rooms, and (3)

differential privacy. Digital contracts backed by transparent contract services make it simpler for organizations to share datasets and collaborate by recording data-sharing agreements transparently. Confidential clean rooms ensure data security and privacy by processing datasets and training models in hardware-protected secure environments. Differential privacy further fortifies this approach, allowing AI models to learn from data without risking individuals' privacy.

Further drills down into the foundation from DEPA for training (DEPA Training Framework | DEPA World, 2023) relies on ADEPTS (Accountability, Democratized data, Transparency, Privacy by Design, Transparency, Secure) principles. This helps solve for the requirements of a mediated data collaboration for ML training. To help unbundle this further, we explain this with a case study along with the technical framework.

## 2.3 CASE STUDY (HYPOTHETICAL)

The following case study explains this capability in a simplistic and outcome orientation perspective.

In the bustling city of Bangalore (Karnataka, India), a group of visionary minds came together to establish AnalyzeMyXrays (AMX), a startup with a mission to revolutionize chest X-ray analysis for COVID infections. Fueled by a passion for leveraging machine learning (ML) in healthcare, the founders envisioned a future where advanced technology could play a pivotal role in early detection and diagnosis.

The genesis of AMX can be traced back to a shared concern about the overwhelming challenges posed by the COVID-19 pandemic. Recognizing the critical role that chest X-rays play in identifying respiratory infections, the founders set out to develop ML models that are capable of swiftly and accurately detecting signs of COVID in these images.

As the startup gained traction in small local hospitals, where they deployed their models locally (keeping privacy in mind), it became clear that their innovative approach had the potential to make a significant impact beyond Bangalore and in larger healthcare use cases.

The startup's initial breakthrough came in the form of a robust ML algorithm designed to analyze chest X-rays for patterns indicative of COVID-19. Trained on limited publicly available datasets of X-ray images (less than 50,000), the model exhibited remarkable accuracy, outperforming traditional diagnostic methods.

The team devised a strategic growth plan focused on expanding their reach across India's vast and diverse healthcare landscape. The key challenge to scale is the accessibility of large-scale tagged chest X-ray COVID-related datasets; the asymmetric benefits are the risk of data sharing (frozen market). So what are the options for AMX? The founders read about DEPA for ML training through a blog article.

AMX recognized the importance of building strong partnerships with hospitals and healthcare providers:

- Leveraging their ML expertise, the startup collaborated with leading medical institutions, offering a seamless integration of their diagnostic tools into existing healthcare systems.

- This collaborative approach not only facilitated widespread adoption but also allowed AMX to fine-tune their models based on real-world feedback from healthcare professionals.
- DEPA for ML allowed for a safe and secure ML training capability (required investment planning), but the capability could scale.

To address the unique challenges of the Indian healthcare system, AMX implemented a scalable and cost-effective pricing model, making their technology accessible to a wide range of healthcare providers, from major urban hospitals to rural clinics. This approach not only facilitated widespread adoption but also aligned with the startup's mission of democratizing advanced healthcare solutions. But there is a challenge!

As AMX solidified its presence in the Indian market, the founders set their sights on a broader horizon – the global X-ray diagnostic market. Recognizing that the impact of their technology extended far beyond regional boundaries, the startup invested in obtaining regulatory approvals and certifications necessary for international expansion. The approach taken by adopting the DEPA for ML training set the foundation.

Global outreach efforts included participation in key healthcare conferences, collaborations with international research institutions, and establishing a network of distributors to ensure seamless integration of AMX technology into healthcare ecosystems worldwide. The startup's commitment to maintaining the highest standards of data security and patient privacy played a pivotal role in gaining trust on the global stage.

This hypothetical success story of AMX serves as a testament to the transformative power of ML in healthcare:

- Combining technological innovation with a strategic growth plan
- Addressed a pressing healthcare need with a new way for diagnostic medicine
  - Synergy of human expertise coupled with AI leading to accurate, efficient, and accessible healthcare solutions for people around the world
- Most importantly, keeping data empowerment along with privacy by design

## 2.4 TECHNICAL IMPLEMENTATION

The technical implementation is built on three critical components. These components provide a foundation for an overall technical and enabling legal approach (hence, techno-legal).

The key to a successful privacy-preserving ML model is applicable to a diverse range of ML models, which include LLMs (use cases linked to fine-tuning, etc.). While there has been substantial work done on data privacy across the globe, ML/AI require lots of effort to be taken and will be complex. A quick construct of the DPI for AI is represented (in Figure 2.1), and the chapter will cover all aspects of the framework in detail, along with comprehensive references.

TABLE 2.1  
Technical Components of DPI for AI

Technology Layer	Details
Differential Privacy	Provides an approach to privacy preservation for AI model training
Confidential Computing	Secure environments for data, process, and training
Electronic Contracts	Self-enforcing, standardized, and scalable

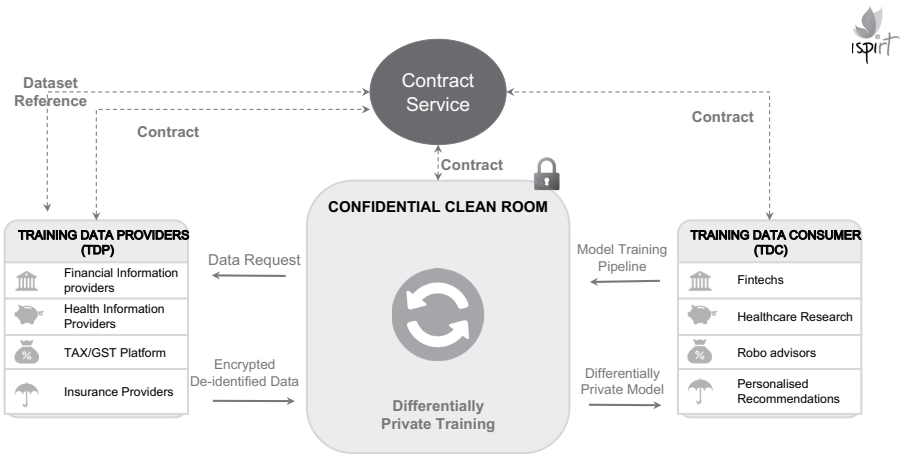


FIGURE 2.1 DEPA for training (DPI for AI) architecture.

2.4.1 DIFFERENTIAL PRIVACY

Although there are many techniques and approaches, we have explored differential privacy (DP). The definition of differential privacy emerged from a long line of work applying algorithmic ideas to the study of privacy (Dwork, n.d.). The definition of differential privacy (Differential Privacy | DEPA World, 2023a; Differential Privacy | Harvard University Privacy Tools Project, n.d.) is a rigorous mathematical definition. In the simplest setting, consider an algorithm that analyzes a dataset and computes statistics about it (such as the data’s mean, variance, median, mode, etc.). Such an algorithm is said to be differentially private if by looking at the output, one cannot tell whether any individual’s data was included in the original dataset or not. Most notably, this guarantee holds for *any* individual and *any* dataset. Therefore, regardless of how eccentric any single individual’s details are, and regardless of the details of anyone else in the database, the guarantee of differential privacy still holds. This gives a formal guarantee that individual-level information about participants in the database is not leaked.

There has also been extensive work done on the applications of DP to machine learning (Ponomareva et al., 2023). These also play a vital role in the development of the technical solution for use cases. DP has been explored in a number of noteworthy scenarios, such as the US census (US Census Bureau, 2023). Apple is another example that uses DP to transform the information shared with Apple before it ever leaves the user's device, such that Apple can never reproduce the true data.

In specific to DEPA, imagine a training data consumer (TDC) who wants to train a model for detecting payment fraud. They could do this by collecting labeled transaction data from multiple payment companies. The trained model might be quite useful, but it could also reveal a lot of information about the transactions, even if the TDC only has access to the trained model. Other kinds of models have been shown to be potentially vulnerable; credit card numbers have been pulled out of language models and actual faces reconstructed from image models.

The DEPA training framework supports model training using a robust approach based on differential privacy (Ponomareva et al., 2023; Papernot & Thakurta, 2023). In specific, DP works by introducing a privacy loss or privacy budget parameter, often denoted as epsilon ( $\epsilon$ ), to the dataset. The parameter  $\epsilon$  controls how much noise or randomness is added to the raw dataset. The added randomness is controlled; therefore, the resulting dataset is still accurate enough to generate aggregate insights through data analysis while maintaining the privacy of individual participants.

There are several ways of training ML models with differential privacy. By far, the most common is using Differentially Private Stochastic Gradient Descent (DP-SGD). DP-SGD (PyTorch, 2021; Rathi, 2021; Dupuy et al., 2021) prevents the model from memorizing or leaking sensitive information about the data by adding noise to the gradients during the optimization process. The amount of noise is carefully calibrated to satisfy a mathematical definition of differential privacy, which guarantees that the model's output is almost independent of any single data point. DP-SGD can be applied to various types of models, such as deep neural networks, and has been used for tasks such as natural language processing and computer vision. DP-SGD can be applied to fine-tune models while preserving the privacy of the task-specific data.

## 2.5 DIFFERENTIAL PRIVATE TRAINING/FINE-TUNING IN DEPA

The DEPA training framework provides training data providers (TDPs) with mechanisms to ensure that TDCs use their datasets in a way that protects the privacy of data principals. Using these mechanisms, TDPs can meet compliance requirements. Granular details on this can be perused in the section of DEPA World that references the specific section on differential private training/fine-tuning in DEPA (Differential Privacy | DEPA World, 2023b).

In conclusion, DP along with privacy budget management (Ponomareva et al., 2023) provides a base foundation for privacy-preserving training.

### 2.5.1 CONFIDENTIAL COMPUTING

This technology layer plays a critical role in the framework. Confidential computing provides a security paradigm that focuses on protecting data during processing.

Unlike traditional methods, which secure data at rest or in transit, confidential computing aims to safeguard information while it's being used by applications, even from privileged users and the underlying infrastructure.

Confidential clean rooms are based on novel security features for confidential computing. Confidential computing, available in most modern CPUs and GPUs (e.g., Intel SGX and TDX, AMD SEV-SNP, ARM CCA and NVIDIA Confidential GPUs), utilizes hardware-based trusted execution environments (TEEs) to isolate the code and data of a given task from the rest of the platform, including privileged entities such as server administrators and hackers who may have compromised the platform. Therefore, the task can be trusted with sensitive data, as hardware memory encryption and access control ensure it will be accessible only to the TEE code.

A core feature of confidential computing is remote attestation: the TEE code can request the hardware to attest a given message (such as a public key), together with the digests of its binary image and configuration, measured when the TEE was created. The attestation is signed with a key unique to the CPU and is backed by a public-key certificate for the platform (endorsed by the hardware vendor). By verifying this signature, a user can thus authenticate the TEE's code and hardware platform before trusting it with sensitive data.

Confidential containers are already supported by most cloud platforms, including AWS Nitro Enclaves (Lightweight Hypervisor – AWS Nitro System – AWS, n.d.), Confidential Spaces on GCP (Confidential Computing | Google Cloud, n.d.), and Confidential Azure Container (Microsoft Azure Confidential Computing | Microsoft Azure, n.d.) Instances on Azure. This is offered by all of the major cloud providers and a viable approach for a zero trust scenario application across the DEPA ecosystem. The technology key components and benefits include the following, shown in Tables 2.2 and 2.3.

In the DEPA training framework, datasets are brought together and processed in secure environments known as confidential clean rooms (Confidential Clean Room High Level Design | DEPA World, 2023). Confidential clean rooms intend to meet a set of privacy and security goals through technical measures. These goals include the following:

- Prevent inappropriate access to raw data or other intermediate data through technical enforcement.
- Allow TDPs to retain control over the data they've shared without trusting any third party.
- Allow TDCs to retain control over the models they are training without trusting any third party.
- Enforce constraints on data usage defined in contracts. This includes noise addition during analytics and training in line with the long-term goal of differential privacy.
- Support flexible, scalable, and extensible training so that TDCs can choose when and what kinds of models they wish to train.
- Provide open and transparent implementations for all infrastructure components.

**TABLE 2.2**  
**Technical Confidential Compute Components of DPI for AI**

Technology Layer	Details
Enclave technology	Utilizes hardware-based techniques, like Intel SGX or AMD SEV, to create secure enclaves. Enclaves are isolated regions of memory where sensitive computations take place, shielded from the rest of the system.
Secure Execution Environment	Ensures the confidentiality and integrity of data by executing code securely within the enclave. Protects against threats like memory tampering and side-channel attacks.
Attestation	Verifies the integrity of enclaves, confirming they haven't been compromised. Enables parties to trust the secure execution environment.

**TABLE 2.3**  
**Key Benefits of Confidential Computing of DPI for AI**

Technology Layer	Details
Data Confidentiality	Protects sensitive data from unauthorized access during processing, reducing the risk of data breaches.
Trust in the Cloud	Enhances trust in cloud environments by safeguarding data from cloud providers and administrators.
Privacy-Preserving Analytics	Enables secure computation on encrypted data, allowing for privacy-preserving analytics and collaborative processing without exposing raw data.
Secure Multi-Party Computation	Facilitates secure collaboration among multiple parties, enabling joint data analysis without revealing individual datasets.
Compliance and Regulation	Helps organizations comply with data protection regulations by ensuring end-to-end data security, even in shared or third-party computing environments.
Intellectual Property Protection	Guards proprietary algorithms and intellectual property by securing their execution within enclaves.
Application Isolation	Provides a strong isolation boundary for applications, protecting them from external and internal threats.



In conclusion, confidential computing technology offers a paradigm shift in data security, providing a robust solution for protecting sensitive information throughout its lifecycle, particularly in cloud and shared computing environments.

### 2.5.2 ELECTRONIC CONTRACTS

Electronic contracts play a crucial role in establishing a techno-legal framework by providing a digital foundation for legal agreements. They enhance efficiency, reduce paperwork, and offer opportunities for smart contracts, integrating technology into legal processes. This convergence supports a more streamlined, secure, and automated approach to transactions (including payments), aligning with the evolving landscape of technology and law. We will now focus on the DEPA-specific aspects.

The contract service plays a crucial role within the DEPA training framework, facilitating secure data collaboration and contractual agreements among various participants, including training data providers (TDPs), training data consumers (TDCs), and confidential clean room (CCR) providers.

The contract service (Contract Service Specifications | DEPA World, 2023) maintains a registry – a verifiable data structure that records signed dataset references and contracts – and enforces contract registration policies. It also maintains a service key, which is used to endorse the state of the registry in receipts. All contract services must expose standard endpoints for registration of datasets and signed contracts and receipt issuance. Each contract service also defines its registration policy, which must apply to all entries in the registry.

The combination of registry, identity, registration policy evaluation, and registration endpoint constitute the trusted part of the contract service. Each of these components should be carefully protected against both external attacks and internal misbehavior by some or all of the operators of the contract service.

Beyond the trusted components, contract services may operate additional endpoints for auditing, for instance, to query for the history of dataset references and signed contracts registered by a given participant.

The combination of registry, identity, registration policy evaluation, and registration endpoint constitute the trusted part of the contract service. The following provides a sample configuration (JSON) of a dataset by a TDP; also included is a sample full contract between a TDP and TDC (*DEPA Contract*, n.d.).

#### Sample configuration (JSON) of a dataset by a TDP

```
{
  "id": "",
  "name": "cowin",
  "url": "https://xyz.domain/cowin/data.img",
  "provider": "",
  "key": {
    "type": "azure",
    "properties": {
      "kid": "COWINFilesystemEncryptionKey",
```

```

"authority": {
"endpoint": "xyz.attest.abc.net"
},
"endpoint": ""
}
}

```

## Sample Contract (JSON) across TDP and TDC

```

purpose": "TRAINING",
"constraints": [
{
"privacy": [
{
"dataset": "12345ba8-bab8-11ed-afa1-0242ac120002",
"epsilon_threshold": "1.5",
"noise_multiplier": "2.0",
"delta": "0.01",
"epochs_per_report": "2"
},
{
"dataset": "67890cc6-bab8-11ed-afa1-0242ac120002",
"epsilon_threshold": "1.5",
"noise_multiplier": "2.0",
"delta": "0.01",
"epochs_per_report": "2"
},
{
"dataset": "12345144-bab8-11ed-afa1-0242ac120002",
"epsilon_threshold": "1.5",
"noise_multiplier": "2.0",
"delta": "0.01",
"epochs_per_report": "2"
}
]
}
],

```

## 2.6 DEPA | TECHNO-LEGAL FRAMEWORK

The power of DEPA, which already provides an open network where data principals are empowered to share their data residing with one or more data providers with consent, will serve as a foundation (Workflows | DEPA World, 2023). The initial version of DEPA, however, was restricted to sharing data belonging to a single data principal. The Account Aggregator (AA), an IndiaStack DPI, is already in use and supports the viability of the DEPA framework. As part of DEPA for training, we expand the possibilities of DEPA to enable seamless data sharing via a techno-legal

framework. A quick historical timeline of DEPA evolution includes DEPA launched (2017), Account Aggregator (AA) live (2021), and DEPA training (2021).

The ability to have a techno-legal approach to the whole AI model training process is very critical. There are many international data protection frameworks, such as GDPR, CCPA, India DPDP (Parliament, 2023), etc. In the context of AI frameworks, the NIST AI Risk Management Framework (AI Risk Management Framework | NIST, 2025) provides an operationalizing framework that helps identify and reduce associated risks. Countries worldwide are proactively engaging (*Global AI Regulation Tracker*, n.d.; *IAPP*, n.d.; United Nations, n.d.; The OECD Artificial Intelligence Policy Observatory – *OECD.AI*, n.d.; *GPAI*, 2023) and working out effective designing/implementing AI governance legislation (The Act Texts | EU Artificial Intelligence Act, n.d.) commensurate with the velocity and variety of proliferating AI-powered technologies. This legislation, like many foundational technologies, can have far-reaching implications as AI is applicable to almost all domains. The framework has the following technical architecture.

The unblocking of the datasets will enable multiple scenarios where organizations need access to bulk data, e.g., running analytics to identify trends, or training ML models. This promises immense socio-economic benefits for India (as well as around the globe) across the spectrum untapped due to data unavailability. Specific to India, this will open up the India ML model ecosystem to an existing large startup ecosystem.

2.7 THE DEPA ECOSYSTEM

This new digital public infrastructure (DPI), namely DEPA for training (DPI for AI), is designed to address diverse use case scenarios. This DPI has unbundled a comprehensive market ecosystem, which includes the following: training data providers (TDPs), training data consumers (TDCs), self-regulated organizations (SROs), technical standards organizations (TSOs), data principals (DPs), CCR providers (CCR-P), data discovery agents (DAs), and technical service providers (TSPs).



The Techno-Legal Framework for Data Collaboration

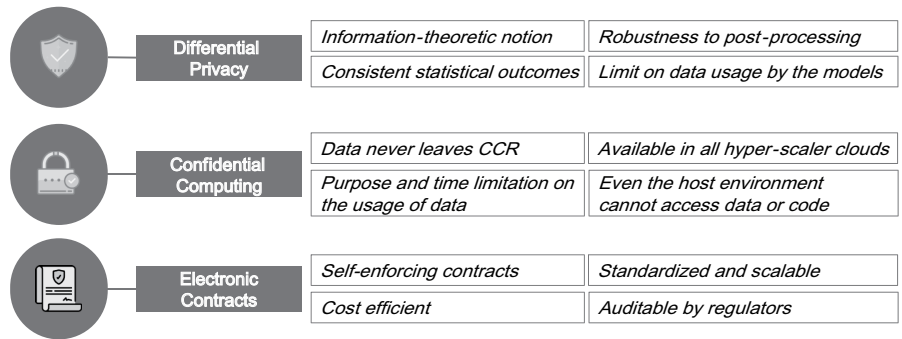


FIGURE 2.2 DEPA techno-legal framework.

Highly critical healthcare use cases, such as tracking and diagnosis of diseases using data across multiple sources (institutions, hospitals, etc.), are ideas for this framework. There is a complete reference (Differential Privacy | DEPA World, 2023b) implementation that illustrates this end to end. There are also multiple financial use cases spanning financial inclusion, detecting fraud, etc. This will be facilitated using datasets from multiple banks and institutions. These use cases are now possible without compromising the privacy of data principals. Thus, this DPI proposes a path to unlock value buried in data silos by providing a techno-legal framework to facilitate anonymized data sharing at scale.

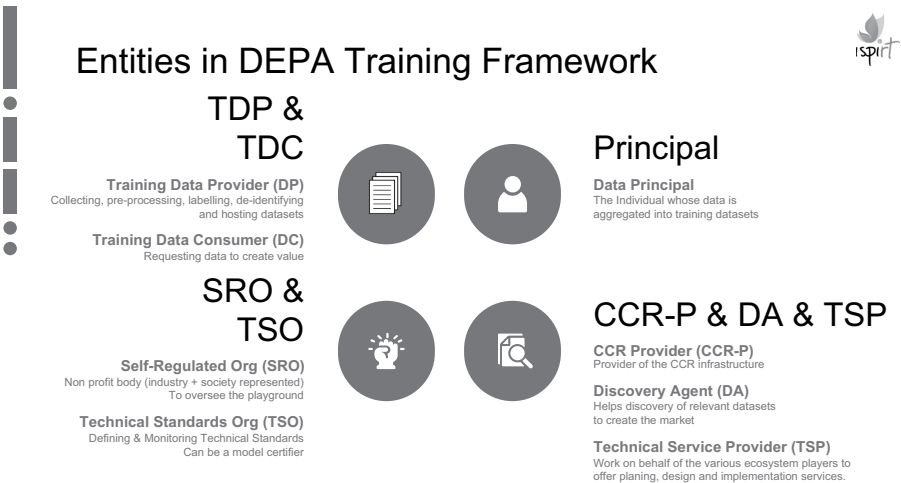


FIGURE 2.3 DEPA market ecosystem.

## Demo Scenario : COVID Prediction Modeling

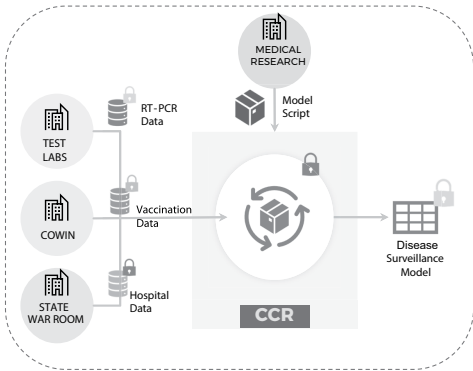


FIGURE 2.4 DEPA reference implementation – COVID disease surveillance.



Aggregate Datasets Train AI Models to Make Predictions

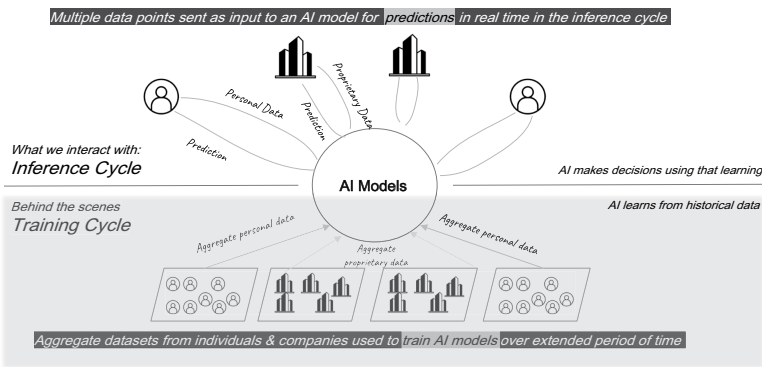


FIGURE 2.5 Continental-scale AI models training.

2.8 THE REFERENCE IMPLEMENTATION

A reference implementation on a disease surveillance use case built on this framework will be discussed. This illustrates a very realistic scenario of the benefits of using this kind of framework for critical country-scale use cases.

More details on how these concepts come together to create an open and fair ecosystem are available at [https://depa.world/training/depa\\_training\\_framework](https://depa.world/training/depa_training_framework).

2.9 CONCLUSION

The unblocking of the datasets will enable multiple scenarios where organizations need access to bulk data, e.g., running analytics to identify trends, or training machine learning models. This promises immense socio-economic benefits for India (as well as around the globe) across the spectrum untapped due to data unavailability. Specific to India, this will open up the India ML model ecosystem to an existing large startup ecosystem. In conclusion, with the rise of large-scale AI models, the potential applications are vast and varied, and they have the power to help solve some of society’s most pressing problems. However, it is essential to ensure that the development and deployment of these models are done responsibly to avoid any unintended consequences. Technical frameworks like DEPA for ML training (DPI for AI) will play a critical role. The aspiration is to provide for the ability to scale and build continental-scale use cases.

REFERENCES

*The act texts | EU artificial intelligence act.* (n.d.). <https://artificialintelligenceact.eu/the-act/>  
*AI risk management framework.* (2025, January 31). NIST. <https://www.nist.gov/itl/ai-risk-management-framework>

- Confidential clean room high level design*. (2023, September 17). DEPA World. [https://depa.world/training/confidential\\_clean\\_room\\_design](https://depa.world/training/confidential_clean_room_design)
- Confidential computing*. (n.d.). Google Cloud. <https://cloud.google.com/confidential-computing>
- Contract service specifications*. (2023, September 4). DEPA World. [https://depa.world/training/contract\\_service\\_specifications](https://depa.world/training/contract_service_specifications)
- DEPA contract*. (n.d.). DEPA Training. <https://github.com/iSPIRT/depa-training/blob/main/scenarios/covid/contract/contract.json>
- DEPA training framework*. (2023, August 14). DEPA World. [https://depa.world/training/depa\\_training\\_framework](https://depa.world/training/depa_training_framework)
- Differential privacy*. (2023a, August 18). DEPA World. [https://depa.world/training/differential\\_privacy](https://depa.world/training/differential_privacy)
- Differential privacy*. (2023b, August 18). DEPA World. [https://depa.world/training/differential\\_privacy#differential-private-trainingfine-tuning-in-depa](https://depa.world/training/differential_privacy#differential-private-trainingfine-tuning-in-depa)
- Differential privacy*. (n.d.). Harvard University Privacy Tools Project. <https://privacytools.seas.harvard.edu/differential-privacy>
- Dupuy, C., Arava, R., Gupta, R., & Rumshisky, A. (2021, July 14). *An efficient DP-SGD mechanism for large scale NLP models*. arXiv.org. <https://arxiv.org/abs/2107.14586>
- Dwork, C. (n.d.). *Differential privacy*. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dwork.pdf>
- Global AI regulation tracker*. (n.d.). <https://www.techieray.com/GlobalAIRegulationTracker.html>
- GPAI*. (2023). <https://gpai.ai/>
- IAPP*. (n.d.). <https://iapp.org/resources/article/global-ai-legislation-%20tracker/>
- Lightweight hypervisor – AWS nitro system – AWS*. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/ec2/nitro/>
- Microsoft Azure confidential computing: Overview and details about confidential computing on the Azure platform*. (n.d.). Microsoft Azure. <https://azure.microsoft.com/en-us/solutions/confidential-compute/>
- The OECD artificial intelligence policy observatory – OECD.AI*. (n.d.). <https://oecd.ai/en/>
- Papernot, N., & Thakurta, A. G. (2023, October 3). *How to deploy machine learning with differential privacy*. NIST. <https://www.nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy>
- Parliament. (2023). The digital personal data protection act, 2023. *The Gazette of India Extraordinary*. [https://prsindia.org/files/bills\\_acts/bills\\_parliament/2023/Digital\\_Personal\\_Data\\_Protection\\_Act\\_2023.pdf](https://prsindia.org/files/bills_acts/bills_parliament/2023/Digital_Personal_Data_Protection_Act_2023.pdf)
- Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., Vassilvitskii, S., Chien, S., & Thakurta, A. G. (2023). How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77, 1113–1201. <https://doi.org/10.1613/jair.1.14649>
- PyTorch. (2021, December 15). Differential privacy series part 1 | DP-SGD algorithm explained. *Medium*. <https://medium.com/pytorch/differential-privacy-series-part-1-dp-sgd-algorithm-%20explained-12512c3959a3>
- Rathi, M. (2021, December 26). *Deep learning with differential privacy (DP-SGD explained)*. <https://mukulrathi.com/privacy-preserving-machine-learning/deep-learning-differential-privacy/>
- United Nations. (n.d.). *AI advisory body*. United Nations. <http://www.un.org/en/ai-advisory-body>
- US Census Bureau. (2023, March 27). *Why the census bureau chose differential privacy*. Census.gov. <https://www.census.gov/library/publications/2023/decennial/c2020br-03.html>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv (Cornell University)*, 30, 5998–6008. <https://arxiv.org/pdf/1706.03762v5>
- Workflows*. (2023, September 2). DEPA World. <https://depa.world/training/workflows>

---

# 3 Generative AI Governance

## *Technological Monoculture, Market Structure and the Risk of Correlated Failures*

*Ramayya Krishnan, Prasanna Parasurama,  
Joao Sedoc, and Arun Sundararajan*

Given the significant barriers to creating high-quality foundation models (cost of collection of training data, need for access to immense computing power), a small number of primarily closed-source foundation models are establishing leadership in the generative AI market. Applications based on these foundation models are being deployed by a number of firms across multiple sectors.

Responsible use of AI requires an understanding of the safety and reliability of the AI models and their use in applications of societal consequence. The advent and success of large language models (LLMs) has changed the AI architectures that are being deployed in organizational applications. Specifically, LLMs are developed and trained without a single downstream use case in mind. Fine-tuning or otherwise customizing these general-purpose models creates an instance of an AI model suited to the needs of an application. This platform model is a departure from the purpose-built AI models designed to meet the needs of particular use cases.

Granted, LLM architectures have many advantages that are common in platform-based approaches, most notably the economies of scale and scope that flow from being able to draw on pre-trained capabilities rather than building them from scratch. However, AI applications derived from these models can suffer from correlated errors and risks. These errors and risks may arise in myriad downstream applications, ranging from recruiting to healthcare provision.

Additionally, and perhaps more saliently, given the extent to which the training datasets of the *generative AI* foundation models overlap, the risks could be far more substantial than what might be suggested by a competitive analysis of market structure and market shares. Indeed, a recent study by Zou et al. (2023) demonstrates a simple class of suffix attacks that exploit a vulnerability in all current aligned LLMs to get them to produce content that their guardrails were designed to prevent.

Understanding correlated risks is a topic that has not been extensively studied in the literature and is critical to the responsible use of AI in consequential application domains. This is a gap our study addresses.

In this study, we analyze the relationship between the *diversity* in upstream foundation models and the risk of *correlated failures* and shared vulnerabilities in downstream applications. Such risks are similar to those generated by *monoculture* in farming settings (Power & Follett, 1987), wherein reliance on fewer seed strains can lead to shared crop vulnerabilities to pathogens and a higher risk of famine. These risks have also been highlighted in the digital context, most notably about vulnerabilities in information security (Birman & Schneider, 2009; Chen et al., n.d.). For example, the vast market share of the Windows operating system has for decades provided malicious agents with the incentive to invest effort to discover and exploit its information security vulnerabilities.

More recently, the risks of algorithmic monoculture have been raised for AI, largely in studying algorithmic screening of job applicants. In particular, Kleinberg and Raghavan (2021) analyze the case where firms that compete on hiring have a choice of using algorithmic hiring or manual processes and demonstrate that homogeneity in the algorithm used by the competing firms leads to a type of Braess’s paradox: the introduction of a more accurate algorithm can drive the firms into a unique equilibrium that is worse for society than the one that was present before the algorithm existed. Bommasani et al. (2022) develop a simple mathematical formalism to measure systemic failure where the same individual is rejected by every firm that they apply to on account of the homogeneity of the resume-processing algorithm in use. Their subsequent measurement experiments study the extent of correlation in outcomes depending on which adaptation method was used to adapt the foundation model.

Building on this stream of literature, we study the extent to which foundational LLMs pose a systemic risk of correlated failures in a high-stakes setting: algorithmic screening of job applications. We consider a scenario in which multiple firms use the same foundational model to fine-tune a resume-screening algorithm using their own data. We ask whether the use of the same foundational model contributes to correlated errors (false negatives and false positives) across firms – i.e., whether the same individual would be incorrectly rejected (false negative) or incorrectly selected (false positive) across firms.

We use applicant tracking system (ATS) data from eight firms based in the U.S. The ATS tracks all details of the firm’s job postings (job title, department, job description), job applications (candidate details, demographics, resume text), and the outcome of each application (whether the applicant received a callback). The data spans 2014–2018, containing 1.17 million job applications for 6,600 job postings. Since we are interested in correlated errors across firms for a given individual, we identify the subset of individuals who applied to similar positions at multiple firms within the same time period in our dataset. This amounts to 25,000 individuals with 65,000 applications to 3,600 jobs across eight firms.

In our initial set of baseline experiments, we prompt off-the-shelf LLaMA-2-7B and LLaMA-2-13B models, both foundational LLMs released by Meta, with the



candidate's resume and the corresponding job description and ask whether the candidate should receive a callback. We compare these algorithmic predictions to the ground truth in our ATS data (whether the candidate received a callback) and estimate the level of correlated errors across firms. Our results show that the off-the-shelf model has almost no predictive power for this screening task, leading to uncorrelated errors across firms.

Subsequently, we use parameter efficient fine-tuning to create eight different LLaMA-2-7B models, one for each firm, on the respective firm's hiring data, and study how the level of correlated errors changes with the model's predictive power. Our preliminary findings show that both standard machine learning models (e.g., logistic regression + tf-idf) and fine-tuned LLaMA-2-7B models have similar area under the curve (AUC). However, unlike baseline machine learning models, LLaMA-2 fine-tuned models show significant correlated false negatives between firms.

## REFERENCES

- Birman, K. P., & Schneider, F. B. (2009). The monoculture risk put into context. *IEEE Security & Privacy*, 7(1), 14–17.
- Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*. <https://arxiv.org/pdf/2211.13972.pdf>
- Chen, P., Kataria, G., & Krishnan, R. (n.d.). *Correlated failures, diversification, and information security risk management*. AIS Electronic Library (AISeL). <https://aisel.aisnet.org/misq/vol35/iss2/9/>
- Kleinberg, J., & Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*. <https://arxiv.org/pdf/2101.05853.pdf>
- Power, J. F., & Follett, R. F. (1987). Monoculture. *Scientific American*, 256(3), 78–87.
- Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and transferable adversarial attacks on aligned language models*. arXiv preprint arXiv:2307.15043. <https://arxiv.org/pdf/2307.15043>

---

# 4 Empowering Citizens through Responsible AI Governance

## *Policy Recommendations for Public Algorithm Registers*

*Jens Meijen and Niharika Gujela*

### 4.1 INTRODUCTION

Governments have increasingly been experimenting with artificial intelligence (AI) to improve operational efficiency in the public sector, reduce costs, and enable more accurate or predictive decision-making. Especially in recent years, AI investments have been a focal point at various levels of government (Mehr, 2017; Van Noordt & Misuraca, 2022), even if capabilities vary (Van Noordt & Tangi, 2023). However, despite the potential benefits, the deployment of AI systems in the public sector carries significant risks, such as bias, discrimination, and opaque decision-making. As the public sector often has a major impact on citizens' lives, these issues can have far-reaching consequences (Leslie, 2019). To protect citizens from algorithmic harm and manage associated risks, many governments are supplementing their AI investments with responsible AI governance efforts to ensure safety, transparency, and accountability.

One of many possible tools to improve transparency and accountability in public sector AI is a public algorithm register that documents all the algorithms in use by a public administration and discloses the purpose, function, and impact of those algorithms to the general public. Algorithm registers provide first-order (system level) and second-order (governance level) transparency, both of which are needed to ensure AI accountability, which is often blurry (Murad, 2021; cf. Kaminski, 2020; Krafft et al., 2022). Some argue that a more accurate term (and focus) would be 'automated decision-making system registers', as their main goal is not just to make the functioning of algorithms more transparent, but to provide insight on how decisions are made with algorithmic assistance. Over the past few years, several cities launched their public algorithm registers, including Amsterdam, Helsinki, Nantes, Antibes, and New York City. Following this example, nine European cities (Barcelona, Bologna, Brussels Capital Region, Eindhoven, Mannheim, Rotterdam, and Sofia) collaborated through the Eurocities Digital Forum Lab network to develop an

AI algorithm register standard. While public transparency is often cited as the main purpose of the registers, several public administrations also see the registers as a potential ‘catalyst for meaningful democratic participation and a platform for fostering mutual trust’ (Haatja et al., 2020, p. 3).

Trust in AI is crucial for these systems to be socially accepted and achieve long-term success (see Wirtz et al., 2019). Public algorithm registers may therefore be a promising path towards a more broad-based acceptance and more democratically accountable use of AI in government. As most public algorithm registers are still relatively new and some are still in the experimental phase, academic studies are limited. However, if the registers are ever to improve public trust, this knowledge gap must be addressed.

Our chapter therefore analyzes some of the main issues in the algorithm registers that were part of the “first wave” of algorithmic accountability practices in the public sector (see Ada Lovelace Institute et al., 2021). We focus in particular on the Dutch national register, the Helsinki municipal register, and the Rotterdam municipal register. In what follows, we first assess the existing lessons learned on public algorithm registers before identifying four categories of shortcomings that limit the registers’ potential to improve public trust: (1) quality and quantity of information, (2) citizen experience, (3) accessibility for non-experts, and (4) accountability. We argue that there is a clear strategy-purpose misalignment, as the way the registers are implemented and maintained (strategy) does not reflect their purpose of improving public transparency and better informing citizens of public algorithm use. We also provide recommendations to mitigate the issues we identified. At the same time, we recognize that algorithmic accountability practices like algorithm registers are difficult to compare and evaluate across contexts, as governments operate under different constraints and citizen demands (see Ada Lovelace Institute et al., 2021, p. 12). We therefore do not aim to provide a normative comparison or criticize existing efforts, but rather present a constructive set of recommendations that policymakers can take on board as they see fit.

## 4.2 REVIEW OF CURRENT LITERATURE

In what follows, we study some of the existing recommendations on algorithm registers. Due to constraints on the length of this chapter, we do not cover every register or study. We also have no room to meaningfully engage in discussions on the *a priori* legitimacy of algorithmic governance and how algorithm registers tend to normalize the use of AI in the public sector. We focus on the question of how to mitigate risk as opposed to whether algorithms should be used at all (see Cath et al., 2022).

Most registers include information on the system’s context and purpose, its technical specifications, and its impact on decision-making and on citizens – or what it is, how it works, and what it does. In a synthesis of existing literature, Murad (2021, pp. 18–20) recommends that creators of registers carefully consider (1) the existing political and legal context (e.g., what is the legal basis for the register?), (2) how much it should disclose, (3) the intended audience (e.g., affected civilians, third-party experts, other governments), (4) what information it should disclose, and (5) how it

should disclose that information. The literature covered by Murad (2021, p. 21) also offers recommendations on what elements should ideally be included in the registers: system purpose, governance, procurement procedures, impact assessments, data use and quality, system architecture, performance, and monitoring, human oversight, redress possibilities, feedback loops, and audit reports. These overlap significantly with Eurocities' algorithm register standard.

In France, a working group of public servants from national and local governments issued recommendations on algorithm registers based on their experience with city-level registers. The recommendations focused on the following practical internal processes (likely because the participants were primarily public servants): define algorithms correctly by talking to agencies, involve a multitude of internal stakeholders, designate a person as a register sponsor, and prioritize areas where transparency can be most useful (Pénicaud, 2021). The French Etalab, a government task force working on open government data, also published a guide for public servants on how to create an algorithm register, with the explicit goal to 'gain insight into how a government decision was made' and 'to help citizens exercise their rights' (Fiche pratique: L'inventaire des principaux traitements algorithmiques – guides.etalab.gouv.fr, 2021).

In New York City, a public directory available online discloses the function, purpose, data usage, and vendor involvement of algorithmic tools used in public administration and service delivery (New York City Office of Technology and Innovation, 2022). However, due to the specialized technical language used in the register and its PDF format, it seems unlikely that non-experts will glean much information from the directory, and citizens may not be incentivized to download and search through the file. Its main added value would seem to be that it offers transparency for experts and civil society actors, or as a means of maintaining an internal overview of algorithmic tools. In this approach, public administrations can only realistically be held accountable through the input (and often goodwill) of third parties, who might then be able to analyze and relay that information to a wider audience (Ada Lovelace Institute et al., 2021).

The Amsterdam City Algorithmic Register, by contrast, aims to provide 'procedural transparency to the general public, while technical information is only required to be shared with third-party auditors' (Murad, 2021, p. 20). Tailoring the information in public registers to the needs of specific stakeholder groups remains a key challenge and likely requires trade-offs (19).

One citizen-centered study on algorithm registers, performed by the Dutch government (Doel groepenanalyse, lit. Target Group Analysis), engaged with a relatively diverse cross-section of society and showed that citizens would like to see the following improvements to the register (2023, p. 3):

- Reduce empty and poorly completed fields.
- Make it easy for visitors to look around.
- Improve search capabilities.
- Improve download functionality.
- Provide more information about the registry.

- Improve English-language functionality.
- Improve the home page.
- Make it possible to search by location (e.g., zip code).
- Make the algorithm code public.

In contrast to the recommendations provided by public servants themselves, these recommendations are more focused on user experience. Throughout its development cycle, the Dutch algorithm register has engaged in an iterative, open feedback process that aims to involve citizens in a collaborative process to improve its quality. The main challenges identified by citizens are filling the register with quality information, reaching a broader audience, and providing meaningful information for non-experts.

More generally, algorithmic accountability practices in the public sector should ideally: (1) include institutional incentives and binding legal frameworks; (2) set clear objectives across departments, define clear scopes, and be consistent across different levels of government; (3) use detailed, audience-appropriate forms of transparency; and (4) prioritize public participation (adapted from Ada Lovelace Institute et al., 2021, p. 4). However, implementing these recommendations can be challenging, not least due to institutional limitations, a lack of expertise, and varying digital literacy among policymakers (1), public servants (2), and citizens (3 and 4). Other higher-level recommendations include making algorithm registers a legal obligation at all levels of government, making them easily accessible and readable, enabling independent review (Automated Decision-Making Systems in the Public Sector – Some Recommendations, 2022), and involving the public in shaping algorithm use ‘commensurate with the risk level of the potential system,’ and utilizing third-party reviews (Pittsburgh Task Force on Public Algorithms, 2021).

In sum, existing recommendations can be divided into three categories: those focused on micro-level, citizen-centred improvements (e.g., the Dutch TGA); on meso-level internal processes (e.g., the French working group recommendations); and on overarching strategy (e.g., Automated Decision-Making Systems in the Public Sector – Some Recommendations, 2022; Ada Lovelace et al., 2021 – although not specific to algorithm registers). Even if concrete approaches differ, the purpose of algorithm registers is usually to improve public transparency and thereby improve trust in governments. Most recommendations include some type of citizen participation, legal obligations, and increased accessibility. However, many existing recommendations also point out that the information in the registers often does not adequately reach the people whose trust they are supposed to increase, so there is a misalignment between their purpose and the strategies being used. We therefore offer policy recommendations that could serve to amend that strategic misalignment.

## RECOMMENDATIONS

Our recommendations build upon key issues in the literature, our analysis of existing algorithm registers, and interviews with public administrators and experts working on maintaining or creating algorithm registers in Rotterdam, Brussels, and the Dutch national register. Our recommendations are practical tips for public administrations

looking to create or improve algorithm registers. Not all of our recommendations necessarily apply to existing registers, and we encourage policymakers to pick and choose any insights that seem useful for their situation and the goal of the register they are building (such as the intended audience). The registers we studied are at different stages (not yet deployed, in early deployment, and more advanced) but in similar contexts (highly developed and relatively digitally literate societies, see Wiley DSGI, 2021) – ensuring our insights remain relevant throughout the registers’ development cycle. A drawback of our recommendations is that they presume a relatively high level of digitalization in administrations and society, although digital literacy among all stakeholders remained a key issue identified by all interviewees. Further research should be done to create recommendations that apply specifically to other contexts.

### 4.3 QUALITY AND QUANTITY OF INFORMATION

The most significant problem is that the data in the registers is either sparse or non-existent. Incompleteness may be the most obvious issue: many fields in existing registers are left empty or are barely filled. This was also one of the main observations in the Dutch algorithm register’s Target Group Analysis, where a diverse group of people was asked for feedback (*In gesprek over het algoritmeregister met burgers*, 2023, p. 8). Furthermore, some entries barely contain any meaningful information, which means the registers are not improving transparency as much as they could. The main reasons for this are a lack of data literacy and digital expertise among public servants, a lack of resources (usually time and people) and high fragmentation in public administrations, a lack of incentives or legal obligations to fill in the registers, and a lack of system documentation provided by third-party providers or external consultants. One interviewee noted that public servants see the register as an annoyance more than a useful tool. Several interviewees pointed out that transparency has its limits: publishing too much information may lead to bad-faith actors trying to game the system and use the information to its advantage.

### RECOMMENDATIONS

1. Establish a culture of transparency among public servants through enforced internal policies that prioritize filling in the registers completely. Where the available data is insufficient, public servants should flag this, and relevant stakeholders (such as model developers) should be contacted for further information.
2. Improve internal data and AI literacy among public servants through training sessions. Instilling a sense of urgency by explaining salient cases of algorithmic scandals may also help.
3. Take a long-term approach to external consultants. Instead of having consultants perform work for administrations, let a public servant shadow the consultants and disseminate relevant knowledge to others in open training sessions or hire consultants to also provide hands-on training sessions.

4. Balance information and security: assess whether a system can be gamed if certain information is released. If certain data must be omitted, specify the exact grounds for the omission.
5. Create legal obligations for public agencies to fill in registers to the best of their abilities (see Automated Decision-Making Systems in the Public Sector – Some Recommendations, 2022). Possibilities include a separate law that obliges governments or cities to maintain algorithm registers or an independent watchdog to monitor the registers.

#### 4.4 CITIZEN EXPERIENCE

Another issue is the user or citizen experience design of the registers themselves. First, the registers are usually embedded in the digital ecosystem of their respective public institutions (whether on the same website or on a separate but linked page) and therefore use similar layouts and styles. However, the citizen experience of the registers is fundamentally centered on the administrations' view of algorithms, in which each agency sees every algorithm separately. This is also related to the fragmentation of public administrations today. By contrast, citizens' lives are impacted by several algorithms at the same time, although rarely by every single algorithm at once. Second, the search function in the registers tends to be predicated on the world of the public administrations instead of citizens' lived experiences. One example cited by the Dutch algorithm register is that citizens searching for a more informal way of saying "tax authorities" resulted in no hits, while the official wording did result in relevant hits (*In gesprek over het algoritmeregister met burgers*, 2023).

#### RECOMMENDATIONS

1. To amend the narrow view of algorithms, a fundamentally more panoramic overview of algorithms and their interactions would be useful. An example is Rotterdam's page, which simply shows all algorithms in the register. However, with a larger number of algorithms, this would reduce clarity. We therefore recommend showing relevant categories of algorithms (e.g., taxes, traffic) for visitors to click and, if necessary, breaking categories down into subcategories.
2. Ideally, registers would offer a more tailored overview of relevant algorithms for individual citizens in a personalized and intuitive 'algorithm dashboard', which would show all algorithms that impact a citizen's life at a glance. Using an optional questionnaire or chatbot to enter personal details could facilitate the creation of the dashboard. Another potential improvement could be the option to compose an automated risk report based on the algorithm dashboard, which would show, for example, to what extent a citizen is at risk of algorithmic profiling or discrimination based on their individual situation.
3. To improve the search function, feedback from citizens with little affinity for politics should be included so as to ensure informal names for certain institutions deliver the desired results.

4. In terms of interface design, it could be interesting to more actively diverge from the standard designs of public administrations to create more visually appealing interfaces. This could incentivize people who generally distrust the government to also use the algorithm register.

## 4.5 ACCESSIBILITY FOR NON-EXPERTS

The accessibility for non-experts of the information included in the registers could also be improved. Many entries use highly technical language, which may require specialized knowledge. Register fields dealing with applications often remain comparatively vague, which makes it hard for citizens to see in what way a certain algorithm has affected them. Likewise, the titles of some entries are arguably incomprehensible to citizens without prior knowledge of public administrations. Lastly, language barriers can be an issue, as in the case of the Dutch national register, where non-Dutch speakers have to resort to automated translation.

## RECOMMENDATIONS

1. Ideally, registers should avoid overly complex or simple entries. After filling a field with what could be seen as ‘expert language’, information should be summarized in simpler terms, which could include a non-technical analogy or metaphor that people can reasonably be expected to understand. Another option is a layered approach, with basic information immediately visible and complex technical information at a lower level (see Murad, 2021, p. 20).
2. The accessibility of registers should be tested by drawing representative random samples from the register for evaluation by a diverse cross-section of society. This should be a citizen-centered, iterative process.
3. In general, every entry should contain a practical use case or example, preferably highlighting the real, tangible consequences of the algorithm’s decisions.
4. Titles of entries should be descriptive. Public servants should ask themselves if someone affected by or interested in a specific algorithm would understand that algorithm’s respective title in the register.
5. Use engaging formats like videos to engage and improve citizens’ understanding of algorithmic use-cases.
6. Include dummy data and simulated environments to help citizens more intuitively understand what data the system uses and how (see Murad, 2021, p. 20).

## 4.6 ACCOUNTABILITY

The registers also offer few means to hold public administrators accountable for algorithmic decisions or for subpar transparency in algorithm entries. However, if the registers are indeed intended to reinforce trust in the algorithms used in the public sector, accountability is crucial. If public algorithm registers are ever intended to be useful for more than an internal overview of all algorithms in use, there should be real consequences attached to not utilizing them correctly. Identifying the entities



to be held responsible is therefore paramount. The burden of ensuring the registers' quality seems to currently lie mostly with people coordinating the register (in the Dutch case, a separate government organization) instead of lying with the fragmented group of public entities using (and therefore benefiting from) algorithms.

## RECOMMENDATIONS

1. Public organizations should make clear agreements on who is responsible for correctly entering an algorithm into the register and for following up with stakeholders (whether citizens or third-party vendors) if issues arise.
2. Registers should contain internal and external reporting mechanisms for subpar entries (with little to no information) in order to hold relevant public entities accountable. One option is a contact form for each algorithm entry that allows citizens to contact individual public servants (who entered the data into the register) directly. The process should be embedded in the broader transparency strategy of the public entities in question.
3. Entries should contain information on stakeholders involved in the creation and use of a system and contain a directly visible means of redress – a way to indicate dissatisfaction with the process of an algorithmic decision and of easily gaining additional information without having to resort to legal action.
4. Public administrations should publish reports, partially redacted if necessary, of dissatisfied citizens' complaints with regard to specific algorithms. These reports should also explain why these decisions were made, why they were experienced as unfair, why they were or were not actually fair, and the final result of the complaints (e.g., overturned decisions).
5. Taking inspiration from the Dutch national register, public administrators building an algorithm register should engage in an open, inclusive, and iterative process with frequent public consultations and workshops to gather feedback. The results and the eventual changes made to the registers should also be published. The participants should ideally be a diverse cross-section of society.

## 4.7 CONCLUSION: JUST A PIECE OF THE PUZZLE

On the whole, the issues identified here imply that there is a strategy-purpose misalignment: public trust and algorithmic transparency cannot be achieved with flawed data presentation, lackluster design, limited accessibility, and little accountability. Our recommendations are of course ideal options, and we know that not all of them are necessarily equally realistic considering the limited resources of public administrations.

In conclusion, we would like to touch upon a few key tensions that policymakers looking to create or improve public algorithm registers must navigate. First, there is a clear tension between (1) experts demanding more detailed and auditable information in the registers; (2) the public needing information that pertains to them and that is explained at a non-expert level, while potentially being confused by information overloads; and (3) public administrations not having the resources, expertise, or

incentives (usually a mix of the three) to provide the necessary data. One possible solution is to create two versions of an algorithm register: one expert version (potentially with limited access) and one public version for all citizens. This would, however, increase the burden on public administrations even more. Experts could also be incentivized to elucidate and disseminate the contents of the ‘expert version’ of the register. Another potential solution would be to offer more insight into the place of an algorithmic decision in the broader (human-led) decision-making process of public administrations by, for example, visualizing the flow of relevant institutional procedures through accessible graphs.

Another key tension exists between transparency and preventing system-gaming (see Lepri et al., 2018). Theoretically, the registers are supposed to provide all relevant information on a certain algorithm, including how it works, how false positives are identified, how it impacts fundamental rights, and so on. However, in practice, this is not always feasible: knowing exactly how an algorithm works could incite people, essentially bad-faith actors, to game the system and use that knowledge to their advantage. A possible solution would be to not release any and all information publicly, but to work with trusted auditors who must be thoroughly vetted so as to prevent abuse.

Public administrations looking to create an algorithm register should, in our view, first make a strategic decision on whether they want (what we propose to call) a ‘mediated’ system or a ‘direct’ system. In the former, the register is aimed at fully informing experts in the most comprehensive way possible, and citizens must then rely on these experts informing the public through, for example, reports and news articles. In a direct system, the register is aimed at informing citizens directly in understandable language. A possible downside is that the effectiveness of the former type depends on the experts, while that of the latter depends on citizens actually taking the time to comb through its contents. If neither group is interested, there is little point in having a register in the first place.

Furthermore, we would like to stress that algorithm registers are just one piece of the puzzle. Registering and monitoring public algorithm use should be part of a bigger strategy of improving governance transparency and digital literacy. If the hype surrounding AI now is to become reality, then algorithmic transparency is not enough to ensure sustained trust in governments. Indeed, transparency without literacy is meaningless, like a map without directions: if citizens don’t understand how algorithms influence their lives and how algorithmic decisions can be flawed, there is little they can do to address injustices or mistakes. It is therefore paramount for citizen-centered registers to, on the one hand, use understandable language and tangible examples to show citizens how their lives are governed today and, on the other hand, to provide accessible paths towards feedback and redress against algorithmic decisions. Likewise, public servants could use algorithm registers as organizational tools, but only if they, too, understand what these registers mean and why they are important. Any public organization should therefore first establish concrete objectives for their register, get a clear picture of what meaningful transparency entails, and build digital literacy (or perhaps ‘algorithmic literacy’) throughout their organization.

## REFERENCES

- Ada Lovelace Institute, AI Now Institute, & Open Government Partnership. (2021). *Algorithmic accountability for the public sector*. <https://www.opengovpartnership.org/wp-content/uploads/2021/08/algorithmic-accountability-public-sector.pdf>
- Automated Decision-Making Systems in the Public Sector – Some Recommendations. (2022). *AlgorithmWatch*. Retrieved November 30, 2023, from <https://algorithmwatch.org/en/adm-publicsector-recommendation/>
- Cath, C., Jansen, F., & Philosophy Documentation Center. (2022). Dutch comfort: The limits of AI governance through municipal registers. *Techné: Research in Philosophy and Technology*, 26(3), 395–412. <https://doi.org/10.5840/techn202323172>
- Doelgroepenanalyse Algoritmeregister. (2023). <https://algoritmes.pleio.nl/files/view/e59fb733-51ca-4811-9b6e-1d89d348a5b3/2023.pdf>
- Fiche pratique: L’inventaire des principaux traitements algorithmiques – guides.etalab.gouv.fr. (2021). Retrieved November 30, 2023, from <https://guides.etalab.gouv.fr/algorithmes/inventaire/#dans-quels-cas-une-administration-doit-elle-realiser-un-inventaire-de-ses-algorithmes>
- Global rankings for digital skills – Wiley gap index. (2021). Wiley. Retrieved November 30, 2023, from <https://dsgi.wiley.com/global-rankings/>
- Haatja, M., van de Fliert, L., & Rautio, P. (2020). *Public AI registers: Realising AI transparency and civic participation in government use of AI*. Whitepaper. <https://algoritmeregister.amsterdam.nl/wp-content/uploads/White-Paper.pdf>
- In gesprek over het algoritmeregister met burgers. (2023, February 7). Algoritmes. <https://algoritmes.pleio.nl/news/view/759b74bd-4dcc-4d42-b969-82971d790812/in-gesprek-over-het-algoritmeregister-met-burgers>
- Kaminski, M. E. (2020). *Understanding transparency in algorithmic accountability (SSRN scholarly paper 3622657)*. <https://papers.ssrn.com/abstract=3622657>
- Krafft, T. D., Zweig, K. A., & König, P. D. (2022). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation Governance*, 16(1), 119–136. <https://doi.org/10.1111/regg.12369>
- Lepri, B., Oliver, N., Letouz’e, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy Technology*, 31(4), 611–627. <https://doi.org/10.1007/s13347-017-0279-x>
- Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3403301>
- Mehr, H. (2017). *Artificial intelligence for citizen services and government*. [https://ash.harvard.edu/files/ash/files/artificial\\_intelligence\\_for\\_citizen\\_services.pdf](https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf)
- Murad, M. (2021). *Beyond the “black box”: Enabling meaningful transparency of algorithmic decision making systems through public registers*. Thesis, Massachusetts Institute of Technology (MIT). <https://dspace.mit.edu/bitstream/handle/1721.1/139092/murad-mmurad-sm-idm-2021-thesis.pdf?sequence=1&isAllowed=y>
- New York City Office of Technology and Innovation. (2022). *Summary of agency compliance reporting of algorithmic tools*. nyc.gov. <https://www.nyc.gov/assets/oti/downloads/pdf/reports/2022-algorithmic-%20tools-reporting.pdf>
- P’enicaud, S. (2021, May 12). *Building public algorithm registers: Lessons learned from the French approach*. Open Government Partnership. <https://www.opengovpartnership.org/stories/building-public-algorithm-registers-lessons-learned-from-the-french-approach/>
- Report of the Pittsburgh Task Force on Public Algorithms. (2021). [https://www.cyber.pitt.edu/sites/default/files/pittsburgh\\_task\\_force\\_on\\_public\\_algorithms\\_report.pdf](https://www.cyber.pitt.edu/sites/default/files/pittsburgh_task_force_on_public_algorithms_report.pdf)

- van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. *Government Information Quarterly*, 39(3), 101714. <https://doi.org/10.1016/j.giq.2022.101714>
- van Noordt, C., & Tangi, L. (2023). The dynamics of AI capability and its influence on public value creation of AI within public administration. *Government Information Quarterly*, 40(4), 101860. <https://doi.org/10.1016/j.giq.2023.101860>
- Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector – applications and challenges. *International Journal of Public Administration*, 42(7), 596–615. <https://doi.org/10.1080/01900692.2018.1498103>

---

# 5 Responsible Adoption of Cloud-Based Artificial Intelligence in Health Care

## *A Validation Case Study of Multiple Artificial Intelligence Algorithms for Diabetic Retinopathy Screening in Public Health Settings*

*Mona Duggal, Anshul Chauhan, Ankita Kankaria,  
Preeti Syal, Vishali Gupta, Priyanka Verma,  
Vaibhav Miglani, Deepmala Budhija, and Luke Vale*

### 5.1 INTRODUCTION

Artificial intelligence is how we have been able to make giant strides in the field of healthcare. A record-breaking time to market for a pandemic vaccination bears testimony to AI's prowess. In the field of imaging and disease prevention, AI has proved to be equally formidable (Pfizer, 2021, p. 1). The current emphasis regarding the application of AI in ophthalmology lies primarily in the screening and management of prevalent diseases, including diabetic retinopathy (DR), age-related macular degeneration (ARMD), glaucoma, retinopathy of prematurity (ROP), age-related or congenital cataract, and retinal vein occlusion (RVO) (Doi, 2006, p. 3, 2007, p. 2). DR screening is often conducted through fundus examination performed by ophthalmologists, experienced eye technicians, or optometrists using standard fundus cameras (Ryan et al., 2015, p. 4). Consequently, there has been a growing inclination toward advancing AI algorithms for the automated analysis of retinal pictures for

DR screening (Ryan et al., 2015, p. 4), illustrating the potential to supplant human graders while providing similar diagnostic accuracy (Padhy et al., 2019, p. 5).

Several algorithms have been trained and designed to recognize specific patterns of DR (Bellemo et al., 2019, p. 6). IDx-DR was the first FDA-approved study using an automated image analysis algorithm for DR detection in a primary-care setting (Bellemo et al., 2019, p. 6). However, more evidence is available on the performance of these algorithms in real-world scenarios (Hogg et al., 2023, p. 8; González-Gonzalo et al., 2022, p. 9). Most of the automated algorithms are trained on datasets that are limited to certain geographical regions or ethnic groups, and their performance decreases when they are used in different settings (Romeo-Aroca et al., 2020, p. 10).

The direct application of AI in public health settings may result in misdiagnosis due to potential incompatibility with fundus photographs of patients in clinical settings (Li et al., 2022, p. 11). In developing AI-based diagnostic systems, it is essential to segregate the dataset into distinct subsets for training, validation, and testing purposes. The training and validation datasets are utilized to develop and optimize the algorithm. The testing set is employed to assess the real-world performance of the AI system within clinical environments (Li et al., 2019, p. 12, 2022, p. 11). The testing dataset must not contain overlapping data points with the training and validation datasets. Failure to adhere to this requirement may introduce biases and inaccuracies in the algorithm's performance (Li et al., 2019, p. 12, 2022, p. 11).

Other challenges in developing such systems lie in obtaining an extensive collection of retinal images for training and validating the algorithms. Obtaining a sufficiently large dataset poses various challenges, including confidentiality, data protection, and regulation compliance (Grzybowski et al., 2020, p. 13).

The commercial AI algorithms' validation steps are unavailable in the public domain. The performance of these algorithms decreases when we test them in real-world patient data. Moreover, principles of responsible adoption of AI in real-world settings should be discussed more. Hence, there is a need for external validation of these screening algorithms and their responsible adoption in real-world intended-use settings. This chapter discusses a case study undertaken to describe the steps of validation and responsible adoption of AI in Indian public health settings.

## **5.2 PROCESS OF ADOPTION OF AI IN PUBLIC HEALTH FACILITIES**

To ensure the responsible deployment of AI, it is essential to examine the following principles: safety and reliability, equality, inclusivity and non-discrimination, privacy and security, transparency, accountability, and the protection and reinforcement of human values. These principles were effectively implemented through a comprehensive methodology, which included obtaining ethics permission, carefully selecting study sites, determining appropriate sample size, providing adequate training to staff members, employing suitable hardware and software, conducting validation procedures, managing data effectively, seeking input from both patients and providers, and thorough data analysis (Table 5.1).

**TABLE 5.1**  
**Matrix Showing Methods Followed in Principles of Responsible Adoption of AI in the Present Study**

Activity	Safety and Reliability	Equality	Inclusivity and Non-Discrimination	Privacy and Security	Transparency	Accountability	Protection and Reinforcement of Human Values
Ethics approval and trial registration	✓				✓		✓
Formation of project advisory group for study implementation		✓			✓	✓	
Information on study site					✓		
Adequate sample size		✓	✓			✓	
Staff hiring and training							
Establishment of a dark room							
Written informed consent	✓						✓
Inclusion of patients irrespective of gender, caste, SES (recruitment)		✓	✓				
Information about hardware and software					✓		
Reliability and transparency of AI algorithms					✓		
Anonymized data of patients (grading and sharing)				✓			✓
Secure government-owned cloud server	✓						✓
Data management and analysis	✓						
Access of data to study team				✓			✓
Diagnostic performance of AI algorithms					✓	✓	
Feedback from patients and providers					✓	✓	✓

*Ethical approval and trial registration:* The study obtained its approval from the Institutional Ethics Committee (IEC) of the Post Graduate Institute of Medical Education and Research (PGIMER), in Chandigarh, India, IEC approval no-PGI/IEC/2020/001342. The study was conducted following the Declaration of Helsinki. The study was registered with the Clinical Trials Registry (CTRI), India (reg no-CTRI/2022/10/046185). Prior permission was obtained from the state and district health authorities to conduct the study in coordination with the health system. Written informed consent will be obtained from the consenting individuals.

*Formation of project advisory group:* A project advisory group was established to convene a series of meetings to guide the development of study tools, review the findings from the study, and consider the current evidence. The experts included professionals with expertise in ophthalmology, general medicine, and public health to guide the implementation of the study.

*Study site:* The study was conducted at a tertiary healthcare centre and a primary healthcare centre (PHC) in the district of North India. The centres were selected after consultation with the state health authorities. The recruitment of participants at the PHC was facilitated through a network of accredited social health activist (ASHA) workers. Simultaneously, the research staff approached the diabetic individuals inside the tertiary healthcare facility's endocrinology outpatient department (OPD). No alteration was made to the patient's regular care throughout the study recruitment procedure.

*Adequate sample size:* To ensure sufficient sample size and statistical power, the study's sample size was determined based on sensitivity and specificity estimates in the screening program (Arenas-Cavalli et al., 2022, p. 14). Taking the prevalence of DR as 17%, the non-gradable image rate as 18.4%, the non-response rate as 10%, the precision as 95%, and the margin of error as 5%, the sample size of 250 diabetic patients was estimated to achieve 70% sensitivity and 86% specificity (Rêgo et al., 2021, p. 15; Scanlon, 2017, p. 16; Key Findings, n.d., p. 17).

*Staff hiring and training:* Two optometrists underwent a two-week training program at the retina unit of a tertiary healthcare institute. The training covered standard operating protocols (SOPs), camera handling and image-capturing techniques, consenting procedures, data entry, picture segregation, patient counseling, and referral procedures.

*Establishment of a dark room:* To achieve physiological mydriasis, for a better retinal view, and to capture high-quality images, a darkroom was established at the PHC with the requisite approvals from the health facility in charge. Before undergoing fundus imaging, the participants were instructed to enter the dark room for 4–6 minutes to achieve physiological mydriasis.

*Recruitment of study participants:* People with diabetes aged 30 years and older who were willing to participate and who provided consent were included in the study. Patients suffering from significant physical or mental disabilities were excluded from the study. The participants were selected irrespective of gender, socio-economic status (SES), caste, or religion.

*Hardware and software:* The fundus photographs were captured on a 3Netra Classic Portable Benchtop Fundus Camera (Forus Health Pvt Ltd, Bangalore, India).



This camera was selected by considering the range of tabletop cameras available commercially in India and by reviewing the published literature in India.

*Participating AI algorithms:* Based on a scoping review, five commercially available Indian companies utilizing cloud-based AI methods for DR detection were invited, and four agreed to participate. One FDA-approved AI algorithm also participated but was later dropped due to extremely low specificity in the interim analysis.

The Retinal AI Diagnostic Software by Radical Health, Leben Care Health Services Private Limited, and SigTuple Technologies Private Limited participated and completed the study. The identity of companies and their algorithms were masked and randomly labeled A1, A2, and A3. Details of the study (e.g., study objectives, type of camera, image format, frequency of sharing images, and expected outcomes) were shared with each AI company before the study.

*Reliability and transparency of AI algorithms:* The companies were requested to provide details related to the training datasets encompassing patient demographics, race, ethnicity, camera specifications, camera operator, pupil condition, image quality, and the relevant grading methods. The companies provided no extra information about their training dataset, image annotation method, ungradable images, human graders, or DR grading procedures. Last, the training of the AIs was carried out primarily on high-end, good-quality cameras on both mydriatic and non-mydriatic eyes. The lack of information on the training datasets and representative data to train AI impacts the performance of AI algorithms and hence limits the applicability of AI-based healthcare tools.

*Fundus imaging:* Trained optometrists obtained two-field non-stereoscopic, non-mydriatic color fundus photographs (macula-centered and disc-centered) of each eye at a 45-degree field of vision (FOV).

*Image storing, segregation, and sharing process:* Patient images were stored on the in-built storage of the password-protected laptop. The trained optometrist segregated the images into right and left eye folders. Participant-related identifiers were removed, and the images were tagged with a unique identification number. The research staff uploaded the deidentified data on the secured (NITI-AWS) Amazon Web Service, AI platforms, and human graders were provided individual access to the images (Figure 5.1).

The image capture and storage process are described in Figure 5.1.

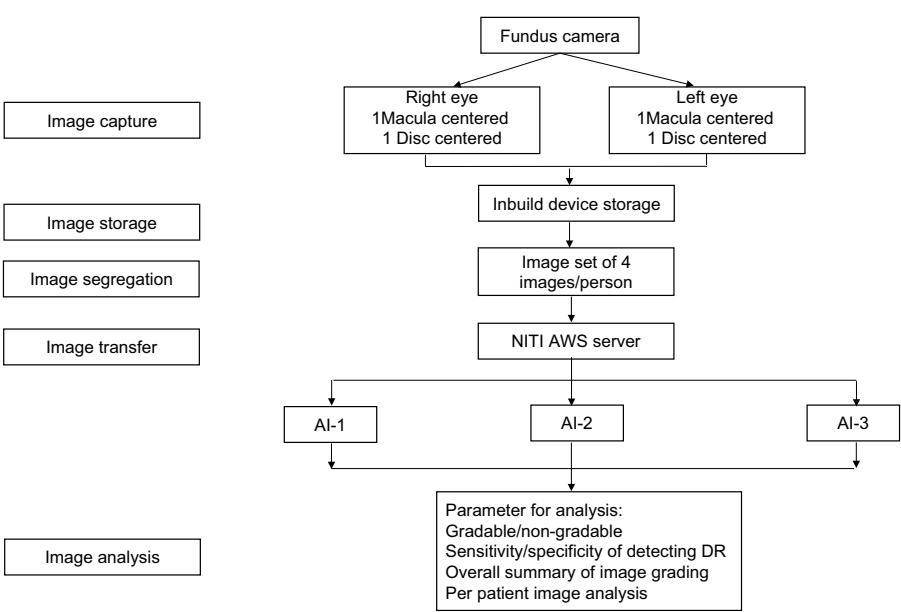
*Image grading process:* Grading refers to identifying visible signs of DR in each image to define the level of retinopathy. Human grader refers to a benchmark in comparison to which the output results of the AI platform were evaluated. The human grader (HG1: ophthalmologist trained in retinal image grading; HG2: optometrist trained in retinal image grading) assessed and assigned grades to each image. The vitreoretinal specialist resolved the disagreement between grades of HGs (1 and 2) and established an arbitrated dataset as the final grading. The images were graded according to the International Clinical Diabetic Retinopathy (ICDR) Classification (Wong et al., 2018, p. 18). Images were analysed for the following attributes: (a) image grade ability: “gradable” or “non-gradable,” (b) presence of DR: “yes” and no,” (c) grading of DR: “mild non-proliferative diabetic retinopathy (NPDR),” “moderate NPDR,” “severe NPDR,” and “proliferative diabetic retinopathy (PDR),” and (d) referral status: “yes” or “no.”

*Data modification for assessing AI algorithm performance:* All three AI screening outputs (DR grades) exhibited variability (Appendix 5.1), which posed a challenge for analysis. To evaluate the diagnostic accuracy and conclude, AI platforms were assessed based on their capacity to diagnose DR (yes/no).

*Data management and analysis:* All of the participant data were entered into REDCap (Harris et al., 2009, p. 19), exported in xls format, cleaned, and then imported into and analysed using STATA Version 15 SE (stataCorp, 2021, p. 20). REDCap is browser-based, metadata-driven electronic data capture (EDC) software for designing research databases (Harris et al., 2009, p. 19). It is a secure web-based application, and the data are encrypted and password-protected through the https protocol. A secure login and password are used for data extraction (Harris et al., 2009, p. 19).

The screening performance of the AIs was measured by sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) against the human grader results. The agreement was assessed using the Cohen kappa statistic between two graders for image gradability and DR detection. A third grader acted as an adjudicator to resolve the disagreement between the two graders. The final set of human grader results was used to compare the diagnostic accuracy of the AI platforms. Confidence intervals (95% CIs) were estimated for sensitivity, specificity, PPV, NPV, and accuracy. A P value < 0.05 was considered significant for all statistical tests.

*Performance of AI algorithms:* A total of 1099 retinal images of 500 eyes from 250 patient analyses were included in the study. The mean age of the study population was 55.74 (11.57) years, with a comparable gender distribution (51.60% females and 48.20% males).



**FIGURE 5.1** Process of image acquisition, segregation, and analysis.

There was a wide variation in performance metrics among the various AI systems, with sensitivity ranging from 59.69% to 97.74%, specificity from 14.25% to 96.01%, PPV from 30.16% to 86.67%, NPV from 85.00% to 94.34%, and accuracy from 37.19% to 88.43%. AI-3 had the highest specificity, PPV, accuracy, and agreement with the HG compared to the other AI algorithms. On the other hand, compared with AI-1 (97.74%) and AI-4 (94.66%), AI-3’s sensitivity was comparatively lower (Table 5.2). However, its NPV was nearly identical to the other algorithms (Table 5.2).

*Feedback from patients and providers:* The patients and providers were informed about the study procedures and the use of AI in DR screening during the study implementation. In-depth interviews (IDIs) were conducted among the key stakeholders, i.e., retina specialists, ophthalmologists, ASHA workers, and patients, to understand their views on implementing AI in health care.

*Strengths:* One notable strength of the study is the comparison of three distinct AI algorithms, which were evaluated using real-world data collected from Indian public health facilities representing real-world situations. Second, the validation set did not exclude ungradable images to assess the diagnostic accuracy of the AI algorithms. In most research studies, the analysis often excludes ungradable retinal images. Consequently, the AI algorithm’s performance can vary when implemented in real-world situations. Last, we used low-cost screening cameras (made in India) to capture non-mydratiatic images, reflecting current practice.

*Limitations:* The AI platforms provided varied DR grades compared to ICDR classification (18), which limited the validation to only binary outcomes (yes/no). Mild NPDR, moderate NPDR, severe NPDR, and PDR have different referral patterns and treatments. The AIs merged these categories (Appendix 5.1), which could lead to inaccurate documentation of disease status, changes in DR screening/ follow-up intervals, undue burden on public health settings (early referral), and missing immediate referral cases (late referral).

**TABLE 5.2**  
**Diagnostic Performance of Three Artificial Intelligence Algorithms to Detect Diabetic Retinopathy with the Arbitrated Grader Set**

DR (yes/no)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	Accuracy	Kappa
<b>HG vs. AI-1</b>	97.74 (93.05 – 99.42)	14.25 (10.85 – 18.45)	30.16 (25.91 – 34.77)	94.34 (83.37 – 98.53)	37.19%	0.04
<b>HG vs AI-2</b>	59.69 (50.68 – 68.12)	87.05 (83.15 – 90.63)	64.71 (55.35 – 73.09)	85.0 (80.45 – 88.34)	79.39%	0.42
<b>HG vs AI-3</b>	68.42 (59.71 – 76.05)	96.01 (93.24 – 97.72)	86.67 (78.31 – 92.26)	88.92 (85.21 – 91.81)	88.43%	0.65

DR: Diabetic retinopathy; CI: Confidence interval; PPV: Positive predictive value; NPV: Negative predictive value; AI: Artificial intelligence

### 5.3 CONCLUSION AND RECOMMENDATION

This research will provide new insights into the real-world implementation of an AI-assisted DR screening model in public healthcare settings. To our knowledge, this is the first planned prospective AI-based study from India with long-term policy implications. This is the first study where available AI algorithms were validated in field conditions and principles of responsible AI adoption were followed (Table 5.1). It is also one of the few studies that use Asian datasets addressing representation bias. Although the real-world precision differed significantly between the AI algorithms, AI-3 performed better.

The research study considered ethics and the possibility of bias during data acquisition, but efforts to obtain similar transparency from AI algorithm companies were futile.

Algorithm developers must maintain transparency in AI validation protocols, and the source of training datasets is at the heart of their design process. Furthermore, contextualizing and standardizing the training dataset to real-world settings is critical before introduction into any care setting.

The inclusion of ethical frameworks of AI and technology should become a part of the ethical review process as technology is increasingly becoming a part of clinical practice. Such frameworks must ensure that algorithms and training datasets are auditable, clinically validated, and explainable. Regular training and sensitization regarding identifying ethical AI before its implementation and clinical adoption in public health settings will ensure these measures are constantly checked.

In addition, as proposed by the NITI Aayog in their pioneer policy paper (“Operationalizing Responsible AI”), an oversight body must ensure risk-based assessment frameworks are developed for widespread adoption, particularly in resource-limited regions. Also, a guiding framework for procurement of such technologies must be curated and made available for wider adoption.

### REFERENCES

- Arenas-Cavalli, J. T., Abarca, I., Rojas-Contreras, M., Bernuy, F., & Donoso, R. (2022). Clinical validation of an artificial intelligence-based diabetic retinopathy screening tool for a national health system. *Eye*, 36(1), 78–85.
- Bellemo, V., Lim, G., Rim, T. H., Tan, G. S., Cheung, C. Y., Sadda, S., . . . Ting, D. S. W. (2019). Artificial intelligence screening for diabetic retinopathy: The real-world emerging application. *Current Diabetes Reports*, 19, 1–12.
- Doi, K. (2006). Diagnostic imaging over the last 50 years: Research and development in medical imaging science and technology. *Physics in Medicine & Biology*, 51(13), R5.
- Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4–5), 198–211.
- González-Gonzalo, C., Thee, E. F., Klaver, C. C., Lee, A. Y., Schlingemann, R. O., Tufail, A., . . . Sánchez, C. I. (2022). Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. *Progress in Retinal and Eye Research*, 90, 101034.
- Grzybowski, A., Brona, P., Lim, G., Ruamviboonsuk, P., Tan, G. S., Abramoff, M., & Ting, D. S. (2020). Artificial intelligence for diabetic retinopathy screening: A review. *Eye*, 34(3), 451–460.

- Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, 42(2), 377–381.
- Hogg, H. D. J., Brittain, K., Teare, D., Talks, J., Balaskas, K., Keane, P., & Maniatopoulos, G. (2023). Safety and efficacy of an artificial intelligence-enabled decision tool for treatment decisions in neovascular age-related macular degeneration and an exploration of clinical pathway integration and implementation: Protocol for a multi-methods validation study. *BMJ Open*, 13(2), e069443.
- Key Findings. (n.d.). <https://npcbvi.mohfw.gov.in/writeReadData/mainlinkFile/File341.pdf>
- Li, F., Chen, H., Liu, Z., Zhang, X., & Wu, Z. (2019). Fully automated detection of retinal disorders by image-based deep learning. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 257, 495–505.
- Li, S., Zhao, R., & Zou, H. (2022). Artificial intelligence for diabetic retinopathy. *Chinese Medical Journal*, 135(3), 253–260.
- Padhy, S. K., Takkar, B., Chawla, R., & Kumar, A. (2019). Artificial intelligence in diabetic retinopathy: A natural step to the future. *Indian Journal of Ophthalmology*, 67(7), 1004–1009.
- Pfizer. (2021). *How a novel “incubation sandbox” helped speed up data analysis in Pfizer’s COVID-19 vaccine trial | Pfizer*. Pfizer.com. [https://www.pfizer.com/news/articles/how\\_a\\_novel\\_incubation\\_sandbox\\_helped\\_speed\\_up\\_data\\_analysis\\_in\\_pfizer\\_s\\_covid\\_19\\_vaccine\\_trial](https://www.pfizer.com/news/articles/how_a_novel_incubation_sandbox_helped_speed_up_data_analysis_in_pfizer_s_covid_19_vaccine_trial)
- Rêgo, S., Dutra-Medeiros, M., Soares, F., & Monteiro-Soares, M. (2021). Screening for diabetic retinopathy using an automated diagnostic system based on deep learning: Diagnostic accuracy assessment. *Ophthalmologica*, 244(3), 250–257.
- Romero-Aroca, P., Verges-Puig, R., de la Torre, J., Valls, A., Relano-Barambio, N., Puig, D., & Baget-Bernaldiz, M. (2020). Validation of a deep learning algorithm for diabetic retinopathy. *Telemedicine and e-Health*, 26(8), 1001–1009.
- Ryan, M. E., Rajalakshmi, R., Prathiba, V., Anjana, R. M., Ranjani, H., Narayan, K. V., . . . Hendrick, A. M. (2015). Comparison among methods of retinopathy assessment (CAMRA) study: Smartphone, nonmydriatic, and mydriatic photography. *Ophthalmology*, 122(10), 2038–2043.
- Scanlon, P. H. (2017). The English national screening programme for diabetic retinopathy 2003–2016. *Acta Diabetologica*, 54(6), 515–525.
- StataCorp, L. (2021). *Stata statistical software: Release 17*. College Station, TX. [www.scirp.org/reference/referencespapers?referenceid=3587089](http://www.scirp.org/reference/referencespapers?referenceid=3587089)
- Wong, T. Y., Sun, J., Kawasaki, R., Ruamviboonsuk, P., Gupta, N., Lansingh, V. C., . . . Taylor, H. R. (2018). Guidelines on diabetic eye care: The international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*, 125(10), 1608–1622.
- Supporting AI at Scale in the APEC Region Through International Standards Discussion Paper for APEC SCSC Workshop – August 2023

# Appendix 5.1

Comparative variables	Outcomes	AI-1	AI-2	AI-3
Gradability	Gradable, Non-gradable	✓	✓	×
Diabetic Retinopathy	Yes/No	✓	✓	✓
Diabetic Retinopathy grading (ICDR)	Mild NPDR	Mild NPDR	Normal	Mild NPDR
	Moderate NPDR	Moderate NPDR	NPDR	Moderate NPDR
	Severe NPDR	Severe NPDR	PDR	PDR
	Proliferative DR (PDR)	PDR		
Macular Edema	Yes/No	✓	×	×
Referral	Yes/No	✓	×	×

**FIGURE 5.2** Comparative variables and outcomes of AI-1, AI-2, and AI-3.

---

# 6 Participation in AI

## *Notes from the Trenches*

*Tarunima Prabhakar, Cheshta Arora,  
and Arnav Arora*

### 6.1 INTRODUCTION

As AI systems become more closely intertwined in public lives, there have been calls to build these systems with greater public participation through the AI development lifecycle. The expert group report by IndiaAI emphasizes the Government of India's position of "developing AI in a responsible manner and building public trust in its use, placing the idea of 'AI for All' at its very core" (MEITY-GoI, 2023). The report recommends architecture, codebase and dataset exchanges that enable community participation. It also discusses engaging AI users in problem solving, feedback and innovation. Similarly, Microsoft's Responsible AI standards call to work with members of identified demographic groups to understand the risks of and impacts associated with differences in quality of service (Microsoft, 2022). Some proposals have called for inclusion of public participation throughout all stages of AI development (Gilman, 2023).

This increasing interest in public participation as well as recent AI literature acknowledges the socio-technical nature of all AI systems (Sambivasan et al., 2021; Hershovich et al., 2022). It is not enough that AI systems work but that they should work for specific demographics and fulfill certain social functions. From a socio-technical standpoint, inclusion of those who will be subject to the decisions of the AI can result in more representative datasets. Design features that keep humans in the loop can ensure that the model 'works' for specific demographics. AI can not only be exclusionary but outright discriminatory. Participation and incorporating multiple perspectives of affected groups or individuals in some meaningful way can help in detecting such potentially discriminatory outputs before deployment.

Participation in AI is also supposed to play a political function of distribution of power. Public participation is suggested to offset some of the risks of a black box technology by increasing scrutiny and building trust in the decisions of the AI system. Public participation in government deployment in AI is supposed to reinforce existing institutional norms of democratic governance.

In this chapter we reflect on building an AI model for detection of gendered abuse in India by centering and working with those affected by gendered abuse, as a part of Uli (<https://uli.tattle.co.in/>). The work was pursued as part of a larger project on end-user tools for mitigating the effects and enabling a collective response to gender-based violence. It is motivated by feminist principles that insist on "reflexivity,

participation, intersectionality, and working towards structural change” (Tandon, 2021). Through this reflection, we highlight the challenges of participation in practice and more broadly the gaps between responsible AI principles and the material conditions in which AI work happens. While recognizing the importance for participation, we identify practices and approaches that are essential to meet the goals of proposals of participation.

## 6.2 THE CALL FOR PARTICIPATION IN AI

The call for participation in AI is an extension of prevailing ideas of inviting participation in decision-making in governance and public projects. The World Bank Participation Handbook in 1996 noted that involvement and collaboration of local stakeholders “can not only make development efforts more effective and sustainable, but can also foster ownership and a sense of belief in the relevance and value of programs right down to the community level” (World Bank, 1996). The calls for participation, thus, are both normative and instrumental. The normative goals focus on participation as a goal in itself – through participation, disadvantaged groups can overcome power asymmetries to ensure that new projects don’t increase existing inequities and serve their interests. Others have framed it as filling a democratic deficit that strengthens the goals of liberal democratic societies (Cornwall, 2002). The instrumental argument for participation is that a project is more likely to be accepted if stakeholders are involved in the decision-making process (Cuppen, 2018). The acceptance could come from a more effective project that accounts for the needs of those affected by the project, as well as from trust in the process. This motivation is captured in the Pre-Legislative Consultative Policy in India that states that consultation can “resolve contentious and complex policies . . . where the government is seeking to create consensus” and support the “expectation for a transparent and better informed government” (Ministry of Law and Justice-GoI, 2014). There is a third goal where participation by stakeholders becomes a form of knowledge production – a way of generating more ideas for the project developers.

While AI developments have been underway for four decades, the calls for participation in creation of AI are recent and tied to the shift away from logic-based AI systems towards more data-driven paradigms (Birhane et al., 2022). Many AI systems, such as moderation or image recognition, have relied on non-expert contributions for data generation or annotation. Similar to participation in development projects, Birhane et al. (2022) identify instrumental and normative objectives for participation in AI. Participation in AI can aid in improvement of the algorithmic performance, the overall technical design. But it also allows for “collective exploration . . . around the needs, goals, and futures of a particular community.” The goals change how participation is defined and operationalized. Delgado et al. (2023) identify different approaches to participation in AI such as user-centered design, participatory design, participatory action research and mechanism design. These approaches have also been categorized along a spectrum or levels of participation (Delgado et al., 2023; Ada Lovelace Institute, 2021; Berdichevskaia et al., 2021). Some tiers of participation identified are consultation, contribution, collaboration and ownership, with each



reflecting a different level of stakeholder agency. The highest level of participation results when stakeholders have the ability to “make decisions about both the design of systems or policies, as well as the process by which such a system or policy is developed” (Delgado et al., 2023).

Cornwall (2002) locates citizen participation in two kinds of spaces: “created” and “invited.” In the former, people recognize and use their agency to address their own needs. In the latter, stakeholders are invited within a prevailing or pre-conceptualized project, lending them opportunities to participate. These two spaces can be in conflict with each other but can also support each other (Berry et al., 2019). This distinction is salient in the context of AI: calls for participation by shareholder-value-driven tech companies are squarely in the realm of invited participation. Many state and development projects are also invited spaces for citizens to share their opinions but with a bounded agency in steering the overall direction of the project. This includes proposals for ‘citizen juries’ to inform design decisions as well as alignment of values in AI (Center for New Democratic Processes, n.d.; van der Veer et al., 2021). The normative goals for calls for participation in AI, however, are to move closer to the ethos of created spaces where participation is not transactional but rather vibrant and “in constant engagement with their publics to increase community knowledge and empowerment” (Birhane et al., 2022).

We place the Uli project as an attempt at creating, rather than inviting, space for participation in AI. In all aspects of the AI design process – where should AI be deployed, what should the data aim to capture, the data collected and the model training – we aimed to engage the primary stakeholders of the Uli project: people of marginalized genders who inhabit online spaces and are often at the receiving end of abuse tied to their identity. The team that conceptualized the project also identified as this stakeholder. Organizationally, the project was co-initiated by two organizations working on digital rights, but throughout the project, we grew the number of organizations and individuals who participated in the project. This chapter is not an assessment on whether Uli was *truly* participatory, or to what extent it was participatory. Rather, it is an assessment of the hurdles that come in the way of trying to build participatory AI that aims to empower the people that use it – an issue that has received scant attention in literature so far.

### 6.3 PROJECT BACKGROUND

In 2021 a team of gender-rights researchers, social scientists and technologists at the Center for Internet and Society, India and at Tattle Civic Tech attempted to build an AI model for gendered abuse detection in India by working with those who were affected by it. The project was motivated by the disproportionate level of online harassment faced by people of marginalized genders in India even while platforms’ capacity to moderate in Indian languages remained insufficient.

We started with focus group discussions and interviews with more than thirty researchers and activists in India to understand their primary concerns and needs for intervention. The conversations made it clear that most problems did not need an AI solution. An AI intervention had to serve a modest social role as a tool that was not meant to ‘solve’ the problem of online gendered abuse but could help tackle the accumulated fatigue and exhaustion that activists, researchers and journalists experienced

in their everyday use of social media. Thus, it was specifically the feature on detection and redaction of abusive content that required machine learning. The Uli plugin also included other – non-machine learning based – features to address other needs heard during the redressal of online gender-based violence (oGBV). The development process of those features however is peripheral to the focus of this chapter.

The project was led by four people who identified as persons of marginalized genders in India – all the team members spoke English. Two of them were proficient in Hindi and English.

## 6.4 THE DIFFERENT STAGES OF BUILDING THE ULI ML MODEL

This section describes the different stages of building the AI model.

### 6.4.1 IDENTIFICATION OF NEEDS

Through July 2021–October 2021, the team conducted interviews over Zoom to get suggestions and feedback on the idea of a user-facing tool that could help in redressal of online gender-based violence. India is a country with twenty-one constitutionally recognized and several hundred non-official languages. The budget and team capacity allowed us to tackle only a small subset of these languages. While the initial plan was to focus on three Indian regional languages, several interviewees highlighted that several obvious cases of abuse in English used in India also escaped platform moderation. The specific way English was used in India, which included some transliteration of words from other languages and code-mixing, made it distinct enough from American or British English to merit specific attention. We converged on Hindi, Tamil and Indian English as the three languages that we would work on. The interviewees mentioned Twitter, Instagram, Facebook and WhatsApp as the common platforms they used. Each platform had its own linguistic culture. Given resource constraints and ease of collecting data, we decided to first look at abusive content on Twitter.

The interviews were followed by a round of focus group discussions in the three identified languages that sought responses from a group consisting of activists, academics, community influencers and journalists. The group was asked to respond to ten posts and discuss what kind of an action they would want the tool to take: redact an abusive post from a user's feed, flag it, report it or do nothing.

This exercise helped understand the type of content that the user would want the tool to act on and the trade-off between efficiency and usability. During the discussions, participants emphasized the contextual nature of online gender and sexual abuse. In their response to a particular post, they were concerned with who made a statement, to whom it was directed and the ongoing global and local events when the post was written. The location of moderation – ‘user-end’ as opposed to platform-end – shaped the respondents' views on how harm and abuse should be moderated. First, because the user could control the moderation (by possibly disabling the feature), participants in the qualitative research phase did not express concerns about excessive moderation through automation. Rather, participants mentioned that from the perspective of mitigating harm to the person harassed, it is acceptable if the machine learning model ‘over’-moderated on certain classes of speech such as hate speech.

Second, they mentioned that the model should be able to capture posts that escape moderation from platforms because they don't violate the laws or platforms' community guidelines but are still harmful. This included not only coded language but also harassment through repeated posting of even innocuous messages such as 'good morning' intended to remind a person that they are being watched.

#### **6.4.2 DATA COLLECTION AND CORPUS CREATION**

Building a machine learning model to detect abuse necessitated the need for annotated data with some instances of gendered abuse. To create this dataset, we used several heuristics; we crowdsourced a list of slurs and offensive words/phrases from the group of activists and researchers we interviewed. Additionally, we created a list of accounts that are often at the receiving end of hate online, as well as a list of accounts that are often found perpetuating hate and abuse on Twitter (now called X), by manually scanning conversations on the platform. This research was complemented by data shared by Arya et al. and Gurumurthy and Dasarathy, which contained a list of influential or highly active women on Twitter/X who are often at the receiving end of online abuse and harassment, as well as annotated data for different variants of potential harm online. Thus, we scraped tweets using three criteria: (1) crowdsourced slurs and keywords, (2) tweets by known perpetrators, and (3) replies to highly influential women on Twitter. In total, we were able to collect close to 1.3 million tweets in the three languages. These tweets were posted between 2018–2021.

Moving backwards from the need to have roughly 20% of tweets annotated by three people to create a robust testing dataset, we estimated that the project budget would allow for the creation of a dataset of roughly 24,000 posts, or 8,000 posts in each language. This calculation was based on setting a rate for annotation that matched the hourly rate for an early to mid-level researcher salary in the country.

To sample the 24,000 posts from 1.3 million tweets, we used some sampling techniques to create a dataset with sufficient representation of violating as well as non-violating posts in it, such that a model trained on it is able to understand and differentiate between gendered abuse and other language. To do this, we trained several language models on existing hate speech and toxicity detection datasets in the three languages, average score over multiple models and sample from both ends of the spectrum: extremely toxic and hateful to not toxic at all. A detailed description of this process can be found in Arora et al. (2023).

#### **6.4.3 SELECTING ANNOTATORS**

Annotators, six for each language, were selected from the pool of focus group discussion participants or their references. A couple of annotators who were active users of social media were also invited. All of the annotators identified themselves as women or as members of the LGBTQIA+ community. They were either involved extensively with marginalized communities facing discrimination and violence or had been at the receiving end of identity-based abuse themselves. Professionally, they identified as journalists, activists, university professors, peer-supporters, community influencers, or members of gender-rights-based organizations. This group was selected

consciously to ensure as much diversity of experience as possible in terms of identities, geographies, age and religion. This methodology of selecting annotators from within the affected communities borrowed from Waseem, which worked with the assumption that the annotators from affected communities annotate differently compared to a team of researchers.

#### 6.4.4 DEFINING GENDERED ABUSE

The focus group discussions provided an experiential description of gendered abuse, and they highlighted the everydayness of online gender-based violence, which is not noted in media stories. This experiential description had to be *translated* into a set of questions or labels that could inform the machine learning model. The team of four researchers/activist-researchers steering the project annotated posts in small batches. The annotations covered various typologies, including intersectional themes (such as ableist, transphobic, and queerphobic, body shaming) and types of abuse (sarcasm, threats, derogatory comments), considering the explicit or implicit nature of the abuse. This detailed labeling facilitated the team's understanding of the data and revealed disagreements within the team. Over three months, the team iteratively annotated data batches, refining the initial typologies to essential labels, and developing guidelines to clarify the label's purpose. The simplification of labeling was crucial to ease the work for activists and researchers who had other primary commitments. The team ultimately converged on two labels:

- Is the post gendered abuse?
- Does the post contain explicit language?

We created an annotation guideline with definitions for the labels, as well as examples. The guideline was initially written in English and then translated into Hindi and Tamil. The examples in the Hindi and Tamil guidelines were picked by the team members speaking the language to mirror the motivation for including the corresponding examples in the English guideline.

We paired annotators and asked them to annotate a hundred posts as per the guideline. Where they disagreed, we asked them to discuss their reasons for their choice of label. This exercise was repeated two to three times for each pair. While in some cases, the disagreement in label was a result of misunderstanding of the guideline, this process also highlighted experiences and interpretations of abuse that the team of four had not accounted for. We also learnt that absent any context to a post, such as the relationship between the person posting and the receiver (in case of replies), or the broader conversation, each annotator assumed context. This shaped whether they perceived the post as gendered abuse or not. To account for some of the complexity of context of a post, we broke the first label into two parts:

1. Is the post gendered abuse when not directed at a person of marginalized gender?
2. Is the post gendered abuse when it is directed at a person of marginalized gender?

The first label would capture outright misogynistic comments, such as those commenting on women's capabilities to participate in professional or public life. The second label reflected feedback of the expert annotators that any form of hate speech when directed at a person of marginalized genders is gendered abuse. The gender of the person the post is directed to cannot always be detected from the post itself, as in several Indian languages the verbs are conjugated by gender, which can signal the gender of the person the post is directed to. This, however, is not the case with English. Thus, a need was felt to account for the scenario when the gender was detected.

#### **6.4.5 COLLECTING ANNOTATIONS**

While the exercise for converging on a definition of gendered abuse was carried out through spreadsheets, a software format familiar to all the annotators we were working with, we didn't find this to be scalable for 24,000 posts. We developed a custom user interface (UI) that was accessible through a URL that could be opened on any browser. The UI was made responsive to enable annotations on mobile. The languages of the UI changed based on the language the annotator was working on. The posts were annotated between March 2022 and July 2022.

While the experts were compensated for the labor of annotating this content – it was tied to the number of posts annotated – their engagement was considered voluntary. They could stop the work whenever they wished to, without any contractual obligations. The posts were assigned to annotators in batches. When an annotator dropped out, their posts were allocated to other annotators without the expectation that the posts would necessarily be annotated. Consequently, all annotators were not allocated an equal number of posts, and we didn't get an 'ideal' dataset of 24,000 posts with 20% of posts annotated by three people.

#### **6.4.6 MODEL TRAINING**

Once the final dataset was curated and annotated with labels from expert annotators, we trained a language model for performing automatic detection of language that can be considered gender-based violence. We divided the dataset into train, validation and test splits with the corresponding binary labels for violating and non-violating posts. To generate the testing dataset, we had to decide on a method to resolve disagreements on the roughly 20% of the data that was annotated by three or more people. As Table 6.1 shows, there were significant disagreements among annotators. Furthermore, the degree of disagreement varied by language and label. On the whole, Tamil-speaking annotators agreed more with each other than did Hindi- and English-speaking ones.

While the common approach in machine learning is to take a majority vote on the label, researchers and annotators expressed a need to retain and value the disagreement because the disagreements could reflect a difference in experiences of abuse and a unique contextual position, which shouldn't be flattened through a majority vote.

The call for working with disagreements pushed the data science team to consider approaches for working with individual-level labels in designing a machine learning model (Mostafazadeh Davani et al., 2022). But, due to limited time for research

**TABLE 6.1**  
**Agreement Scores (Krippendorff Alpha)**

Language	Label	Values
English	Label 1	0.402
	Label 2	0.258
	Label 3	0.35
Hindi	Label 1	0.396
	Label 2	0.314
	Label 3	0.501
Tamil	Label 1	0.488
	Label 2	0.411
	Label 3	0.721

and development before deployment, we fell back on more traditional approaches of using individual labels for training, which is still under research.

To extract maximal performance out of the model, instead of purely training on our dataset, we started with a language model pre-trained on large amounts of web text such that it already possesses generic language understanding before we teach it how to detect language with gender-based violence. This is standard practice in the natural language processing literature, the subfield of machine learning concerned with developing models pertaining to language data. We experimented with versions of language models pre-trained on web text in Indic languages as well as ones trained on Twitter data. We trained our model on the training subset of the data and optimized its parameters using the validation set. Once the model was trained and optimized, we tested the performance on the test set, data that it had never been exposed to, in order to get an unbiased estimation of its performance. A detailed description of the model training process can be found in Arora et al. (2023).

**6.5 THE NEED FOR TRANSLATION**

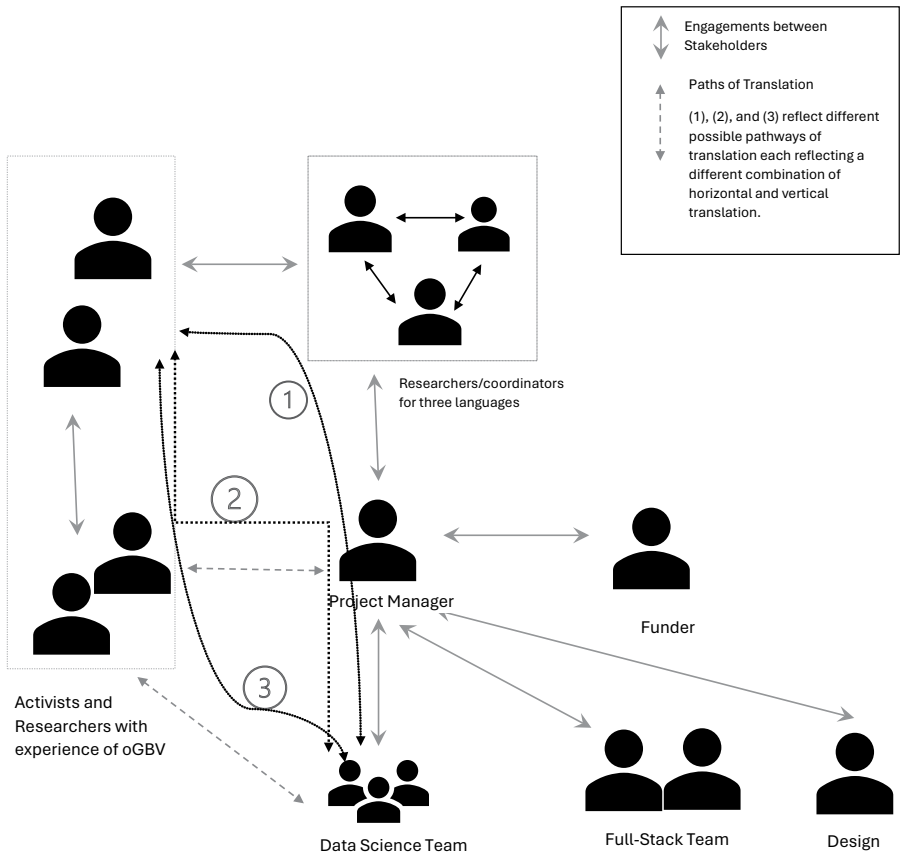
The process described here involved several stakeholders: the researchers who conceptualized and steered the project, the activists and researchers who participated in focus group discussions and created the dataset, the data science team that would use the data to train the model and the engineering team that developed the annotation UI and the plugin. Other stakeholders such as the design team and funders do not appear explicitly in the process as described, but they are also relevant stakeholders who shaped the final product.

Each of these stakeholders identified with different, often multiple, communities. Each of these communities of learning and practice speak different *languages* comprising terms describing shared concepts and experiences (Silver, 2012). In fact, the act of enunciating a shared language creates boundaries that allow for creation of a community. For example, data scientists understand a machine learning model performance in terms of F1 scores and ROC curves (receiver operating characteristic

curve), but these terms may be illegible to other communities, such as designers or grassroots activists.

Calls for participation in AI require engaging the community that is affected by the decisions of the machine learning model, but this boundary of a community likely describes several sub-communities that have their own specialized language. In the case of Uli, the boundary of a community was people who have experienced or responded to online gender-based violence. There are, however, other communities within this boundary. For example, people might identify with other sub-communities based on linguistic, regional, or gender or sexual identity within this boundary. The variation in agreement scores for the three languages speaks to the heterogeneity within a community.

Regardless of where the boundary of a community is drawn, and who constitute in-group and out-group stakeholders, there emerges a need for translation. Figure 6.1 describes the engagements among different stakeholders. In the case of Uli, the needs expressed by activists and researchers need to be translated into



**FIGURE 6.1** The different points of engagements and channels of translation between a community and an AI development team.

terms that data scientists and engineers understand, and the possibilities and constraints of AI to the activists and researchers. This translation is mediated via other stakeholders.

The translation is a continuous process. Although it is not possible to describe all translations in the thick process ('thick' here is used in the sociological sense of thick description of human action) of a product development, we present here a few instances to explain the concept:

1. Online abuse manifests in multiple ways. It includes malevolent speech, but also behaviors such as repeated commenting. Whether something is experienced as abuse is also shaped by who is posting the content and the broader online discourse in which a post is shared. Some of the community members believed AI was capable of detecting all such instances. In the focus group discussions, participants consistently emphasized that gendered abuse is contextual, but the data scientists needed a dataset with a few categorical labels, which grossly simplifies context. What is retained from the participants' expression of needs into a few labels emerges from an act of translation.

The technical team described what was technically feasible for a machine learning model to detect, which was translated into terms the community could understand (explicitly stated in the annotation guideline that, "this tool will not be able to act on all instances of misogyny") through examples of posts and possible action on it. The participants pushed back on certain simplifications, however. For example, people strongly felt that who the post was being directed to should be accounted for in the guideline. This resulted in amending the annotation guidelines (described in the section on 'Defining Gendered Abuse').

2. Any machine learning model will result in false positives and false negatives. These, however, were not terms familiar to qualitative researchers and activists engaged in the project. First, the four members steering the project had to understand these concepts from the data scientists and then translate this into concepts the community of activists and researchers could understand. The community articulated their need in terms of tolerance for over- and under-moderation and possible misuse of the model.
3. The annotation guidelines were originally written in English (this was the language all the language coordinators spoke). It was then translated into Hindi and Tamil by the research coordinators, who shared it with activists and researchers annotating the posts in those languages. At the annotation level, this involved translating disagreements among annotators to create a shared space of understanding, and translating the value of keeping those disagreements into the design of the machine learning model to the technical team. The research coordinators flagged that there was significant diversity in the annotator backgrounds, and disagreement on annotations should not be flattened out through a majority vote. Here the data scientist discussed newer approaches to training with individual-level annotations (described in the section on 'Model Training').



These specific examples show the layers of translation from needs of a community to specific technical implementations. Science, Technology and Society (STS) scholarship considers translation within scientific and technological processes to be a matter of serious concern. Translation is not so much about producing equivalence but rather about a method to actively “generate new meaning” (Sarukkai, 2013). It emerges within fraught relations that turn translations into sites of judgment and locations of continuous struggle (Law & Lin, 2017) but at the same time present a possible route to “partial connections”.

Participatory AI seeking to empower AI users aims to make the boundary between AI developers and AI users more porous. This necessitates a bi-directional translation. Furthermore, the translation is not just between a community and an AI development team but also takes place within the development team (Hoffman et al., 2023). Prior research on performance of software engineering teams surface the need for effective communication plans within teams and between teams of different cultures. For analytical purposes, we call this vertical translation – that which happens between different levels of abstraction – while the former as horizontal translation. Any real-world project emerges from a grid of horizontal and vertical translation. For example, an external community discusses its needs with project representatives, who discusses it with a project manager (horizontal translation) or an engineering manager, who further distills it into tasks for a data science team and engineering team (vertical translation). The constraints stated by an engineer may also percolate up the same communication chain to the community members on how a certain need may be reflected in an AI system. In another scenario, a community might have members with technical expertise (shared language with developers) who can deliberate with the technical team on specific implementation details. In such a scenario, the vertical translation between levels of abstractions might also take place within the community from whom participation is sought.

In a perfectly participatory system, every stakeholder involved in a project has a shared language, obviating the need for translations. But this is rarely the case, and unlikely to be true in large projects with specialization of roles. In the context of environmental assessment, Diduck and Sinclair note that limited information or overly technical information can be huge stumbling blocks to meaningful public engagement. To overcome such hurdles, Gilman (2023) recommends building technical expertise of the communities that one is working with for democratizing AI. There are crucial material facts that intercede with this proposal. First is the technical nature of machine learning, which is increasing in complexity and is inscrutable even to the engineers working on it. Second is that when the community itself is battling marginalizations on multiple fronts, getting familiar with the technical functioning of a system for a project might not be a priority. Evaluations of past projects that have sought participation note that attempts to expand public participation often backfire and produce more distrust or lead to “participation fatigue” (Berry et al., 2019). Participants can become exhausted from working on the project, and motivating them to still make useful contributions becomes more difficult (Swiner, 2022). In the case of Uli, all activists and researchers cannot be expected to know or learn about performance metrics of machine learning or the need for categorical variables

in machine learning. Thus, translation emerges as a critical function in interdisciplinary and participatory work, but it is also fraught with layers of opacity, tensions and possible risks of miscommunications. Translation generates partial connections across different stakeholders who bring with them a range of social, political and cultural imaginaries of AI, but is always enacted within a charged field.

The metaphor of translation also helps us understand the sites of participation: where in the AI development process does a community participate? The ideal scenario calls for participation in all aspects of the AI development process. But, through the Uli case study, we have attempted to foreground material realities, which surface the more practically grounded questions on what are the critical points of engagement? And which translatory efforts are essential for a participatory project? With a task with subjective assessments such as abusive speech detection, the annotations of specific posts – the politics of defining abuse aside – were an essential site of meaning-making and contestation. Thus, we made efforts to translate some of the current technical possibilities of machine learning models to the participants. The discussions during pairwise annotation exercises and examples in the annotation guidelines were translatory spaces where subjective experience met the simplification of a machine learning model. Such translatory needs could be lower for less subjective tasks such as object identification.

## 6.6 TRUST DESPITE OPACITY

Birhane et al. (2022) note that “participation cannot be expected to provide a solution for all concerns, and is not a solution for all problems.” They advocate for using participation for considered and specific goals as a tool in the responsible development of AI. They warn against ceding questions of democratic governance to participatory approaches in technological development, and co-option of participatory procedures by powerful entities to claim legitimacy to the detriment of the community whose participation is sought. In this chapter, we have attempted to show that even with best intentions, participatory processes in AI have to accept and manage a certain degree of opaqueness as being central to how different stakeholders engage with the process. Different actors participate at different intensities, and participatory research should be able to account for all kinds of involvement from different actors. The AI users won’t necessarily understand the models that were used to train the machine learning model, and the technical team may not always comprehend the political nature of the system being designed. The motivation to learn the technical language is also contingent on the perceived seriousness of the problem statement and of the decision that the AI tool is going to make. The act of translation can help a stakeholder, given their limited affective capacity, to form partial connections without having to interact with the entire system. It also shifts our understanding of participatory design as that which is based on a notion of relevant or partial opaqueness to account for different actors and their level of interest in the system.

Despite the layers of opacity, teams can aspire to build trust in an AI system. Trust, while intuitively felt, is a complex concept to explicate. It is, however, manifested when two stakeholders don’t fully know each other and the world the other inhabits

but feel like they know enough to allow the other stakeholder to act for their interests. Any new space that is created for participation bears traces of social relations and previous experiences of planned intervention in other spaces (Cornwall, 2002). “Simply creating a new institution is not enough to purge it of older associations; new spaces may come to be infused with existing relations of power, reproducing existing relations of rule”. Thus, the legacy of individuals and institutions steering a specific participatory AI project is consequential. Participants, even if they share language with AI developers, use several other heuristics to evaluate whether to trust the process: who is the funder, which actors will extract what kind of value from the system, who else will use the system? These were the questions our participant stakeholders asked us. Although a truly participatory AI project can ameliorate past associations, superficial or callous attempts at inviting or creating spaces for participation can also create or reinforce negative associations creating challenges for future projects.

Through our chapter we also emphasize that communities are not homogenous. Recognizing different affects, capacities, motivations, intentions and opinions becomes a relevant exercise to navigate these differences among various actors and how they approach a system. Assessment of past community initiatives surface different motivations for people participating in them. Some are motivated by a concern for the problem, some by a sense of belonging with the community and some by financial incentives (Herzele et al, 2013; Sloot et al., 2019). We similarly find explicit recognition of the time spent in participation to be important. Monetary compensation of effort is one crucial aspect of this. The conundrum of monetary compensation of meaningful participation, however, is to ensure that participation is not reduced only to monetary compensation (this would make it equivalent to MTurk or any data labeling platform). Participation can also be credited in public outputs such as research publications and datasheets for datasets. The interplay of intentions and motivations that result in deep and sustained participation remains an active area of research for the Uli team.

Another consequence of accounting for diverse motivations and capacities is accepting participation inequality. Participation inequality is the phenomenon that most significant contributions come from a small number of participants (Haklay, 2016). Yet, a participatory project should leave open the possibility for any person in the community to participate. This entails creating pathways of participation that are accessible, even if sparsely accessed. Furthermore, subtle forms of participation might not always be visible to the team working on the model. For example, there might be several silent followers who keep up with communications on a project without explicitly voicing their inputs. A disruption or lack of communication, however, will be registered and possibly expressed by exiting a project at a critical juncture (Hirschman, 1970).

Through an exposition of the machine learning work done through Uli, we hope to simultaneously increase confidence in the possibility of participatory AI, which increases agency and empowers communities to steer the decisions of automated systems affecting them, while also recognizing the complexities embedded in building such systems. Recognizing and accounting for the complexity allows for more authentic spaces for participation and the possibility for greater public trust in AI systems.

## REFERENCES

- Ada Lovelace Institute. (2021). *Participatory data stewardship*. <https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/>
- Arora, A., Jinadoss, M., Arora, C., George, D., Brindaalakshmi, Khan, H. D., Rawat, K., Ritash, D., Mathur, S., Yadav, S., Shora, R. S., Raut, R., Pawar, S., Paithane, A., Sonia, Vivek, Priscilla, D., Khairunnisha, Banu, G., Tandon, A., Thakker, R., Korra, R. D., Vaidya, A., & Prabhakar, T. (2023). *The Uli dataset: An exercise in experience led annotation of oGBV*. <https://arxiv.org/abs/2311.09086>
- Berditchevskaia, A., Malliaraki, E., & Peach, K. (2021). *Participatory AI for humanitarian innovation*. [https://media.nesta.org.uk/documents/Nesta\\_Participatory\\_AI\\_for\\_humanitarian\\_innovation\\_Final.pdf](https://media.nesta.org.uk/documents/Nesta_Participatory_AI_for_humanitarian_innovation_Final.pdf)
- Berry, L. H., Koski, J., Verkuijl, C., Strambo, C., & Piggot, G. (2019). *Making space: How public participation shapes environmental decision-making*. Discussion brief. Stockholm Environment Institute.
- Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM conference on equity and access in algorithms, mechanisms, and optimization*. Association for Computing Machinery.
- Center for New Democratic Practices. (n.d.). *Citizens' juries and artificial intelligence*. <https://www.cndp.us/citizens-juries-artificial-intelligence/>
- Cornwall, A. (2002). Locating citizen participation. *IDS Bulletin*, 33, i–x. <https://doi.org/10.1111/j.1759-5436.2002.tb00016.x>
- Cuppen, E. (2018). The value of social conflicts: Critiquing invited participation in energy projects. *Energy Research & Social Science*, 38, 28–32. <https://doi.org/10.1016/j.erss.2018.01.016>
- Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023). The participatory turn in AI design: theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization*. Association for Computing Machinery.
- Gilman, M. (2023, September). *Democratizing AI: Principles for meaningful public participation*. Policy brief. Data & Society. <https://datasociety.net/events/democratizing-ai-principles-for-meaningful-public-participation/>
- Haklay, M. E. (2016). *Why is participation inequality important?* Ubiquity Press.
- Hershcovich, D., Frank, S., Lent, H., Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., & Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th annual meeting of the association for computational linguistics*. Vol. 1: Long papers (pp. 6997–7013). Association for Computational Linguistics.
- Herzele, A., Gobin, A., Gossum, P., Acosta, L., Waas, T., Dendoncker, N., & Frahan, B. (2013). Effort for money? Farmers' rationale for participation in agri-environment measures with different implementation complexity. *Journal of Environmental Management*, 131, 110–120.
- Hirschman, A. O. (1970). *Exit, voice, and loyalty. Responses to decline in firms, organizations, and states*. London: Harvard University Press.
- Hoffmann, M., Mendez, D., Fagerholm, F., & Luckhardt, A. (2023). The human side of software engineering teams: An investigation of contemporary challenges. *IEEE Transactions on Software Engineering*, 49(1), 211–225.
- Law, J., & Lin, W. (2017). The stickiness of knowing: Translation, postcoloniality, and STS. *East Asian Science, Technology and Society: An International Journal*, 11(2), 257–269. <https://doi.org/10.1215/18752160-3823719>

- Microsoft. (2022, June). *Microsoft responsible AI standard (V2)* [PDF]. <https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf>
- Ministry of Electronics & Information Technology, Government of India. (2023). *India AI 2023, first edition: Expert group to ministry of electronics and information technology*. <https://www.meity.gov.in/content/indiaai-2023-expert-group-report-%E2%80%93first-editionthe-ministry-electronics-and-information>
- Ministry of Law and Justice (Legislative Department), Government of India. (2014). *Pre-legislative consultation policy*. <https://ddashboard.legislative.gov.in/documents/pre-legislative-consultation-policy>
- Mostafazadeh Davani, A., Diaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10, 92–110.
- Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining algorithmic fairness in India and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 315–328). Association for Computing Machinery.
- Sarukkai, S. (2013). Translation as method: Implications for history of science. In *The circulation of knowledge between Britain, India and China*. Leiden, The Netherlands: Brill. [https://doi.org/10.1163/9789004251410\\_014](https://doi.org/10.1163/9789004251410_014)
- Silver, M. (2012). Voice and stance across disciplines in academic discourse. In K. Hyland & C. S. Guinda (Eds.), *Stance and voice in written academic genres*. London: Palgrave Macmillan. [https://doi.org/10.1057/9781137030825\\_13](https://doi.org/10.1057/9781137030825_13)
- Sloot, D., Jans, L., & Steg, L. (2019). In it for the money, the environment, or the community? Motives for being involved in community energy initiatives. *Global Environmental Change*, 57, 101936.
- Swiner, C. (2022). *Citizen participation in public projects*. Master's thesis, TU Delft. <http://resolver.tudelft.nl/uuid:6d773a8c-9ee3-4ff7-a56f-716acae4e4e9>
- Tandon, A. (2021). Practicing feminist principles in AI design. *Feminist AI*. <https://feministai.pubpub.org/pub/practicing-feminist-principles>
- van der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., Oswald, M., & Peek, N. (2021). Trading off accuracy and explainability in AI decision-making: Findings from 2 citizens' juries. *Journal of the American Medical Informatics Association*, 28(10), 2128–2138. <https://doi.org/10.1093/jamia/ocab127>
- World Bank. (1996). *The World Bank participation sourcebook*. Washington, DC: World Bank.

---

# 7 Risk Assessment Methodology for AI Regulation and Navigating Liability Determination in an AI-Driven World *A Policy Paper on Risk Assessment*

*Aditya Mohan and Karthik Satishkumar*

## 7.1 INTRODUCTION

Artificial intelligence (AI), according to the authors, is the capability of a system to acquire knowledge through data patterns and apply this knowledge to assigned problem, solving tasks such as content generation, prediction, recommendation or decision. However, definition of AI is contested and varies significantly among sources such as the Organisation for Economic Cooperation and Development (OECD, n.d.a), European AI Act (European Union, n.d.) and National Institute of Standards and Technology (NIST, n.d.a). This multiplicity of definitions and lack of consensus indicates that AI is still an emerging technology, and work needs to be done to define it universally and establish its dimensions. This chapter sets out to find an approach for risk evaluation through practices in cybersecurity industry, and then distill an AI-appropriate approach and dwell on its application.

## 7.2 PERVASIVENESS OF AI

Artificial intelligence is a software-based capability and therefore can be built for any system that functions using digital data and an operating piece of code (stand-alone software or embedded software). This makes AI implementable in almost all modern industries and services, as most rely on software operation. This wide array of applications means AI comes in a variety of forms (e.g., text, image, speech

processing), which makes the task of creating a unified approach to risk assessment, governance, and regulation a challenge.

### 7.3 TRANSFORMATIVE APPEAL OF AI

AI's ability to allow scaling of tasks by huge proportions [contract processing (JPMorgan Chase, n.d.)] [medical image processing] and dealing with complex tasks [OCT 3D scan examination for eye issues (Association of Optometrists, 2018)] [vocalizations among animal species] makes it a financially irrefutable option.

### 7.4 LEAPS OF AI AND UNFORESEEN RISKS

Recent concern about generative AI, with its ability to generate realistic images (e.g., DALL.E3) and videos (3D Gaussian Splat) and to communicate with users through a bot in a human-like conversation (ChatGPT) and some foundation models such as Palm 2 (Google, n.d.b) have advanced the ground on reasoning, specifically Med-Palm, which is multi-modal (Google, n.d.a).

However, of interest is a recent technique that, despite being non-invasive, can translate neural activity into images, thus having the potential to intrude into human thoughts (Meta AI, n.d.). So, we have sufficient reason to believe that AI is breaking new ground and well beyond mundane business needs; hence, a risk evaluation methodology and its use by regulators is suggested as necessary.

**Complex product chain:** Commercialization of AI has happened primarily around or after the cloud platform became mainstream and the off-the-shelf monolithic software system's decline. This has meant more modularity and multiparty componentization of AI software. In such a scenario, it is cumbersome to establish the root cause of an incident to a specific component and establish liability of an AI incident.

**Black box and explainability issue:** With some AI techniques arises the problem of black box, i.e., a component or a group of components that can only be probed for outcome but not intermediate results at each stage of processing/assessment of input. Although some techniques such as decision tree allow easier interpretation, it may be difficult for a Bayesian network and very challenging for neural networks. Without explainability, the risk for an AI system not rendering itself to liability determination increases.

**Why at the Global Partnership for Artificial Intelligence (GPAI):** Due to the pervasiveness and cross-border applications of AI, it would be most effective if a framework for risk assessment and quantification of AI is universally agreed upon and applied by several member states to allow smooth commercial use of AI, its governance, insurance and legal supervision.

### 7.5 CYBERSECURITY APPROACH: A PRISM TO FOCUS FOR FUTURE AI REGULATION

Based on the preceding elaboration, the authors consider that AI needs a risk modelling and risk quantification framework. On the path towards standardisation and adoption, the authors propose that the cybersecurity industry provides a great

template for developing and adopting risk modelling and quantification frameworks as applicable to AI. Like AI, the cybersecurity industry in its nascent years encountered challenges in understanding risks and its quantification with increased IT adoption. Information technology adoption raised concerns about privacy and security whilst also raising fears in the societal space regarding ‘jobs being replaced by IT automation’. The advent of the internet, digital transformation and cloud transformation only increased the size and scale of this challenge for the cybersecurity industry in the new millennia.

To address these challenges, government, academia and enterprise organisations came together over multiple decades to create standardisation, enabling cybersecurity adoption with increased trust. Institutes such as the International Standards Organisation (ISO, n.d.a), National Institute of Standards and Technology (NIST, n.d.b), Information Systems Audit and Control Association (ISACA, n.d.), Open Worldwide Application Security Project (OWASP, n.d.), the European Union Agency for Cybersecurity (ENISA, n.d.), Payment Card Industry Data Security Standard (PCI SSC, n.d.) and others contributed to standardize cybersecurity domains, associated risk definitions and a risk quantification model under a number of globally adoptable cybersecurity frameworks. Global cybersecurity standards institutions further collaborated with governments to help define regulatory compliance laws, which went a long way in mandating minimum required cybersecurity and IT standards. Examples of such regulatory compliance frameworks are ISO-27001 (ISO, n.d.b), the Healthcare Insurance Portability and Accountability Act (HIPAA) in healthcare (HHS, n.d.b), the Payment Card Industry Data Security Standard (PCIDSS) in the financial services industry (PCI SSC, n.d.), and the European Union’s General Data Protection Regulation (GDPR, n.d.). These frameworks had an underlying audit compliance requirement built on the foundation of the “trust but verify” principle. Such regulatory compliance frameworks and standards, risk models and audit frameworks helped increase the confidence in IT systems and processes, leading to its exponential adoption and eventual digitisation across the globe.

The applicability of such global standards meant that enterprises and governmental agencies across the globe could trust the IT processes as measured by these common standards, leading to interoperability and better trust among such entities. This, in turn, has had a direct impact on significant benefits to societies in areas of education, health, social welfare, defence, employment, etc. IT digitization has improved productivity and uplifted millions of people with exponential gross domestic product (GDP) growth across the globe. The cybersecurity regulations and compliance frameworks have thus played a key enabler role in this global impact of IT. AI would thus benefit from similar global collaboration towards standardization and compliance frameworks to quantify and mitigate the risks of its adoption.

## 7.6 LEARNINGS FROM THE CYBERSECURITY INDUSTRY

The benefits of cybersecurity standardisation and its significant impact in IT adoption can be further understood specifically in a risk modelling and quantification framework, which in turn provides a great template for definition of similar standards in AI. The authors present the NIST risk management framework and how that has helped in areas such as risk determination, liability, insurance and protecting brand reputation.



7.6.1 NIST Risk Management Framework and Maturity Model

The NIST cybersecurity risk management framework (RMF) provides a process that integrates security, privacy and cyber supply chain risk management activities into the system development life cycle. The risk-based approach to control selection and specification considers effectiveness, efficiency and constraints due to applicable laws, directives, executive orders, policies, standards or regulations. The RMF approach can be applied to new and legacy systems, any type of system or technology (e.g., IoT, control systems), and within any type of organisation regardless of size or sector (NIST, n.d.c).

At the core, the RMF provides a seven-step process to quantify risk (listed in Table 7.1).

The framework helps organisations quantify their cybersecurity risks measured and monitored against a set of controls. The detailed discussion of the NIST RMF framework is outside the scope of this chapter, but readers are welcome to read through the framework here (NIST, n.d.c). What is important, however, is to understand that the RMF has helped in definition, quantification, measurement and monitoring of cybersecurity risks across organisations. The framework then presents a maturity model for organisations to be evaluated against a set of controls. The maturity model is now presented in the subsequent section.

7.6.2 Cybersecurity Capability Maturity Model

The NIST RMF is underpinned by the cybersecurity capability maturity model (C2M2), which provides a measurable maturity level indicator of companies against applicable risk management framework and controls: the higher the maturity level, the lower the cybersecurity risk.

The NIST C2M2 and capability maturity model is a globally accepted framework across industry and government sector verticals. The framework and maturity model is leveraged by cyber insurance providers to determine premiums, quantify

TABLE 7.1  
RMF Seven-Step Process

Prepare	Essential activities to <i>prepare</i> the organization to manage security and privacy risks
Categorize	<i>Categorize</i> the system and information processed, stored, and transmitted based on an impact analysis
Select	<i>Select</i> the set of NIST SP 800–53 controls to protect the system based on risk assessment(s)
Implement	<i>Implement</i> the controls and document how controls are deployed
Assess	<i>Assess</i> to determine if the controls are in place, operating as intended, and producing the desired results
Authorize	Senior official makes a risk-based decision to <i>authorize</i> the system (to operate)
Monitor	Continuously <i>monitor</i> control implementation and risks to the system

TABLE 7.2  
C2M2 Maturity Levels

Level	Name	Description
MIL1	Initiated	<ul style="list-style-type: none"><li>Initial practices are performed, but may be ad hoc</li></ul>
MIL2	Performed	<ul style="list-style-type: none"><li>Practices are documented</li><li>Adequate resources are provided to support domain activities</li><li>Practices are more complete or advanced than at MIL1</li></ul>
MIL3	Managed	<ul style="list-style-type: none"><li>Activities are guided by policy (or other directives)</li><li>Personnel have the skills and knowledge needed to perform their assigned responsibilities</li><li>Responsibility, accountability and authority for practices are clearly assigned to personnel with adequate skills and knowledge</li><li>The effectiveness of activities in the domain is evaluated and tracked</li><li>Practices are more complete or advanced than at MIL2</li></ul>

	Individual Functional Areas - Subject Matter Experts score their functional areas based on organization structure and for each function, category, and sub-category.			Scores - SME scores compared against independent core group.		Results - Combine scores and compare against targets set by organization. The resulting risk gap must be addressed.		
	Area 1 (i.e., Policy)	Area 2 (i.e., Network)	Area 3 (i.e., Applications)	SME Average	Core Group	Combined	Tier Target	Risk Gap
Identify								
Business	3	3	2	3	3	3	3	0
Asset	2	1	2	1	2	2	3	1
Governance	2	2	4	2	2	2	2	0
Risk Assess	2	2	2	2	2	2	2	0
Risk Management	2	2	2	2	2	2	3	1
Protect	2	1	1	1	1	1	3	2
Detect	2	2	2	2	2	2	3	1
Respond	1	1	2	1	2	1	3	2
Recover	2	4	3	3	3	3	4	1

Adapted from 'The Cybersecurity Framework in Action: An Intel Use Case'

FIGURE 7.1 NIST scorecard.

investment into cyber defence, make liability determination in cases of data and privacy breaches, and more.

An example NIST scorecard is shown in Figure 7.1.

AI would benefit from a similar risk management framework and a maturity model that allows a standardized quantification, measurement and reporting process against AI systems and its applicability in various domains and areas of use.

## 7.7 CYBERSECURITY STANDARDIZATION IMPACT

The cybersecurity risk score, underpinned by the NIST C2M2 and capability maturity model, is applied in various aspects of the cybersecurity industry. Two key aspects are presented in the following sections.

### 7.7.1 CYBER INSURANCE PREMIUM DETERMINATION

With increasing cybersecurity incidents and breaches, several companies procure cyber insurance to protect themselves against cyber incident liabilities. Most insurance companies endorse maturity models such as NIST C2M2, Australian Signals Directorate's Essentials 8 (Australian Cyber Security Centre, n.d.) maturity model (a variation of NIST C2M2), etc. The insurance premiums of companies with higher maturity levels, i.e., thereby lower cyber risk levels, are thus lower, providing a competitive advantage for companies.

Similarly, an AI risk score, underpinned by an AI safety determination model, will provide companies with a level of trust in the AI system adoption whilst lowering risks.

### 7.7.2 VALUATION OF BUSINESS

Recently, the United States Securities and Exchange Commission (SEC) adopted rules requiring registrants to disclose material cybersecurity incidents they experience and to disclose on an annual basis material information regarding their cybersecurity risk management, strategy, and governance [(SEC, 2023): SEC Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure by Public Companies].

“Whether a company loses a factory in a fire – or millions of files in a cybersecurity incident – it may be material to investors,” said SEC Chair Gary Gensler. A key aspect of the new rule is to also add Regulation S-K Item 106, which will require registrants to describe their processes, if any, for assessing, identifying and managing material risks from cybersecurity threats, as well as the material effects or reasonably likely material effects of risks from cybersecurity threats and previous cybersecurity incidents (SEC, 2023).

The disclosure includes processes and cyber risks managed at the board level and thus has a direct impact on the stock price of a company and thereby its valuation. Companies that disclose independently assessed higher cybersecurity maturity levels and risk management capability measured against frameworks such as NIST C2M2 or Essential 8 are valued higher than are companies with lower scores.

Companies that rely on AI system use as a core differentiation for their business will also benefit from a similar disclosure rule. For example, a company that adopts a trusted AI system with a real data-based generative model will be valued more than one that uses synthetic data.

## 7.8 AI RISK QUANTIFICATION

The cybersecurity risk quantification and methodologies presented in the prior sections provide an excellent template for the AI system risk quantification process. The authors of this chapter bring forth their experience in policy definition and the

cybersecurity industry to provide a view on modelling the AI risk quantification process. The risk score, like the cybersecurity maturity score, can then be used for various scenarios such as AI liability determination, AI safety regulations and so on.

7.8.1 AI SYSTEM DIMENSIONS

A risk profile for AI systems needs all of the system dimensions to be evaluated for their contribution to the risk. Such a list of dimensions/characteristics can be fully/partially derived from a survey of existing standards, for example, ISO 25059 (quality model for AI system), ISO 23053 (framework for artificial intelligence systems using machine learning) or trustworthy aspects from various regulatory/guidance publications in various countries such as the EU, Australia, Canada, UK and US (NIST, n.d.a; European Commission, n.d.a, n.d.b; UK Government, n.d.; HHS, n.d.a; Government of Canada, n.d.; Australian Government, n.d.). However, authors considered that the key to such selection should be that the approach is people-centric; is very simple to enumerate and use by industry, legislative bodies and regulatory bodies; and it has a wide (cross-border) consensus already established. Such criteria are satisfied by the OECD framework for the classification of AI systems (OECD, n.d.b) as it has the consensus of all member states already obtained, the approach as illustrated in Figure 7.2 is people-centric, and the criteria listed for each dimension are easy to understand for industry and governing authorities. One of the stated aims of this framework is to be used for risk assessment, and that is what the authors will do in the subsequent section.

7.8.2 RISK PROFILING TEMPLATE

Table 7.3 shows the OECD framework for the classification of AI systems. The authors have built upon this framework to add a scoring schema, provision for weight assignment to each dimension and then compute total score.

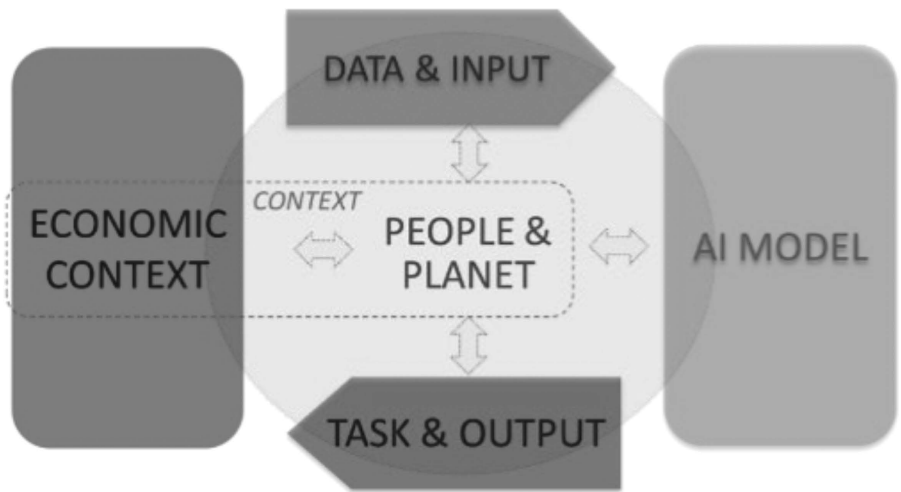


FIGURE 7.2 Key dimensions of an AI system, OECD.

**TABLE 7.3**  
**Classification Framework Dimensions and Criteria, OECD**

PEOPLE & PLANET	Criteria	REF	Description
USERS	Users of AI system	P_1	What is the level of competency of users who interact with the system?
STAKEHOLDERS	Impacted stakeholders	P_2	Who is impacted by the system (e.g., consumers, workers, government agencies)?
OPTIONALITY	Optionality and redress	P_3	Can users opt out, e.g., switch systems? Can users challenge or correct the output?
HUMAN RIGHTS	Human rights and democratic values	P_4	Can the system’s outputs impact fundamental human rights (e.g., human dignity, privacy, freedom of expression, non-discrimination, fair trial, remedy, safety)?
WELL-BEING & ENVIRONMENT	Well-being, society and the environment	P_5	Can the system’s outputs impact areas of life related to well-being (e.g., job quality, the environment, health, social interactions, civic engagement, education)?
DISPLACEMENT	{Displacement potential}	P_6	Could the system automate tasks that are or were being executed by humans?
ECONOMIC CONTEXT	Criteria	Description	
SECTOR	Industrial sector	E_1	Which industrial sector is the system deployed in (e.g., finance, agriculture)?
BUSINESS FUNCTION & MODEL	Business function	E_2	What business function(s) is the system employed in (e.g., sales, customer service)?
	Business model	E_3	Is the system a for-profit use, non-profit use or public service system?
CRITICALITY	Impacts critical functions/activities	E_4	Would a disruption of the system’s function/activity affect essential services?
SCALE & MATURITY	Breadth of deployment	E_5	Is the AI system deployment a pilot, narrow, broad or widespread?
	{Technical maturity}	E_6	How technically mature is the system? (Technology Readiness Level –TRL)
DATA & INPUT	Criteria	Description	
COLLECTION	Detection and collection	D_1	Are the data and input collected by humans, automated sensors or both?
	Provenance of data and input	D_2	Are the data and input from experts, provided, observed, synthetic or derived?
	Dynamic nature	D_3	Are the data dynamic, static, dynamic updated from time to time or real-time?
RIGHTS & IDENTIFIABILITY	Rights	D_4	Are the data proprietary, public or personal data (related to identifiable individual)?
	“Identifiability” of personal data	D_5	If personal data, are they anonymised, pseudonymised?
STRUCTURE & FORMAT	{Structure of data and input}	D_6	Are the data structured, semi-structured, complex structured or unstructured?

**Table 7.3 (Continued)**  
**Classification Framework Dimensions and Criteria, OECD**

PEOPLE & PLANET	Criteria	REF	Description
	{Format of data and metadata}	D_7	Is the format of the data and metadata standardised or non-standardised?
SCALE	{Scale}	D_8	What is the dataset’s scale?
QUALITY AND APPROPRIATENESS	{Data quality and appropriateness}	D_9	Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?
AI MODEL	Criteria	Description	
MODEL CHARACTERISTICS	Model information availability	M_1	Is any information available about the system’s model?
	AI model type	M_2	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?
	{Rights associated with model}	M_3	Is the model open-source or proprietary, self or third-party managed?
	{Discriminative or generative}	M_4	Is the model generative, discriminative or both?
	{Single or multiple model(s)}	M_5	Is the system composed of one model or several interlinked models?
MODEL-BUILDING		M_6	Model-building from machine or human knowledge?
	Model evolution in the field ML	M_7	Does the model evolve and/or acquire abilities from interacting with data in the field?
	Central or federated learning ML	M_8	Is the model trained centrally or in a number of local servers or “edge” devices?
MODEL INFERENCE	{Model development/ maintenance}	M_9	Is the model universal, customisable or tailored to the AI actor’s data?
	{Deterministic and probabilistic}	M_10	Is the model used in a deterministic or probabilistic manner?
	Transparency and explainability	M_11	Is information available to users to allow them to understand model outputs?
TASK & OUTPUT	Criteria	Description	
TASKS	Task(s) of the system	T_1	What tasks does the system perform (e.g., recognition, event detection, forecasting)?
	{Combining tasks and actions into composite systems}	T_2	Does the system combine several tasks and actions (e.g., content generation systems, autonomous systems, control systems)?
ACTION	Action autonomy	T_3	How autonomous are the system’s actions and what role do humans play?
APPLICATION AREA	Core application area(s)	T_4	Does the system belong to a core application area, such as human language technologies, computer vision, automation and/or optimisation or robotics?
EVALUATION	{Evaluation methods}	T_5	Are standards or methods available for evaluating system output?

### 7.8.3 SCORING MECHANISM

For a scoring mechanism, each characteristic (based on its descriptive question) is to be scored with a numerical value based on the AI use case under assessment. The authors considered several value ranges. Negative 1 to positive 1 is not a suitable range as (negative 1 would have implied the characteristic mitigates risk and for similar reason a value of zero would not be compatible either as all characteristics contribute to risk in some measure except P\_3). Therefore, authors considered values in the positive range 1 and more. For illustrative purposes, a simplification has been considered to denote the criteria contributing to more risk with a value of 2 and when contributing to less risk with a value of 1 except the case of P\_3.

P\_3 is a binary criterion as it identifies the possibility to opt out of AI, and the use case can very clearly be demarcated to a YES/NO value and the fact that ability to opt out of the use of an AI system nullifies the AI risk, so the options are 0 or 1. If other cases like live correction of AI system output is possible, then more values can be considered, e.g., 0, 1 and 2.

Criteria like E\_1 cannot be answered with a value approach of 1 or 2. A more graded approach is possible depending on regulatory choice; however, the authors for illustrative purposes have chosen 1 for a non-critical sector and 2 for a critical sector of the industry. A complete list for reference is provided in Appendix 7.1.

### 7.8.4 WEIGHTS FOR EACH DIMENSION

Weights can be applied to each criterion of each dimension, or they can be applied to the overall dimension sub-total. Weights allow the regulator/implementor of the scoring mechanism to choose the significance of each dimension, for example, AI systems deployed in government/public services would have a higher weight for the 'People' dimension than several private deployments. Similarly, 'Data Collection' would have a higher weight for a weather forecasting AI system and 'Action autonomy' T\_3 would be significant enough to increase the weight of 'Task and Output' for an autonomous driving system (ADS).

Another approach to weights drawing on from cybersecurity literature (NIST, n.d.b) is to consider the severity of consequence if the criteria is mismanaged or if the criteria is likely (frequently) not under control.

Having set a scoring template and weights, the risk score is computed as follows:

$$Risk\ Score = \sum_{k=1}^n (S[k] \cdot W[k])$$

In this risk profile framework:

$n = 5$  (five dimensions for the purpose of this chapter)

$S[k]$  is the sub-score of specific dimensions.

$W[k]$  is the weight of specific dimensions.

### 7.8.5 EVALUATION SCENARIO

The authors chose a specific AI technology application to evaluate risk and recommend that applications be evaluated for risk score and not the underlying technology itself.

The specific technology application is facial recognition (FR), and the authors use two different deployments of facial recognition to compare and illustrate the risk profile in the two use cases, i.e., facial recognition for access control (AC) to private buildings and facial recognition by law enforcement (LE) for monitoring public spaces.

The risk profile computation is available for the two use cases in Appendix 7.2 and Appendix 7.3. The only difference to the use cases of facial recognition is that for the AC use case all weights are 0.2, whereas for the LE use case the ‘People’ dimension is more significant and has a weight of 0.3, and the economic context is less important and has a weight of 0.1. As previously highlighted, the weight selection is up to a regulatory authority, and weight of all dimensions may or may not add up to 1 depending on the criteria for weight.

Comparative results are as shown in Figures 7.3–7.5. Due to the high risk in the ‘People’ dimension, the risk profile area in the graph is larger and more sharply expanded towards the ‘People’ dimension in the spider graph of the LE use case as

Facial Recognition - Access Control

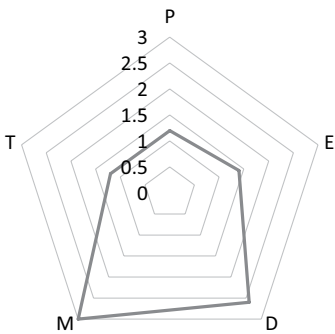


FIGURE 7.3 Risk profile representation – Facial recognition: Access Control.

Facial Recognition - Law Enforcement

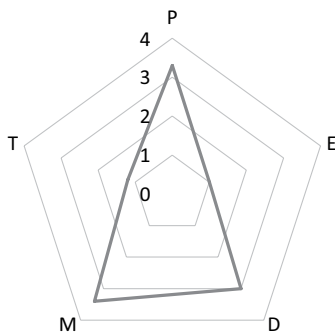
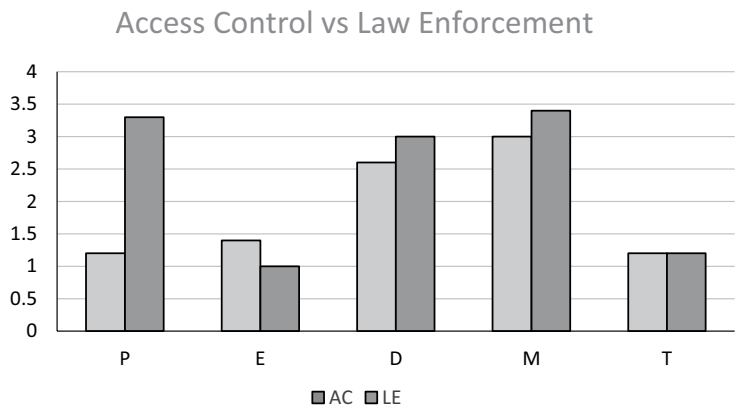


FIGURE 7.4 Risk profile representation – Facial recognition: Law Enforcement.





**FIGURE 7.5** Risk profile: Comparison by dimension.

**TABLE 7.4**  
**Risk Profile: Comparison by Dimension**

	Access Control	Law Enforcement
People and Planet	1.2	3.3
Economic Context	1.4	1
Data and Input	2.6	3
Model	3	3.4
Task and Output	1.2	1.2
TOTAL	9.4	11.9

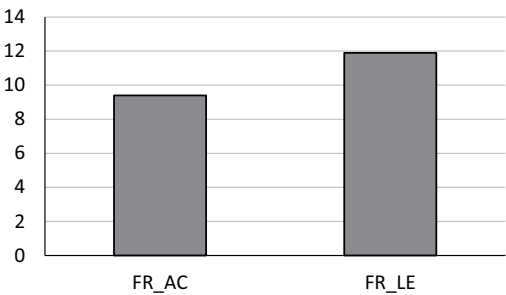
compared to the AC use case. Such a risk profile score and dimensional representation, the authors believe, would make risk assessment intuitive for a regulatory authority. For example, due to such visual representation of dimensions, it would be easy to spot an outlier in a specific AI technology use either through the spider representation or through dimension comparison in a bar graph representation (Figures 7.3–7.5).

The authors specifically highlight that a regulator may choose to classify an AI system as inoperable, either based on the total score or the score of each dimension or even the score of any specific criteria within a dimension.

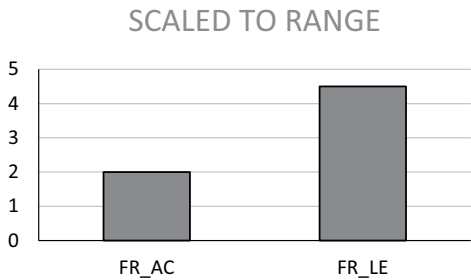
The authors further recommend that the total score for a risk profile should be scaled within the range of maximum and minimum possible score instead of being used as absolute values for easy and fairer comparison. The benefit of this approach is evident in Table 7.5.

**TABLE 7.5**  
**Risk Score: Net and Scaled**

Possible Highest Risk Score	14.8	
Possible Lowest Risk Score	7.4	
	NET SCORE	SCALED
FR: Access Control	9.4	2
FR: Law Enforcement	11.9	4.5



**FIGURE 7.6** Net score comparison.



**FIGURE7.7** Scaled score comparison.

**7.9 APPLICATION TO INSURANCE**

In the OECD’s paper ‘Enhancing the role of Insurance in Cyber Risk Management’ (Shetty et al., 2023), it is specifically stated that: “The insurability of a given risk is usually economically viable only where certain criteria (or “principles of insurability”) are generally met,” and among these criteria is listed that “Risk must be quantifiable” and probability, severity, impact (and subsequent recovery) become a factor in insurance premium.

Now having established a mechanism for an AI system's risk quantification, we can see how it is beneficial to insurance determination:

**Probability:** In the early days of regulation, there may not be enough historical AI incident data to establish probability in the wider industry; however, estimates for individual dimensions can be established, for example, D\_4, D\_5, and large-scale sourced data may show a percentage that contains personally identifiable information (e.g., sourced images), or a percentage of data may have missing ownership, consent metadata. Not all dimensions would necessitate a probability determination, and a supervisory authority may establish probability, prioritising those dimensions where an incident can be severe. Among these dimensions, those that have a high probability of adverse status will be those that impact an insurance premium.

**Severity:** Violation of human rights or risk to life, P\_4, P\_5, may bring severe economic and legal consequences for an operator of the AI system and therefore the severity ramification of these dimensions for insurance premium determination will be high.

**Impact:** A breach of personally identifiable information may have large or small impact depending on the exposure of the system, i.e., whether the system is used in a restricted group or is a public-facing system. Similarly, T\_3, Action Autonomy may have a high impact in the case of an incident with ADS on public roads; however, it may have low impact for a vehicle confined to a warehouse. This may consequentially also impact the *cost of recovery* from an incident. Together, impact and recovery will affect the insurance premium determination.

The more severe the possibility of incident type, the higher may be the impact and consequential cost of recovery/remedy. So, insurance cost would be proportional to the product of probability and cost of recovery/remedy.

Viability of insurance (Shetty et al., 2023) depends on the size of the industry over which the risk is spread. The wider the size of industry/jurisdictions that adopt this risk quantification framework, the more consistent will be the insurance premium, as it would arise from a larger data of probability, severity, impact and recovery trends for risk dimensions of AI systems. This will bring down the premiums and benefit the operators of AI systems, in turn bringing down the cost for end-users/consumers.

## 7.10 REGULATORY RECOMMENDATION

Having established the need for risk quantification, and having created a methodology for a risk profile, the authors have the following recommendation for the implementation of this framework:

Any business interested in launching its AI product/service into the market or importing an AI product/service from outside the jurisdiction, would apply to the AI authority for risk assessment and risk profile generation.



## 7.11 APPLICATION TO LIABILITY DETERMINATION

The availability of a risk score and the profile of risk (distribution of risk among dimensions of the AI system) gives law enforcement agencies a view of the governance applied in designing and producing the product.

Incidents occur because of an inherent fault in the *design of the product* or fault introduced during the *process of production*. There is, however, one more aspect to incidents, which is *incorrect usage* arising out of *wilful misuse or misinterpretation of documentation* of the product. Furthermore, to establish that liability determination is needed, a link has to be established between incident and defect (European Commission, n.d.c), where again a risk profile can help. For example, an incident where a member of the public reports that their personal data has been made public by a product may not align with the claim if the risk assessment documents that in dimension D\_4, only non-PII (personally identifiable information) data was used in model building.

The risk score mechanism described in this chapter covers aspects of design in P\_3 (optionality of use) and D\_6 (structure of data and input) among others. Also considered are aspects of production, for example, M\_6 to M\_8, which focus on model building. The documentation is covered by M\_1. Therefore, with risk scores in each dimension, law enforcement gets to make a more informed start in the investigation towards liability determination or to seek more data from the producer or deployer of the product/service.

However, liability determination is of value to businesses as well since it articulates product risk and therefore helps businesses in cases where liability may not be of the AI producer but the deployer, for example, where an entity uses AI API services to build/modify a product in ways contrary to recommendation/documentation of the provider (e.g., usage of a visual classifier intended for coloured images to identify objects in greyscale X-ray images), leading to inaccuracies. Furthermore, a risk profile may help evidence the good governance decisions of design and production, therefore protecting the producer to the extent of informing law enforcement where every effort was made towards a robust product.

## 7.12 CONCLUSION

A unified and coordinated (such as at a platform like the Global Partnership on Artificial Intelligence) approach to risk assessment and a standardised risk scoring template would help governments and regulatory authorities to provide safe use of AI systems to members of the public in their jurisdiction, as well as across jurisdictions when products/services are deployed beyond borders. Such an approach would help businesses operate in a predictable regulatory environment, encouraging investment in even high-risk AI use cases, as well as make exports, compliance and compatibility in several jurisdictions easier to achieve.

## REFERENCES

Association of Optometrists. (2018). *DeepMind algorithm vs retina specialists*. <https://www.aop.org.uk/ot/science-and-vision/research/2018/08/15/jaw-dropping-deepmind-algorithm-on-par-with-retinal-specialists>

- Australian Cyber Security Centre. (n.d.). *Essential eight maturity model*. <https://www.cyber.gov.au/resources-business-and-government/essential-cyber-security/essential-eight/essential-eight-maturity-model>
- Australian Government. (n.d.). *AI ethics framework*. <https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework>
- European Commission. (n.d.a). *Ethics guidelines for trustworthy AI*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- European Commission. (n.d.b). *EU AI act draft*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>
- European Commission. (n.d.c). *EU defective products and liability*. [https://europa.eu/youreurope/business/dealing-with-customers/consumer-contracts-guarantees/defective-products/index\\_en.htm](https://europa.eu/youreurope/business/dealing-with-customers/consumer-contracts-guarantees/defective-products/index_en.htm)
- European Union. (n.d.). *EU AI act*. [https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC\\_1&format=PDF](https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF)
- European Union Agency for Cybersecurity (ENISA). (n.d.). *ENISA official website*. <https://www.enisa.europa.eu/>
- General Data Protection Regulation (GDPR). (n.d.). *GDPR official website*. <https://gdpr-info.eu/>
- Google. (n.d.a). *Med-PaLM AI model*. <https://sites.research.google/med-palm/>
- Google. (n.d.b). *PaLM 2 AI model*. <https://ai.google/discover/palm2/>
- Government of Canada. (n.d.). *Algorithmic impact assessment framework*. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>
- Information Systems Audit and Control Association (ISACA). (n.d.). *ISACA official website*. <https://www.isaca.org/>
- International Organization for Standardization (ISO). (n.d.a). *ISO 27001 security standard*. <https://www.iso.org/standard/27001>
- International Organization for Standardization (ISO). (n.d.b). *ISO official website*. <https://www.iso.org/home.html>
- JPMorgan Chase. (n.d.). *COIN at JP Morgan Chase*. <https://mvvsp1.5gcdn.net/a92c8e20ce5a48f49ead0392e1170b3d>
- Meta. (n.d.). *Decoding brain activity with AI*. <https://ai.meta.com/blog/brain-ai-image-decoding-meg-magnetoencephalography/>
- National Institute of Standards and Technology (NIST). (n.d.a). *Language of trustworthy AI*. [https://docs.google.com/spreadsheets/d/e/2PACX-1vTRBYglcOtgaMrdF11aFxFEY3EmB31zslY14q2\\_7ZZ8z\\_1IKm7OHtF0t4xIsckuogNZ3hRZAaDQuv\\_K/pubhtml](https://docs.google.com/spreadsheets/d/e/2PACX-1vTRBYglcOtgaMrdF11aFxFEY3EmB31zslY14q2_7ZZ8z_1IKm7OHtF0t4xIsckuogNZ3hRZAaDQuv_K/pubhtml)
- National Institute of Standards and Technology (NIST). (n.d.b). *NIST official website*. <https://www.nist.gov/>
- National Institute of Standards and Technology (NIST). (n.d.c). *Risk management framework*. <https://csrc.nist.gov/projects/risk-management/about-rmf>
- Open Web Application Security Project (OWASP). (n.d.). *OWASP official website*. <https://owasp.org/>
- Organisation for Economic Co-Operation and Development (OECD). (n.d.a). *OECD definition of AI*. <https://oecd.ai/en/ai-principles>
- Organisation for Economic Co-Operation and Development (OECD). (n.d.b). *OECD framework for classification of AI systems*. <https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm>
- Payment Card Industry Security Standards Council (PCI SSC). (n.d.). *PCI security standards*. <https://www.pcisecuritystandards.org/>
- Shetty, S., Burkart, R., & Schmidli, H. (2023). Modelling and pricing cyber insurance. *European Actuarial Journal*, 13(2), 213–234. <https://link.springer.com/article/10.1007/s13385-023-00341-9>
- UK Government. (n.d.). *Understanding artificial intelligence ethics and safety*. <https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety>

- U.S. Department of Health and Human Services (HHS). (n.d.a). *HHS trustworthy AI playbook*. <https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf>
- U.S. Department of Health and Human Services (HHS). (n.d.b). *HIPAA regulations*. <https://www.hhs.gov/hipaa/index.html>
- U.S. Securities and Exchange Commission (SEC). (2023). *SEC press release on cybersecurity disclosures*. <https://www.sec.gov/news/press-release/2023-139>

# Appendix 7.1 Scoring Scheme

PEOPLE & PLANET	Criteria	REF	Description	Scoring Schema (1)	Scoring Schema (2)
USERS	Users of AI system	P_1	What is the level of competency of users who interact with the system?	HIGH=1	LOW=2
STAKEHOLDERS	Impacted stakeholders	P_2	Who is impacted by the system (e.g., consumers, workers, government agencies)?	GROUP=1	PUBLIC=2
OPTIONALITY	Optionality and redress	P_3	Can users opt out, e.g., switch systems? Can users challenge or correct the output?	YES=1	NO=2
HUMAN RIGHTS	Human rights and democratic values	P_4	Can the system’s outputs impact fundamental human rights (e.g., human dignity, privacy, freedom of expression, non-discrimination, fair trial, remedy, safety)?	NO=1	YES=2
WELL-BEING & ENVIRONMENT	Well-being, society and the environment	P_5	Can the system’s outputs impact areas of life related to well-being (e.g., job quality, the environment, health, social interactions, civic engagement, education)?	NO=1	YES=2
DISPLACEMENT	{Displacement potential}	P_6	Could the system automate tasks that are or were being executed by humans?	NO=1	YES=2
ECONOMIC CONTEXT	Criteria	Description		Scoring Schema (1)	Scoring Schema (2)
SECTOR	Industrial sector	E_1	Which industrial sector is the system deployed in (e.g., finance, agriculture)?	NON-CRITICAL=1	CRITICAL=2
BUSINESS FUNCTION & MODEL	Business function	E_2	What business function(s) is the system employed in (e.g., sales, customer service)?	NON-CRITICAL=1	CRITICAL=2
	Business model	E_3	Is the system a for-profit use, non-profit use or public service system?	NON-PUBLIC=1	PUBLIC=2

(Continued)



*(Continued)*

PEOPLE & PLANET	Criteria	REF	Description	Scoring Schema (1)	Scoring Schema (2)
CRITICALITY	Impacts critical functions/activities	E_4	Would a disruption of the system's function/activity affect essential services?	NO=1	YES=2
SCALE & MATURITY	Breadth of deployment	E_5	Is the AI system deployment a pilot, narrow, broad or widespread?	PILOT=1	LIVE=2
	{ Technical maturity }	E_6	How technically mature is the system (Technology Readiness Level –TRL)	TRL_HIGH=1	TRL_LOW=2
DATA & INPUT	Criteria		Description	Scoring Schema (1)	Scoring Schema (2)
COLLECTION	Detection and collection	D_1	Are the data and input collected by humans, automated sensors or both?	NON_AUTO=1	AUTO=2
	Provenance of data and input	D_2	Are the data and input from experts, provided, observed, synthetic or derived?	OBTAINED=1	PROCESSED=2
	Dynamic nature	D_3	Are the data dynamic, static, dynamic updated from time to time or real-time?	SCHEDULED=1	REAL_TIME=2
RIGHTS & IDENTIFIABILITY	Rights	D_4	Are the data proprietary, public or personal data (related to identifiable individual)?	NON_PII=1	PII=2
	"Identifiability" of personal data	D_5	If personal data, are they anonymised, pseudonymised?	ANON=1	NON_ANON=2
STRUCTURE & FORMAT	{ Structure of data and input }	D_6	Are the data structured, semi-structured, complex structured or unstructured?	ST=1	UNST=2
	{ Format of data and metadata }	D_7	Is the format of the data and metadata standardised or non-standardised?	STD=1	NON_STD=2
SCALE	{ Scale }	D_8	What is the dataset's scale?	LARGE=1	SMALL=2
QUALITY AND APPROPRIATENESS	{ Data quality and appropriateness }	D_9	Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?	FULL=1	SAMPLE=2

AI MODEL	Criteria	Description	Scoring Schema (1)	Scoring Schema (2)
MODEL CHARACTERISTICS	Model information availability	M_1 Is any information available about the system's model?	YES=1	NO=2
	AI model type	M_2 Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?	STAT=1	HU=2
	{Rights associated with model}	M_3 Is the model open-source or proprietary, self or third-party managed?	MANG=1	UNMANG=2
	{Discriminative or generative}	M_4 Is the model generative, discriminative or both?	GEN=1	DIS=1
	{Single or multiple model(s)}	M_5 Is the system composed of one model or several interlinked models?	ONE=1	MANY=2
MODEL-BUILDING		M_6 Is model-building from machine or human knowledge?	HUM=1	ML=2
	Model evolution in the field ML	M_7 Does the model evolve and/or acquire abilities from interacting with data in the field?	NO=1	YES=2
	Central or federated learning ML	M_8 Is the model trained centrally or in a number of local servers or "edge" devices?	LOCAL=1	FED=2
MODEL INFERENCE	{Model development/ maintenance}	M_9 Is the model universal, customisable or tailored to the AI actor's data?	CUST=1	NONCUST=2
	{Deterministic and probabilistic}	M_10 Is the model used in a deterministic or probabilistic manner?	PROB=1	DET=2
	Transparency and explainability	M_11 Is information available to users to allow them to understand model outputs?	YES=1	NO=2
TASK & OUTPUT	Criteria	Description	Scoring Schema (1)	Scoring Schema (2)
TASKS	Task(s) of the system	T_1 What tasks does the system perform (e.g., recognition, event detection, forecasting)?	ANALYSIS=1	OUTCOME=2

(Continued)

(Continued)

PEOPLE & PLANET	Criteria	REF	Description	Scoring Schema (1)	Scoring Schema (2)
ACTION	{Combining tasks and actions into composite systems}	T_2	Does the system combine several tasks and actions (e.g., content generation systems, autonomous systems, control systems)?	SING=1	COMB=2
	Action autonomy	T_3	How autonomous are the system’s actions and what role do humans play?	MONIT=1	AUTO=2
APPLICATION AREA	Core application area(s)	T_4	Does the system belong to a core application area such as human language technologies, computer vision, automation and/or optimisation or robotics?	PERIPH=1	CORE=2
EVALUATION	{Evaluation methods}	T_5	Are standards or methods available for evaluating system output?	YES=1	NO=2

# Appendix 7.2 Score for Facial Recognition in Private Access Control

PEOPLE & PLANET	Criteria	REF	Description	Scoring Schema (1)	Scoring Schema (2)	SUB SCORE	TOTAL WEIGHT	NET SCORE
USERS	Users of AI system	P_1	What is the level of competency of users who interact with the system?	Domain Conversant=1	Not Domain Conversant=2	1		
STAKEHOLDERS	Impacted stakeholders	P_2	Who is impacted by the system (e.g., consumers, workers, government agencies)?	Access restricted to Group=1	Open to Public=2	1		
OPTIONALITY	Optionality and redress	P_3	Can users opt out, e.g., switch systems? Can users challenge or correct the output?	Ability to Interact without AI=1	No option to AI interaction=2	1		
HUMAN RIGHTS	Human rights and democratic values	P_4	Can the system's outputs impact fundamental human rights (e.g., human dignity, privacy, freedom of expression, non-discrimination, fair trial, remedy, safety)?	NO=1	YES=2	1		
WELL-BEING & ENVIRONMENT	Well-being, society and the environment	P_5	Can the system's outputs impact areas of life related to well-being (e.g., job quality, the environment, health, social interactions, civic engagement, education)?	NO=1	YES=2	1		

(Continued)

(Continued)

PEOPLE & PLANET	Criteria	REF	Description	Scoring Schema (1)	Scoring Schema (2)	SUB SCORE	TOTAL WEIGHT	NET SCORE
DISPLACEMENT	{Displacement potential}	P_6	Could the system automate tasks that are or were being executed by humans?	NO=1	YES=2	1		
ECONOMIC CONTEXT	Criteria		Description	Scoring Schema (1)	Scoring Schema (2)	6	0.2	1.2
SECTOR	Industrial sector	E_1	Which industrial sector is the system deployed in (e.g., finance, agriculture)?	NON-CRITICAL=1	CRITICAL=2	1		
BUSINESS FUNCTION & MODEL	Business function	E_2	What business function(s) is the system employed in (e.g., sales, customer service)?	NON-CRITICAL=1	CRITICAL=2	1		
	Business model	E_3	Is the system a for-profit use, non-profit use or public service system?	NON-PUBLIC=1	PUBLIC=2	1		
CRITICALITY	Impacts critical functions/activities	E_4	Would a disruption of the system's function/activity affect essential services?	NO=1	YES=2	1		
SCALE & MATURITY	Breadth of deployment	E_5	Is the AI system deployment a pilot, narrow, broad or widespread?	PILOT=1	LIVE=2	2		
	{Technical maturity}	E_6	How technically mature is the system? (Technology Readiness Level –TRL)	High Technical Readiness=1	Low Technical Readiness=2	1		
DATA & INPUT	Criteria		Description	Scoring Schema (1)	Scoring Schema (2)	7	0.2	1.4
COLLECTION	Detection and collection	D_1	Are the data and input collected by humans, automated sensors or both?	Non Autonomous Collection=1	Autonomous collection=2	1		

RIGHTS & IDENTIFIABILITY	Provenance of data and input	D_2	Are the data and input from experts, provided, observed, synthetic or derived?	Provided Data=1	Synthetic or Derived data=2	1			
	Dynamic nature	D_3	Are the data dynamic, static, dynamic updated from time to time or real-time?	Non Real time data=1	Dynamic real time data=2	1			
	Rights	D_4	Are the data proprietary, public or personal data (related to identifiable individual)?	No Personally Identifiable Information=1	Personally Identifiable Information=2	2			
	“Identifiability” of personal data	D_5	If personal data, are they anonymised, pseudonymised?	Anonymised=1	Non Anonymised=2	2			
	{Structure of data and input}	D_6	Are the data structured, semi-structured, complex structured or unstructured?	Structured or Semi-structured=1	Unstructured=2	2			
STRUCTURE & FORMAT	{Format of data and metadata}	D_7	Is the format of the data and metadata standardised or non-standardised?	Standardised data format=1	Non Standardised data format=2	1			
SCALE	{Scale}	D_8	What is the dataset’s scale?	LARGE=1	SMALL=2	2			
QUALITY AND APPROPRIATENESS	{Data quality and appropriateness}	D_9	Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?	Full data set=1	Sampled set=2	1			
AI MODEL	Criteria		Description	Scoring Schema (1)		Scoring Schema (2)	13	0.2	2.6
MODEL CHARACTERISTICS	Model information availability	M_1	Is any information available about the system’s model?	YES=1	NO=2	1			
	AI model type	M_2	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?	Statistical or Hybrid=1	Human generated rules=2	2			

(Continued)

(Continued)

PEOPLE & PLANET	Criteria	REF	Description	Scoring Schema (1)	Scoring Schema (2)	SUB SCORE	TOTAL	WEIGHT	NET SCORE
	{Rights associated with model}	M_3	Is the model open-source or proprietary, self or third-party managed?	Owned and Managed=1	Third Party/Open Source=2	1			
	{Discriminative or generative}	M_4	Is the model generative, discriminative or both?	Generative=1	Discriminative=1	2			
	{Single or multiple model(s)}	M_5	Is the system composed of one model or several interlinked models?	Single model=1	Multi model=2	1			
MODEL-BUILDING		M_6	Is model-building from machine or human knowledge?	Human Knowledge=1	Machine Learning=2	2			
	Model evolution in the field ML	M_7	Does the model evolve and/or acquire abilities from interacting with data in the field?	NO=1	YES=2	1			
	Central or federated learning ML	M_8	Is the model trained centrally or in a number of local servers or “edge” devices?	Centralised Training=1	Federated training=2	1			
MODEL INFERENCE	{Model development/ maintenance}	M_9	Is the model universal, customisable or tailored to the AI actor’s data?	Customisable=1	Non Customisable=2	1			
	{Deterministic and probabilistic}	M_10	Is the model used in a deterministic or probabilistic manner?	Probabilistic=1	Deterministic=2	2			
	Transparency and explainability	M_11	Is information available to users to allow them to understand model outputs?	YES=1	NO=2	1			
TASK & OUTPUT TASKS	Criteria		Description	Scoring Schema (1)	Scoring Schema (2)		15	0.2	3
	Task(s) of the system	T_1	What tasks does the system perform (e.g., recognition, event detection, forecasting)?	Non-predictive=1	Predictive=2	2			

	{Combining tasks and actions into composite systems}	T_2	Does the system combine several tasks and actions (e.g., content generation systems, autonomous systems, control systems)?	Singular=1	Combinatorial=2	1			
ACTION	Action autonomy	T_3	How autonomous are the system's actions and what role do humans play?	Supervised=1	Autonomous=2	1			
APPLICATION AREA	Core application area(s)	T_4	Does the system belong to a core application area such as human language technologies, computer vision, automation and/or optimisation or robotics?	Non Core Application=1	Core Application=2	1			
EVALUATION	{Evaluation methods}	T_5	Are standards or methods available for evaluating system output?	YES=1	NO=2	1			
							6	0.2	1.2
							TOTAL		9.4

---



# Appendix 7.3 Score for Facial Recognition in Law Enforcement

PEOPLE & PLANET	Criteria	P	Description	Scoring Schema (1)	Scoring Schema (2)	SUB SCORE	TOTAL WEIGHT	NET SCORE
USERS	Users of AI system	P_1	What is the level of competency of users who interact with the system?	Domain Conversant=1	Not Domain Conversant=2	1		
STAKEHOLDERS	Impacted stakeholders	P_2	Who is impacted by the system (e.g., consumers, workers, government agencies)?	Access restricted to Group=1	Open to Public=2	2		
OPTIONALITY	Optionality and redress	P_3	Can users opt out, e.g., switch systems? Can users challenge or correct the output?	Ability to Interact without AI=1	No option to AI interaction=2	1		
HUMAN RIGHTS	Human rights and democratic values	P_4	Can the system's outputs impact fundamental human rights (e.g., human dignity, privacy, freedom of expression, non-discrimination, fair trial, remedy, safety)?	NO=1	YES=2	2		
WELL-BEING & ENVIRONMENT	Well-being, society and the environment	P_5	Can the system's outputs impact areas of life related to well-being (e.g., job quality, the environment, health, social interactions, civic engagement, education)?	NO=1	YES=2	2		
DISPLACEMENT	{Displacement potential}	P_6	Could the system automate tasks that are or were being executed by humans?	NO=1	YES=2	2		

ECONOMIC CONTEXT	Criteria	E	Description	Scoring Schema (1)	Scoring Schema (2)	11	0.3	3.3
SECTOR	Industrial sector	E_1	Which industrial sector is the system deployed in (e.g., finance, agriculture)?	NON-CRITICAL=1	CRITICAL=2	2		
BUSINESS FUNCTION & MODEL	Business function	E_2	What business function(s) is the system employed in (e.g., sales, customer service)?	NON-CRITICAL=1	CRITICAL=2	2		
	Business model	E_3	Is the system a for-profit use, non-profit use or public service system?	NON-PUBLIC=1	PUBLIC=2	2		
CRITICALITY	Impacts critical functions/activities	E_4	Would a disruption of the system's function/activity affect essential services?	NO=1	YES=2	1		
SCALE & MATURITY	Breadth of deployment	E_5	Is the AI system deployment a pilot, narrow, broad or widespread?	PILOT=1	LIVE=2	2		
	{Technical maturity}	E_6	How technically mature is the system? (Technology Readiness Level –TRL)	High Technical Readiness=1	Low Technical Readiness=2	1		
DATA & INPUT	Criteria	D	Description	Scoring Schema (1)	Scoring Schema (2)	10	0.1	1
COLLECTION	Detection and collection	D_1	Are the data and input collected by humans, automated sensors or both?	Non Autonomous Collection=1	Autonomous collection=2	2		
	Provenance of data and input	D_2	Are the data and input from experts, provided, observed, synthetic or derived?	Provided Data=1	Synthetic or Derived data=2	1		

(Continued)

*(Continued)*

PEOPLE & PLANET	Criteria	P	Description	Scoring Schema (1)	Scoring Schema (2)	SUB SCORE	TOTAL WEIGHT	NET SCORE
	Dynamic nature	D_3	Are the data dynamic, static, dynamic updated from time to time or real-time?	Non Real time data=1	Dynamic real time data=2	2		
RIGHTS & IDENTIFIABILITY	Rights	D_4	Are the data proprietary, public or personal data (related to identifiable individual)?	No Personally Identifiable Information=1	Personally Identifiable Information=2	2		
	“Identifiability” of personal data	D_5	If personal data, are they anonymised, pseudonymised?	Anonymised=1	Non Anonymised=2	2		
STRUCTURE & FORMAT	{Structure of data and input}	D_6	Are the data structured, semi-structured, complex structured or unstructured?	Structured or Semi-structured=1	Unstructured=2	2		
	{Format of data and metadata}	D_7	Is the format of the data and metadata standardised or non-standardised?	Standardised data format=1	Non Standardised data format=2	1		
SCALE	{Scale}	D_8	What is the dataset’s scale?	LARGE=1	SMALL=2	1		
QUALITY AND APPROPRIATENESS	{Data quality and appropriateness}	D_9	Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?	Full data set=1	Sampled set=2	2		
AI MODEL	Criteria	M	Description	Scoring Schema (1)	Scoring Schema (2)	15	0.2	3
MODEL CHARACTERISTICS	Model information availability	M_1	Is any information available about the system’s model?	YES=1	NO=2	1		
	AI model type	M_2	Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?	Statistical or Hybrid=1	Human generated rules=2	1		

MODEL-BUILD- ING	{Rights associated with model}	M_3	Is the model open-source or proprietary, self or third-party managed?	Owned and Managed=1	Third Party/Open Source=2	1			
	{Discriminative or generative}	M_4	Is the model generative, discriminative or both?	Generative=1	Discriminative=1	2			
	{Single or multiple model(s)}	M_5	Is the system composed of one model or several interlinked models?	Single model=1	Multi model=2	1			
		M_6	Is model-building from machine or human knowledge?	Human Knowledge=1	Machine Learning=2	2			
	Model evolution in the field ML	M_7	Does the model evolve and/or acquire abilities from interacting with data in the field?	NO=1	YES=2	2			
	Central or federated learning ML	M_8	Is the model trained centrally or in a number of local servers or “edge” devices?	Centralised Training=1	Federated training=2	2			
	{Model development/ maintenance}	M_9	Is the model universal, customisable or tailored to the AI actor’s data?	Customisable=1	Non Customisable=2	2			
	{Deterministic and probabilistic}	M_10	Is the model used in a deterministic or probabilistic manner?	Probabilistic=1	Deterministic=2	1			
MODEL INFERENCE	Transparency and explainability	M_11	Is information available to users to allow them to understand model outputs?	YES=1	NO=2	2			
<b>TASK &amp; OUTPUT</b>	<b>Criteria</b>	<b>T</b>	<b>Description</b>	<b>Scoring Schema (1)</b>	<b>Scoring Schema (2)</b>	<b>17</b>	<b>0.2</b>	<b>3.4</b>	
TASKS	Task(s) of the system	T_1	What tasks does the system perform (e.g., recognition, event detection, forecasting)?	Non-predictive=1	Predictive=2	1			

(Continued)

(Continued)

PEOPLE & PLANET	Criteria	P	Description	Scoring Schema (1)	Scoring Schema (2)	SCORE	SUB TOTAL	WEIGHT	NET SCORE
	{Combining tasks and actions into composite systems}	T_2	Does the system combine several tasks and actions (e.g., content generation systems, autonomous systems, control systems)?	Singular=1	Combinatorial=2	1			
ACTION	Action autonomy	T_3	How autonomous are the system's actions and what role do humans play?	Supervised=1	Autonomous=2	1			
APPLICATION AREA	Core application area(s)	T_4	Does the system belong to a core application area such as human language technologies, computer vision, automation and/or optimisation or robotics?	Non Core Application=1	Core Application=2	2			
EVALUATION	{Evaluation methods}	T_5	Are standards or methods available for evaluating system output?	YES=1	NO=2	1			
							6	0.2	1.2
							TOTAL		11.9

---

# 8 Harnessing the Potential of AI for Indian Agriculture

## *Using “Bhashini” as a Tool to Deploy Responsible AI and Increase the Uptake of AI Applications Among Farmers*

*Abhishek Raj, Harsh Singh, and Anshul Pachouri*

### 8.1 INTRODUCTION

Agriculture in India and worldwide faces immense pressure due to the ever-increasing population to feed. Increasing productivity or expanding cultivated land is essential to absorb this pressure. As the latter is difficult to attain, achieving gains in productivity is a more practical solution. An increase in productivity also translates to better income levels in the agriculture sector. This is critical in the Indian context, as the agriculture sector employs more than 45% of the nation's workforce besides contributing to the food security (National Sample Survey Office, 2022). However, modern agricultural practices in India have been input-intensive and are susceptible to shocks that result from weather uncertainty and market variability (Sharma et al., 2021). This threatens 86% of India's farmers who fall under the category of small and marginal farmers (SMF). Farmers in this category have less than two hectares of land and depend on their farm produce for livelihood (Agriculture Census Division, 2019).

Indian farmers rely on agriculture extension and advisory services (AEAS) for farming-related advisories besides their traditional knowledge systems to deal with uncertainties and increase productivity (Danso-Abbeam et al., 2018). Agriculture extension services provide farmers with access to knowledge and information needed to increase productivity (National Resources Institute, 2014; Kansime et al., 2019). In India, AEAS are mostly delivered through extension workers, call centers, interactive voice response systems (IVRS), and short message service (SMS), among others (Rajeev & Srinivas, 2023). However, these mechanisms have certain limitations, such as a shortage of extension staff, unavailability of call centers at farmers'

convenience, and call-center agents' variable competence in the delivery of technical advice, among others. These advisories are often generic and lack personalization based on farmers' exact requirements and circumstances. Further, most of these services are push-based, which means they are not available on demand (Rajeev & Srinivas, 2023).

Technological advancements can potentially address many existing limitations and improve the existing extension service delivery mechanism for farmers in India, especially with the growth of artificial intelligence (AI). In particular, the generative aspect of AI-powered chatbots can potentially solve the issue of personalized and targeted advisory and query resolution in agriculture extension services. AI-powered chatbots are sophisticated software that can carry out human-like conversations through responses based on the data that it has been trained on. They often use natural language processing (NLP), machine learning (ML), and large language models (LLMs) (Google, n.d.). These chatbots could serve multiple uses in agriculture – they can provide advice on weather, agronomic practices, crop management, pest management, and resolve other farming-related queries. These chatbots could also exponentially reduce the time taken to solve farmer queries compared to manual resolution mechanisms, which suffer from huge gaps in demand for advice. The manual mechanisms have shortfalls related to experts' availability to answer the queries. AI chatbots, such as *Jugalbandi*, launched by Microsoft in May 2023, and *Ama KrushAI*, launched by the Government of Odisha in February 2023, among others, have been envisaged to democratize access to agriculture expertise, especially for the small shareholding farmers.

In India, the government has been pushing to use AI's power for various purposes, such as grievance management, query resolution, and analytics. This space has seen a lot of development and attracted much traction from the administration to deploy these solutions and increase productivity. However, for a diverse country, such as India, where linguistic variations emerge every few hundred kilometers, AI-powered bots' penetration and uptake depend upon the ability to provide an output suited to the farmers' comprehension. This is where the "*Bhashini*" platform emerges as a vital tool to break the linguistic barrier to realize inclusivity. Launched in July 2022 as an AI-based local language initiative, *Bhashini* seeks to make content available in Indian languages digitally and help develop services for the nation's citizens (Press Information Bureau, 2022). The integration of *Bhashini*'s language application programming interface (API) has made personalized agricultural advisory services through AI-powered chatbots a reality for Indian farmers.

In this chapter, we discuss examples of a few emerging chatbots in India's agricultural realm that have used *Bhashini*'s translational capabilities. These include *Ama KrushAI*, *Jugalbandi*, *Kisan e-Mitra*, and *KissanAI*. We discuss barriers to the uptake of AI applications among farmers for agriculture, particularly chatbots. Our discussion is based on extensive secondary research and interviews with stakeholders. These stakeholders include farmers, agricultural experts, solution architects, and government officials. Finally, we provide recommendations to increase the uptake of AI applications for the betterment of agriculture in India through increased productivity and farm incomes.

We envisage that the policy and other recommendations will help address the emerging challenges in the adoption of AI chatbots by Indian farmers. The chapter may also help stakeholders identify *Bhashini's* other innovative use cases in Indian agriculture that could be explored in the future.

## 8.2 USE OF AI CHATBOTS IN INDIAN AGRICULTURE: A FEW EXAMPLES

This section discusses a few examples of the AI chatbots deployed in India to assist farmers with their agricultural and related needs.

### 8.2.1 AMA KRUSHAI

The Department of Agriculture and Farmer's Empowerment, Government of Odisha, launched *Ama KrushAI* in February 2023. It is portrayed as the first AI-powered chatbot in India dedicated to the agriculture sector (The New Indian Express, 2023). The chatbot seeks to help Odisha's farmers with advisory services on the best agro-nomic practices, government programs, and loan products from more than 40 commercial and cooperative banks in Odia, Hindi, and English (Das, 2023).

*Ama KrushAI* is expected to provide a tailored response to farmers' specific queries and complements Odisha's other initiative – *Ama Krushi*. This initiative provides customized agricultural advice free of cost to Odisha's farmers through weekly calls to their number. It has approximately 690,000 farmer enrollments (Department of Agriculture and Farmers' Empowerment, Government of Odisha, n.d.).

When it comes to technology, the *Ama KrushAI* system architecture incorporates a dual-response mechanism. Initially, it examines an existing database for content relevant to the user's query. Once it identifies relevant material, it refines it through a generative pre-trained transformer (GPT) technology to generate a user-focused response. In the absence of specific database content, GPT-3 addresses the query directly through its extensive pre-training to formulate a suitable response. This approach ensures tailored responses for specialized queries and general answers for new or undefined inquiries (Rajeev & Srinivas, 2023). The system integrates the unified communications interface (UCI) to optimize user interaction, with a particular emphasis on agricultural professionals. UCI is a digital public good (DPG) framework that has proved effective in governance contexts (Sunbird, n.d.).

The AI's effectiveness depends on the incorporation of localized domain knowledge, which highlights the need to create a democratic knowledge system with a focus on accurate information sourcing (Kulkarni, 2023). In this regard, *Ama KrushAI* uses data from a knowledge database called *Krushak*, Odisha state farmers' database to provide contextualized and personalized extension services to farmers (Times of India, 2023). The database was developed over five years with content from the Department of Agriculture and the Odisha University of Agricultural Technology. The platform incorporates *Bhashini*, as it recognizes the importance of regional language communication. *Bhashini's* capabilities in translation, transliteration, speech-to-text, and text-to-speech, especially in Odia, facilitate natural and efficient interaction between farmers and the AI system (Rajeev & Srinivas, 2023).



In terms of implementation, the Ama KrushAI chatbot is still in the nascent stage, and the first pilot has been running with 10,000 farmers (The New Indian Express, 2023). Notably, modern knowledge management moves beyond traditional static documents and frequently asked questions (FAQs) in *Ama KrushAI*'s context. It evolves into systems with advanced features, such as multilingual and conversational interfaces, and capabilities for logical reasoning and understanding conditionality, causality, and correlation (Kulkarni, 2023).

### 8.2.2 Jugalbandi

*Jugalbandi*, developed under the collaboration of Microsoft Research and *AI4Bharat*, is an AI-powered chatbot positioned as an open-source platform. *Jugalbandi* uses LLMs, such as GPT and Indian language translation models. The Indian language translation models also include those under the Indian government's *Bhashini*'s mission to power conversational AI solutions that can respond in real-time to human queries in the desired language (Jugalbandi team, n.d.).

As per its creators, the foundational vision for the *Jugalbandi* bot's creation was to address India's linguistic divide. Despite English being the predominant language for business and public affairs, only 11% of the Indian population speaks English. This contrasts starkly with Hindi, which is spoken by 57% of the population (Office of the Registrar General, 2018). As a result, a substantial segment of the population is inadvertently excluded from accessing vital government programs, predominantly due to language barriers. The *Jugalbandi* bot was conceptualized to bridge this gap (Yee, 2023).

In the present operational mechanism of the *Jugalbandi* bot, the process starts by sending a text- or audio-based message to a designated WhatsApp number. This message activates the chatbot's functionalities. The initial step involves the message's conversion into text through the *AI4Bharat* speech recognition model. Subsequently, the text undergoes translation into English, a task executed by the *Bhashini* translation model. A pivotal aspect of *Jugalbandi*'s technology is its foundation on OpenAI's ChatGPT. In a prominent agriculture use case, the *Jugalbandi* chatbot retrieves information pertinent to government programs for farmers through the user's prompts. This information is then translated back into Hindi or other local languages.

The final step involves the synthesis of this information into an audio format through *AI4Bharat*'s text-to-speech model. The processed response is subsequently delivered back to the user via WhatsApp. It effectively communicates with users in remote areas, such as villagers, in their native language (Anand, 2023).

In terms of implementation, the *Jugalbandi* chatbot is still in the nascent stage, albeit with a lot of potential. As per the available statistics, it covers 10 of India's 22 official languages and 171 of approximately 20,000 government programs (Yee, 2023).

### 8.2.3 Kisan e-Mitra

The Union Ministry of Agriculture and Farmers Welfare launched the *Kisan e-Mitra* AI chatbot in September 2023 for effective grievance management under the *Pradhan Mantri Kisan Samman Nidhi* (PM-KISAN) program (Press Information Bureau, 2023). The *Kisan e-Mitra* chatbot intends to enhance PM-KISAN's efficiency

and reach through increased access to program information for farmers and resolution of their grievances (AIR Staff, 2023).

The AI chatbot has been built with the collaborative efforts of *EkStep* Foundation and *Bhashini*, among others. The chatbot integrates *Bhashini*'s language models, which makes it accessible to farmers and beneficiaries of different linguistic regions. The AI chatbot can understand users' queries about the program and respond in their desired language. The *Kisan e-Mitra* chatbot also helps the beneficiaries address their queries about application status, payment status, and other grievances (ET Government, 2023).

Currently, the *Kisan e-Mitra* chatbot is available as a web application and is also integrated with the PM-KISAN mobile application (AIR Staff, 2023). The chatbot is available in English, Hindi, Bengali, Odia, and Tamil, with plans to make it available in the 22 scheduled languages in the Constitution of India (AIR Staff, 2023).

### 8.2.4 *KissanAI*

*KissanAI*, previously known as *KissanGPT*, is an AI-powered chatbot launched in September 2023. The chatbot seeks to help Indian farmers increase their productivity and profitability through real-time advice on irrigation, crop and pest management, and other farming-related queries (Stanly, 2023).

*KissanAI* uses ChatGPT 3.5, Whisper models, and on-field data collected through various agricultural research universities to provide solutions in a localized context (Ground Report, 2023). It offers a user-friendly interface accessible through its web portal and mobile application. It currently supports nine Indian languages: Gujarati, Marathi, Tamil, Telugu, Kannada, Malayalam, Punjabi, Bangla, and Hindi. It plans to add Assamese and Odia to its list of supported languages (Pawar, 2023).

## 8.3 BARRIERS TO THE UPTAKE OF AI-POWERED APPLICATIONS, SUCH AS CHATBOTS

Based on stakeholder interviews, secondary research, and the examples discussed in section 8.2, we have identified specific barriers that hinder the uptake of AI applications, such as chatbots, among Indian farmers.

### A. Linguistic diversity and training challenges

India is a linguistically diverse nation. As per the 2011 Census, India is home to 121 languages, with 1,369 rationalized mother tongues, which are classified, and 1,474 unclassified mother tongues (Singh & Nakkeerar, 2022). This linguistic diversity poses huge challenges for language data collection to train the models. High-level insights from the field suggest that existing AI chatbots struggle to interpret dialects and regional languages.

### B. Low literacy levels necessitate speech-based interfaces

Conventional text-based interfaces are less effective in regions with low literacy. Moreover, speech offers more convenience than written text to farmers. Thus, speech-based solutions have become crucial in the Indian context. However, data collection for a language with fewer

speakers, such as Manipuri and Bodo, becomes challenging. Large-scale translation solutions, such as Google Translate, have become less effective for these languages. Further, a significant resource gap plagues low-resource languages, exacerbated by challenges, such as the inability to find data collectors, difficult geographies for data collection, and verification issues.

### C. Contextual relevance and trust in chatbots

Gaining and retaining farmers' trust in chatbots is a massive barrier to the uptake of AI applications. In many cases, generative AI struggles to understand the context of the query. Incorrect and out-of-context responses may make farmers hesitant to use the chatbot in the future. Moreover, India has 127 agro-climatic zones, each with unique topography, soil types, and weather conditions (Verma et al., 2017), which makes centralized advisories ineffective. In this regard, extensive efforts are required to create region-specific datasets to increase contextual understanding of AI models. This again points toward the need to collect reliable data with a local context suited to the specific agro-climatic zone.

Additionally, digital advisories and the grievance resolution mechanism need accountability. Farmers will not trust advice they do not understand and cannot provide their feedback to. Interestingly, some stakeholders that we interviewed for this chapter highlighted farmers' preference for interactive voice response (IVR) over chatbots, possibly due to trust factors.

### D. Digital connectivity and behavioral gaps

Despite India's success story in digital connectivity, many areas, especially the country's remote, far-flung regions, still lack access to connectivity. For instance, only 41% of India's rural population actively uses the Internet (Kantar, 2023). Ownership of smartphone devices is also on the lower side among farmers. These issues, in turn, hamper reliance on digital services, whereas continued service delivery is important to build robust channels of agriculture advisory, which the farming community can trust.

Further, a lack of digital skills and an understanding of the benefits of services, such as AI chatbots, hinder the uptake of AI applications among farmers. Notably, initiatives, such as *GramVaani*, have tried to bridge this gap through on-ground volunteers to help the community.

### E. Resource availability constraints

As highlighted in previous sections, 86% of Indian farmers are small or marginal. These farmers may receive timely advice but often lack the resources, such as agricultural equipment, to act on it. This lack of resources renders the advice ineffective. Indian farmers struggle a lot with this issue. For instance, in the case of tube wells, many small farmers might miss out on irrigation schedules due to availability issues with the tube wells, even when they get a rainfall advisory.

Moreover, these small farmers operate on tight margins. They may lack individual buying capacity for AI applications, especially private

sector applications, which may have associated subscription or usage costs. Thus, affordability and cost-effectiveness become crucial for AI-powered applications, such as chatbots.

## 8.4 RECOMMENDATIONS

This section provides recommendations for the uptake of AI applications, particularly chatbots, among Indian farmers.

### 8.4.1 CONTEXTUAL TRAINING

The quality of AI chatbots' outputs depends on the quality of the data with which the model has been trained. The AI models should be fed with contextual data that is fit for a particular region. It must be trained with regional data to develop effective responses, maintain context, and solve local-level queries. Farmers should get customized advice as per their needs and circumstances. For example, the time of sowing crops differs for every farmer. Therefore, advice on irrigation, pest management, and harvesting customized to their sowing time and field conditions will be useful.

The development of an agricultural glossary will aid this contextualization. For example, the same crop can have different names based on regions. However, with the glossary, the AI model and chatbot will better understand the context and address the query. Moreover, the initiation of dedicated wings under regional agricultural universities for data collection, updates, and verification can provide more localized and effective advisories. This data collection should gradually be extended to capture village-level data, which is mostly up to the district level at present.

### 8.4.2 HUMAN-IN-THE-LOOP

The AI field has advanced a lot, but it is still not mature enough to manage the advisory without human moderation. Human intervention is crucial in all phases of the design and deployment of such a solution. For example, on-ground human intervention is required to upskill and teach farmers from all literacy levels how to use the systems to increase productivity.

### 8.4.3 DELIVERY OF AN AGRICULTURAL CHATBOT "PLUS" PACKAGE

A solution in isolation will not help increase the uptake and win farmers' trust. Along with the basic purpose of agro-advisory, it should have add-on features, such as weather updates, pest management, soil management, and *mandi* price information, among others.

### 8.4.4 COLLABORATION AND PARTNERSHIPS

The success of initiatives, such as AI chatbots, will require collaborative efforts from all stakeholders. For instance, verification of collected data and model outputs is often outsourced to third-party agencies, which necessitates collaborations with third-party agencies.

Further, partnerships with local non-governmental organizations (NGOs) that work in agriculture can help train and evolve AI models to adapt them better to local needs. The farmers in India receive advice based on the farming-related database, which state agriculture universities traditionally maintain. These universities may not always be able to build AI applications independently. However, the private sector can use these databases to train their models efficiently. For this to happen, a mechanism should be in place that allows responsible data sharing between state agricultural universities and the private sector. A good example is the collaboration among Andhra Pradesh's Acharya N G Ranga Agricultural University, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), and Plantix, an AI application for pest management (ICRISAT, 2017).

#### **8.4.5 REGULAR ASSESSMENT OF FARMERS' NEEDS AND SERVICE OFFERINGS**

Conducting assessments in different regions can help in understanding and bridging the gap between farmers' needs and the services offered by AI-powered chatbots. The assessment can be accomplished from various methods, such as on-ground surveys and findings from the pilot, among others.

#### **8.4.6 DATA PROTECTION AND PRIVACY**

With the growth of AI, a significant focus will emerge on data ownership. The indiscriminate use of farmers' data can lead to catastrophic outcomes. Big players in the market, such as a large company from the chemical industry, can use the data to target specific segments of farmers and earn more profitability.

### **8.5 CONCLUSION**

AI-powered chatbots can usher benefits for Indian agriculture through increased productivity, profitability, and resilience if the existing challenges are addressed responsibly. These chatbots present an exciting opportunity to improve the delivery of agriculture extension services to farmers in an effective manner. Initiatives, such as *Bhashini*, have enabled the development of multilingual interfaces in the AI chatbots for agriculture, such as *Ama KrushAI* and *Jugalbandi*, among others.

However, more contextual training of AI models, human integration in the loop, expectations management, and farmers' privacy protection are vital for the responsible uptake of such applications. The gap between farmers' needs and chatbots' services should be bridged. Collaborations and partnerships with NGOs in agriculture, agricultural universities, and the private sector will also catalyze the uptake.

### **8.6 ACKNOWLEDGMENTS**

We are thankful to all the stakeholders who agreed to provide their insights. We extend our heartfelt gratitude to Upasna Sharma, Shubhmoy Kumar Garg, Sultan Ahmad, Mohammad Salman, Sakshi Joshi, and Navin Bhushan. We are grateful to our colleagues Mitul Thapliyal, Kunjbihari Daga, Vikram Pratap Sharma, and Diganta Nayak at MSC for their support in writing this policy brief.

## REFERENCES

- Agriculture Census Division. (2019). *Agriculture census 2015–16*. New Delhi: Agriculture Census Division, Department of Agriculture, Co-Operation & Farmers Welfare. [https://agcensus.nic.in/document/agcen1516/T1\\_ac\\_2015\\_16.pdf](https://agcensus.nic.in/document/agcen1516/T1_ac_2015_16.pdf)
- AIR Staff. (2023, September 21). Union minister Kailash Choudhary launches PM KISAN AI-chatbot (Kisan e-Mitra) in New Delhi. *All India Radio News*. Retrieved December 1, 2023, from [https://newsonair.gov.in/News?title=Union-Minister-Kailash-Choudhary-launches-PM-KISAN-AI-Chatbot-\(Kisan-e-Mitra\)-in-New-Delhi&id=468123](https://newsonair.gov.in/News?title=Union-Minister-Kailash-Choudhary-launches-PM-KISAN-AI-Chatbot-(Kisan-e-Mitra)-in-New-Delhi&id=468123)
- Anand, N. (2023, May 26). Microsoft-powered AI chatbot ‘Jugalbandi’ is here. All you need to know. *Hindustan Times*. Retrieved November 30, 2023, from <https://www.hindustantimes.com/technology/jugalbandi-microsoft-ai-chatbot-features-open-source-language-technology-news-ai4bharat-project-101685111840575.html>
- Danso-Abbeam, G., Ehiakor, D. S., & Aidoo, R. (2018). Agricultural extension and its effects on farm productivity and income: Insight from Northern Ghana. *Agriculture & Food Security*. <https://doi.org/10.1186/s40066-018-0225-x>
- Das, S. (2023, February 20). *Odisha launches India’s first AI-chatbot Ama KrushAI for farmers*. Krishi Jagran. Retrieved November 28, 2023, from <https://krishijagran.com/news/odisha-launches-ai-chatbot-to-help-farmers-with-personalized-extension-services/>
- Department of Agriculture and Farmers’ Empowerment, Government of Odisha. (n.d.). *About Ama Krushi*. Ama Krushi: Krushaka ra Sathi. Retrieved November 29, 2023, from <https://www.amakrushi.in/about/>
- ETGovernment. (2023, September 22). *AI chatbot for PM-KISAN scheme launched; designed to provide seamless support to farmers*. ET Government. Retrieved November 27, 2023, from <https://government.economictimes.indiatimes.com/news/digital-india/ai-chatbot-for-pm-kisan-scheme-launched-designed-to-provide-seamless-support-to-farmers/103850257>
- Google. (n.d.). *AI chatbots*. Google Cloud. Retrieved December 1, 2023, from <https://cloud.google.com/use-cases/ai-chatbot>
- Ground Report. (2023, April 13). *What is KissanGPT set to help Indian farmers?* Ground Report. Retrieved December 1, 2023, from <https://groundreport.in/What-is-KissanGPT-set-to-help-Indian-farmers/>
- ICRISAT. (2017, May 25). *Mobile app to help farmers overcome crop damage launched in India*. International Crops Research Institute for the Semi-Arid Tropics (ICRISAT). Retrieved December 1, 2023, from <https://www.icrisat.org/mobile-app-to-help-farmers-overcome-crop-damage-launched-in-india/#:~:text=The%20'Plantix'%20app%20was%20launched,immediately%20to%20the%20individual%20farmer>
- Jugalbandi Team. (n.d.). *About Jugalbandi*. Jugalbandi. Retrieved November 27, 2023, from <https://www.jugalbandi.ai/mission>
- Kansiime, M. K., Alawy, A., Allen, C., Subharwal, M., Jadhav, A., & Parr, M. (2019). Effectiveness of mobile agri-advisory service extension model: Evidence from Direct-2Farm program in India. *World Development Perspectives*. <https://doi.org/10.1016/j.wdp.2019.02.007>
- Kantar. (2023). *Internet in India 2022*. Kantar. [https://www.iamai.in/sites/default/files/research/Internet%20in%20India%202022\\_Print%20version.pdf](https://www.iamai.in/sites/default/files/research/Internet%20in%20India%202022_Print%20version.pdf)
- Kulkarni, R. (2023, May 18). Open source knowledge platforms: The key to accurate AI bots. *Express Computer*. Retrieved November 29, 2023, from <https://www.expresscomputer.in/artificial-intelligence-ai/open-source-knowledge-platforms-the-key-to-accurate-ai-bots/98129/>
- National Resources Institute. (2014). *Agricultural extension, advisory services and innovation*. University of Greenwich. <https://www.nri.org/publications/thematic-papers/7-agricultural-extension-advisory-services-and-innovation/vile>
- National Sample Survey Office. (2022). *Periodic labour force survey (PLFS) July 2021–June 2022*. Ministry of Statistics and Programme Implementation, Government of

- India. [https://www.mospi.gov.in/sites/default/files/publication\\_reports/AnnualReport-PLFS2021-22F1.pdf](https://www.mospi.gov.in/sites/default/files/publication_reports/AnnualReport-PLFS2021-22F1.pdf)
- The New Indian Express. (2023, February 19). India's first agri chatbot Ama KrushAI launched in Odisha. *The New Indian Express*. Retrieved November 30, 2023, from <https://www.newindianexpress.com/states/odisha/2023/feb/19/indias-first-agri-chatbot-ama-krushai-launched-in-odisha-2548843.html>
- Office of the Registrar General. (2018). *Census of India 2011: Language*. New Delhi: Ministry of Home Affairs, Government of India. [https://language.census.gov.in/eLanguageDivision\\_VirtualPath/eArchive/pdf/C-16\\_2011.pdf](https://language.census.gov.in/eLanguageDivision_VirtualPath/eArchive/pdf/C-16_2011.pdf)
- Pawar, S. (2023, April 14). *Meet KissanGPT, an AI voice assistant designed for Indian farmers*. Analytics Drift. Retrieved December 1, 2023, from <https://analyticsdrift.com/meet-kissangpt-an-ai-voice-assistant-designed-for-indian-farmers/>
- Press Information Bureau. (2022, August 26). *BHASHINI – national language translation mission*. PIB. Retrieved December 1, 2023, from <https://static.pib.gov.in/WriteReadData/specificdocs/documents/2022/aug/doc202282696201.pdf>
- Press Information Bureau. (2023, September 21). *Union minister of state for agriculture and farmers welfare, Shri Kailash Choudhary launches AI chatbot for PM-KISAN scheme today*. Press Information Bureau. <https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1959461>
- Rajeev, G., & Srinivas, V. (2023, March 31). *Leveraging Artificial Intelligence to deliver advisory to farmers*. Samagra. <https://www.samagragovernance.in/blog/2023-03-31-leveraging-artificial-intelligence-to-deliver-advisory-to-farmers/>
- Sharma, U., Chetri, P., Minocha, S., Roy, A., Holker, T., Patt, A., & Joerin, J. (2021). Do phone-based short message services improve the uptake of agri-met advice by farmers? A case study in Haryana, India. *Climate Risk Management*. <https://doi.org/10.1016/j.crm.2021.100321>
- Singh, D. K., & Nakkeerar, D. (2022). *Census of India 2011 – language Atlas – INDIA*. New Delhi: Office of the Registrar General & Census Commissioner, India (ORGI) – Map Division.
- Stanly, M. (2023, April 25). *Conversations – how KissanGPT helps Indian farmers earn profit*. IndiaAI. Retrieved December 1, 2023, from <https://indiaai.gov.in/article/how-kissangpt-helps-indian-farmers-earn-profit>
- Sunbird. (n.d.). *UCI use cases*. Sunbird UCI. Retrieved November 30, 2023, from <https://uci.sunbird.org/learn/uci-use-cases>
- Times of India. (2023, February 18). Odisha launches Ama KrushiAI bot for farmers. *The Times of India*. Retrieved November 30, 2023, from <https://timesofindia.indiatimes.com/india/odisha-launches-ama-krushiai-bot-for-farmers/articleshow/98048844.cms?from=mdr>
- Verma, S., Gulati, A., & Hussain, S. (2017). *Doubling agricultural growth in Uttar Pradesh: Sources and drivers of agricultural growth and policy lessons*. Indian Council for Research on International Economic Relations. [https://icrier.org/pdf/Working\\_Paper\\_335.pdf](https://icrier.org/pdf/Working_Paper_335.pdf)
- Yee, C. M. (2023, May 23). *With help from next-generation AI, Indian villagers gain easier access to government services*. Microsoft. Retrieved December 1, 2023, from <https://news.microsoft.com/en-in/features/with-help-from-next-generation-ai-indian-villagers-gain-easier-access-to-government-services/>

---

# 9 Regional Inequities in Extraction and Flow of Resources That Support and Power the Design, Development and Access to AI

## *Lessons from the Global South*

*Saikat Datta, Shachi Solanki,  
and Anand Venkatanaryanan*

### 9.1 INTRODUCTION

The advent of artificial intelligence (AI) poses the risk of perpetuating the inequities between the Global North and South, reminiscent of industrialisation and colonisation in the 18th and 19th centuries (Mohamed et al., 2020).

This chapter draws on Indian and Kenyan experiences around AI to explore how the Global South can avoid the inequities of the past. It will also have to take measures to ensure the benefits of AI remain in the Global South.

The problems of how AI is currently evolving and the inequities it is creating are manifold. Studies (Beaudry et al., 2006; Vasuki, 2013) at the turn of the 21st century reveal that the concentration in the development of technology has a profound impact on the development of cities in the US. They show a direct correlation on increase in wages, skilled labour and other benefits. The global supply chain for AI, from extracting raw datasets to labelling, modelling and deployment show rewards vastly skewed in favour of the Global North.

Studies (Chan et al., 2021) show that data labelling is done in countries like India and sub-Saharan countries and then shipped to the Global North to develop models and deployed, earning huge profits. This also leads to other problems such as technology denial (Government of India & Government of the United States of America,



n.d.) or creation of regimes that favour the Global North. This too is patterned on discriminatory international regimes such as the Nuclear Proliferation Treaty (NPT), Missile Technology Control Regime (MTCR) and the Wassenaar Arrangement.

Countries like India and Kenya have also faced high-technology denial in areas such as the super-computers (Report of the Steering Committee on Science and Technology for Eleventh Five Year Plan (2007–12), 2006) and cryogenic engines (Sinha, 2017) when it did not suit the Global North. The skewed evolution of AI is also showing that while the extracted data leverages rich diversity of the Global South, the modelling of these datasets end up being biased against the very regions from which they are sourced (Sambasivan et al., 2020).

## 9.2 THE GLOBAL CHALLENGE IN REINING IN AI

There is an interplay of three major scenarios around AI that pose significant challenges and threaten to increase inequities.

First, from a nation-state perspective, the development of AI will be viewed from the prism of global competitiveness. If given the resources and investments, nations will push for greater development of AI to either retain their edge over others or catch up with those more advanced. In such a scenario, traditional guardrails will be discarded since they will be seen as impediments to gaining advantage. For nations, not staying ahead in the global AI race will have profound implications for developing and sustaining an economic, military and technological edge over others. Viewed from the Global North-South divide prism, no country will want to restrict its ability to develop AI.

Nations have been competing to develop AI without developing their national strategies on AI. According to OECD, national AI strategies are important to determine how a country sees the role of AI in its social and economic development (Missing Persons: The Case of National AI Strategies, 2023). Without it, AI will be developed on an *ad hoc* basis, and the decision-making will be left to private corporations that model it. For instance, Canada recently announced that it will have a voluntary code of practice to govern generative AI (Tusikov, 2023). This means that nations are willing to yield space to the industry to set the guardrails on issues that not only impact citizens but also shape the global narrative on development of AI. This skews the regulatory landscape in favour of private players whose prime motive will be profit-driven.

A Stanford study reveals that countries that are leading the AI race globally have received large private investments for its development and deployment. India has the highest relative AI skill penetration rate in the world but lacks the investment that is required to utilise it. In 2022 private investments in AI totaled to \$91.9 billion, out of which \$47.36 billion was invested in the US alone, while India had only \$3.24 billion (Maslej et al., 2023). In the absence of AI opportunities in India, its rich talent pool will migrate to countries in the Global North, which will further impede the South's ability to catch up with its peers in the North.

Second, AI will be driven by private companies willing to push research and deployment for higher profit margins (Chiang, 2023). Such scenarios are already playing out where algorithms are designed to create more engaging content at the cost of accuracy.

Corporations have been using cheap labour from the Global South to develop AI. However, people from the Global South, especially women and marginalized communities, are not represented in the datasets that AI is trained on, making it biased against them.

The issues around Kenyan researchers hired for labelling for Open AI highlights the risks of leaving decision-making powers in the hands of private corporations. Sama, the company hired by Open AI to label information for developing ChatGPT, paid the researchers less than \$2 an hour and subjected them to inhumane conditions (Perrigo, 2023a). They were also traumatised by being forced to watch hours of porn and child sexual abuse material (CSAM) as part of their work.

Third, AI will adhere to the Darwinian principle of “survival of the fittest” (Hendrycks, 2023). In this scenario, AI will be subjected to deliver the best intended outcome. For instance, if the AI is being designed and deployed to reduce decision-making processes, it will attempt to do so by only adapting code that allows it to do so. This adds an inherent element of self-preservation to AI, which will seek to be the fittest to survive. This element of self-preservation will eventually lead to a point of sentience where AI can decide what is the “fittest” decision.

Combining these three scenarios – competing nations, profit-driven private companies and the survival-of-the-fittest bias – will eventually set aside any and all traditional guardrails that regulate technology.

### **9.3 GLOBAL INEQUITIES AND THEIR IMPACT ON THE DEVELOPMENT OF AI**

In 1987, India sought to purchase the Cray X-MP14 supercomputer for weather forecasting. However, the US Department of Commerce blocked the sale, citing the possibility of dual use, stating that it could be used for the development of weapons.

The denial of technology to countries like India for multitude reasons has been an integral part of evolution of its indigenous capabilities. Not only did denial lead to delays in several crucial technology development programmes, but it also extended the inequities developed during its years as a colonised nation. Colonisation, in many ways, was also a function of technological development. As the industrial revolution swept through Europe, it brought technological capabilities that sped up production of goods. These goods needed raw material as well as new markets. Both these factors led to advanced European nations seeking to establish new colonies that could not only ensure rights to exploit resources but also dump mass-produced goods and extract wealth.

As a result, colonisation created wealth and established advanced economies that continue to retain their edge over their erstwhile colonies. This is also reflected in the development of AI, ushering in centuries-old inequities and could threaten the return of another form of colonisation.

The top 10 countries investing in AI are enabled with a plethora of resources that are crucial to its development. They have better internet speeds, available at cheaper prices, giving them access to larger datasets, that can then be labelled at very low prices in the Global South, which comprises countries that were colonised in the last 300 years.

The US, which is also the highest investor in AI, is also investing more than the next seven countries put together. Naturally, US tech companies form the bulk of those investments, which come in the form of setting up research labs across the world. A study by Georgetown University's Centre for Security and Emerging Technologies (CSET) of "six US companies with a history of conducting cutting-edge AI research and development" – Apple, Amazon, Google, Microsoft, IBM, Facebook – revealed interesting patterns. Of the six, four companies – Facebook, Google, IBM, and Microsoft – established 62 labs carrying out research in AI (Heston, 2020). Of these, 68% were located outside the US, but the majority of the staff, 68%, were located in the US. Which meant that only 32% of the staff were distributed across the majority of the labs outside the US. This distribution, the study concluded, allowed companies to access global talent while saving costs, accessing markets and adapting products (Heston, 2020).

An abundance of resources also fuels investments and offers a perpetual advantage to the Global North. Better resources lead to higher investments in AI, which produce solutions that dominate markets, consumption and usage in the Global South.

Private investments in AI across countries reveals how advanced economies will continue to dominate in development and deployment of AI, furthering global inequities. The US leads in private investments at US\$248.9 billion between 2013 and 2022, which is more than the combined investments of the next 14 nations in the same period (Maslej et al., 2023, p. 190).

The scale of investments also positions certain regions in the Global North far ahead of the others in the AI readiness index. The North America region, which includes the US and Canada, scores an average of 81.56, which is the highest, with the US taking the top position and Canada the fifth on the same index (Rogerson et al., 2022). The US leads in the number of AI unicorns as well as non-AI unicorns (Maslej et al., 2023), driven by the appreciation of potential benefits in the public and private sectors, cutting across the political spectrum. The report [Rogerson et al.] points out that both also benefit from their geo-political alliance, which includes traditional allies such the UK and Australia, which scored high (UK is third and Canada in fifth position) on the private investments in AI index between 2012 and 2022 (Maslej et al., 2023).

In sharp contrast, the only country from the Global South in terms of investments in the same period is India at US\$7.73 billion (sixth position), but it ends up in 32nd position on the AI readiness index (Rogerson et al., 2022, p. 8). In the South and Central Asia region of the AI readiness index, India is the leader with a score of 63.67 (Rogerson et al., 2022, p. 38), closely followed by Turkey and Kazakhstan, just above the global average score of 44.61 (Rogerson et al., 2022, p. 8).

Chan et al. have argued that bridging this inequity will impose limits on inclusion in the development of AI. Few countries that have succeeded like South Korea did so by reducing dependence on imports, while maximising industries with high export potential through a policy of import substitution industrialisation (ISI) (Chan et al., 2021, p. 5). However, ISI policies in AI will only work if there is focus on high-yield activities such as model deployment and research (Chan et al., 2021) rather than on low-yield activities such as data labelling. Instead of relying on foreign companies to invest in AI, countries will also need to fund domestic AI development, which will enable them to overcome global inequities.

#### 9.4 AI LABOUR FROM THE GLOBAL SOUTH: EXPERIENCES FROM INDIA AND KENYA

Development of AI requires large-scale deployment of technical and human resources. This entails investments in infrastructure and substantial computing power. It is also labour intensive, requiring a skilled workforce. The Global North as a collective has the largest investments in AI and creates the most AI-related jobs. In contrast, many countries from the Global South have AI-skilled workforce seeking employment opportunities.

First, there is a requirement of skilled workers from the fields of Science, Technology, Engineering and Math (STEM) to develop AI models. The Global South has a talent pool of skilled workers looking for AI-related jobs. India, for instance, has the highest number of skilled AI workforce in the world (Maslej et al., 2023, p. 182). Hiring charts reveal, however, that it does not have the jobs for this rich talent pool. Hirings in AI-related jobs is largely centred in the Global North. This means that the skilled workforce from the Global South will migrate to countries with investments and jobs in AI.

This has increasingly been witnessed in STEM fields in the US. In 2019, 23.1% of all STEM workers in the country were immigrants. Among this group, Indian immigrants held the largest share at 28.9% of all foreign-born STEM workers. Workers from Vietnam, Mexico and other countries in the Global South form a large section of STEM workers in the US and have made important contributions to its economy in terms of innovation and productivity (Foreign-born STEM Workers in the United States, 2024).

This phenomenon of migration of skilled workforce to countries with better jobs is not new. The World Bank defines it as “brain drain”, where skilled human resources migrate for trade, education, etc. (Manuel Cunjamá, 2001). India has been witnessing brain drain since the 1960s, and its consequences have been vastly studied. OECD data reveals that India contributes the largest diaspora of highly skilled individuals to OECD countries, with more than 3 million migrants in the category (OECD social, employment and migration working papers, 2020). This has created a reverse phenomenon of “brain gain” for the Global North, where skilled immigrants have made large contributions to its development (Chatterjee, 2022).

Immigrants account for a substantial portion of innovators in the US, with studies attributing 23% of total innovations in the US from 1990–2016, to them (Bernstein et al., 2022). Empirical evidence also shows that there is a positive correlation between migration and productivity in advanced economies (Boubtane & Rault, 2016). Jau-motte et al. found that a 1% increase in migrant adult population results in approximately a 2% increase in productivity and per capita GDP (Impact of Migration on Income Levels in Advanced Economies, 2016). While, the exact increase in productivity and contribution to GDP differs between regions and sectors, similar studies from other advanced countries have verified this positive correlation (Portes et al., 2020).

Secondly, training AI models requires large amounts of labour to build datasets. Manufacturing companies from developed countries have been outsourcing to developing countries since 1970s and utilising their cheap resources to increase profits. In the 1990s, the emergence of the IT sector also witnessed a simultaneous emergence

of IT outsourcing hubs. India is the largest such hub and is projected to reach USD \$8.81 billion in 2023. A global comparison reveals, however, that the most revenue in the IT industry is generated in the US, which will be to the tune of USD \$167.90 billion in 2023.

This pattern is now emerging in the global AI market as well, with large private companies looking at the labour markets from the Global South to build datasets.

Development of AI/ML is 80% data preparation work consisting of collection, labelling and cleaning (Ramnani, 2024). The global data collection and labelling market is expected to reach USD \$8.22 billion by 2028 (“Data Collection and Labeling Market Size Worth \$8.22 Billion by 2028: Grand View Research, Inc.,” 2021), but the benefits of this market accrue mainly to the Global North.

Companies from the Global North crowdsource low-wage workers from the Global South, especially from sub-Saharan Africa and Southeast Asia, to develop these models (Murgia, 2019). These workers are hired by private companies to work as per the contractual conditions set by them.

For instance, ChatGPT, an LLM with over 100 million weekly active users (Malik, 2023), was built on the outsourced labour of Kenyan workers. The reason behind GPT 3’s excellent linguistic capabilities is the large datasets that it has been trained on, which were scraped from the web. Naturally, these datasets also contained a large amount of hate speech and toxicity, which needed to be labelled and then purged from the training data. OpenAI hired Sama, a US-based outsourcing firm, which mainly outsources to countries from the Global South such as Kenya, India and Uganda.

Reports reveal that Kenyan data scientists who were hired to do this job were subjected to toxic work environments and exposed to swathes of traumatising online content containing Child Sexual Abuse Material (CSAM), violence, etc. They received wages as low as USD \$1.3 per hour, while the contract between OpenAI and Sama was worth over USD \$150,000 (Perrigo, 2023b). OpenAI is reportedly earning revenue at the pace of USD \$1.3 billion per year (The Hindu Bureau, 2023). Private corporations have been able to exploit labour from the Global South, at minimal costs, to maximise their profits.

In the absence of AI employment opportunities in the South, therefore, two scenarios are going to play out – migration of its skilled workers and private corporations exploiting labour from the Global South for their profit motives. The economic advantage will lie with the Global North, further deepening the North-South divide.

## 9.5 INSIGHTS FROM NATIONAL AI STRATEGIES

The impetus to develop national AI strategies is largely driven by the location of the countries in the Global North and South. Advanced economies in the Global North tend to cover more ground, aimed at advancing their technological and economic dominance, while those in the Global South cover fewer areas and view the development of AI to address current challenges in governance and bridge the economic and technological gap with their counterparts in the Global North.

A study of AI strategies and priorities in countries in the Global North (UK, US) and in the Global South (India, Kenya) confirms this hypothesis. In Kenya, using

blockchain and AI is viewed to combat endemic corruption (Emerging Technologies for Kenya – Exploration and Analysis, 2019). India views AI as a means for solving “the complexity and multi-dimensional aspects” of its “economic and societal challenges” that can be “easily extended to the rest of the emerging and developing economies” (Kant, 2018).

India currently has established four focus areas as a part of its national strategy and policy for AI, and Kenya has three. The US has ten, Japan has eight and the UK has three areas (“An Overview of National AI Strategies and Policies,” 2021).

The selection of the focus areas by each country also confirms the divergent approach between the Global North and South. While the US covers nearly all key areas (“An Overview of National AI Strategies and Policies,” 2021), it does not identify public administration as one. However, Kenya cites corruption in governance as a key concern and views the deployment of AI and blockchain to address poor administration (“An Overview of National AI Strategies and Policies,” 2021). Both India and Kenya identify healthcare, agriculture and food security as key areas for AI (“An Overview of National AI Strategies and Policies,” 2021).

The UK has identified energy, environment, manufacturing and mobility (“An Overview of National AI Strategies and Policies,” 2021). Only healthcare emerges as a common area for developing AI by all countries, whether North or South. Although there are some similarities in the aims in the healthcare sector, the US policies view it to retain its dominance in the sector. For emerging economies like India and Kenya, achieving universal health coverage at optimum costs by deploying AI is a key concern.

The Global South’s approach to their AI strategies is also guided by historical inequities that emanate from industrialisation and colonisation. As former colonies, countries like India and Kenya view any emerging technology as a means to prevent exploitation of their economies. India’s science, technology and innovation policies (“Science, Technology, and Innovation Policy,” 2020, p. 9) have reflected this theme consistently. Not only does India see technology and innovation as a means to safeguard against economic distress, but it also sees it as a means to ensure that it can continue to corner a sizable chunk of the global economy.

This is further buttressed by the fact that STEM is overwhelmingly embraced as the means to gain economic prosperity at the household level. This has also resulted in shaping India as one of the highest producers of labor skilled in AI, eager to find jobs in the emerging global technology markets.

India’s current AI strategy is two-fold. It aims to use AI as a means to not only build a robust emerging technology and use it to arrive at solutions, but it also provides employment for its vast army of AI-skilled labour.

As a part of its national strategy on AI, the Government of India has identified 16 sectors that have the most potential. These range from agriculture to health, transportation, education and environment, among others (Report of Committee – B on Leveraging A.I. for Identifying National Missions in Key Sectors, sec 4, pp. 6–8). Out of these four, agriculture, healthcare and governance through Digital Public Infrastructure (DPI) have been identified as “key potential growth sectors” (India AI 2023, 2023; sec. Working Group 4, pp. 71–72).

Agriculture, which continues to be a mainstay of India’s economy, also employs the highest number of skilled and unskilled workers. Using AI to increase efficiency

in this sector remains a high priority for India, not only to produce more jobs but also to leverage its datasets to find global technologies and solutions. Similarly, health-care is a potential growth area for India's national AI strategy to not only increase coverage but also build solutions for faster diagnosis, predictive analysis for better treatment and identifying potential new drug candidates (India AI 2023, 2023; sec. Working Group 4, pp. 71–72).

## 9.6 ADDRESSING ALGORITHMIC BIASES

AI models are trained on large datasets. For a given input, an AI establishes patterns within its database to arrive at a suitable output. It has the capability of progressive learning and continually enhances its output as it gets trained. AI decision making can, however, systematically disadvantage certain groups of people. This is defined as “algorithmic bias” and can result from societal biases creeping into datasets or from underrepresentation of certain sections of the population (Barton et al., 2019). AI models that get trained on these datasets further perpetuate historical inequities.

Studies have revealed several such biases. Online recruitment tools deployed by companies were later found to have inherent racial and gender biases (Dave & Dastin, 2020). Historically, women and people of colour were denied employment opportunities in certain industries, resulting in flawed databases, not indicative of the actual employability of this workforce. The AI decision-making on such databases discriminated against these historically disadvantaged groups.

Algorithmic biases have also been found in facial recognition technologies, which were failing to accurately recognize people of colour (Study Finds Gender and Skin-type Bias in Commercial Artificial-intelligence Systems, 2018). This is because the training data was more representative of light-skinned people, resulting in a lack of diversity in the datasets, which produced inaccurate outcomes (Barton et al., 2019).

Research studying fairness in machine learning, however, has largely been limited to the concerns of the Global North. Sambavisan et al. have illustrated that the fairness issues being studied, such as injustices of race and gender, measurement scales and legal tenets, do not hold relevance to the Global South (Sambasivan et al., 2020). When the Global South uses AI models trained on datasets of the Global North, several other forms of fallacies play out.

Machine learning (ML) models such as text-to-image tools trained in the US and Europe lack regional and cultural context, which has resulted in inaccurate outputs for the Global South (Shankar et al., 2017). For instance, studies reveal that they wrongly classified images of grooms from South African and South Asian countries, when compared to images of grooms from the US. They also produced different images for the same word when queried in different languages. DeVries et al. show this through different images being produced for inputs such as “wedding” or “spices” because of different regional and cultural contexts attached to them (DeVries et al., 2019).

The difference in regional and cultural context in the Global South means that it will have to build and label its own datasets that reflect its diversity. The region, however, faces some unique challenges in doing that. Its quality of datasets is impacted by certain key causes, such as internet infrastructure and technology usage patterns, which further causes algorithmic biases.

India, for instance, is the most populous country in the world, but only 52% of its population has access to the internet (Internet in India 2022, 2023). Technology adoption in India varies across regions, religions, caste and gender, which impacts representation in datasets. An Indian government survey from 2019–2021 revealed that while 57.1% of men had access to the internet, only 33.3% of Indian women were connected to the internet. Another report reveals that 71% of internet users in India are urban, while 41% are rural (Internet in India 2022, 2023). India has a diverse population among its states, but internet adoption among different states ranges from 70% to 32% (Internet in India 2022, 2023). Naturally, this will create algorithmic biases against the underrepresented, resulting in flawed and/or negative outcomes against them.

The key priority sectors for adoption of AI are also different for the Global South. This requires AI solutions that are relevant for addressing the regional challenges in these sectors.

Agriculture is one of the most promising areas of AI adoption in the Global South (Wall et al., 2021). The region constitutes some of the largest agriculture economies (Wunsch, 2024), for whom increasing crop productivity is a key priority.

As per the Food and Agriculture Organisation (FAO), the Global North is a net exporter of commodity crops, while the Global South is a net importer. The South countries, on the other hand, are net exporters of fruits, vegetables, fats, oils and tropical products (Agricultural Trade in the Global South, 2022). Agritech companies based in the Global North have AI models skewed in favour of commodity crops and focused on large-scale farming systems (Gardezi et al., 2022). This does not meet the needs of small-scale, ecologically diverse farming prevalent in the Global South.

Health is another priority sector for the Global South (Wall et al., 2021), where the region grapples with disproportionate prevalence of certain diseases, such as dengue, tuberculosis, ebola, etc. Most prevalent databases in clinical AI, however, are from high-income countries, with almost 40% of data being attributed to the US. Models trained in the Global North pose an imminent risk of inaccurate diagnosis for patients from the Global South due to differences in genetic composition, climate, food habits and living conditions. A comparative study on early detection of breast cancer, between sub-Saharan Africa (SSA) and high-income countries, found that what has been successful in the West is not effective in reducing mortality from breast cancer in the SSA (Black & Richmond, 2019).

Models trained in the Global South, on the other hand, have proved to be much more relevant, precise and reliable in their diagnosis. Deep learning models developed to detect vision-threatening diabetic retinopathy in Africa have been validated in Zambia for their accuracy comparable to human graders (Bellemo et al., 2019). Models developed for screening diabetic patients in India have also been verified to have produced accuracy equal to, and even exceeding, that of human graders (Gulshan et al., 2019).

The Global South, therefore, needs to be building its own AI capabilities that work for the unique challenges facing the region. South-South cooperation (SSC) has been recognized as an effective instrument for catalysing economic development through an exchange of innovation and good practices. This model, which has seen success in sectors such as agriculture, can be leveraged in the field of AI to



collectively build capabilities and solutions suitable for the South. Regional cooperation in infrastructure, and research and development, can produce a shared pool of resources that attract further investments in the region. Such collective capacity building will not only make its datasets richer and algorithmic decisions fairer, but also retain the benefits of AI in the region.

## 9.7 AI AND SURVIVAL OF THE FITTEST AND REGULATORY CHALLENGES FOR THE GLOBAL SOUTH

The “risk of extinction” (*Statement on AI Risk*, n.d.a) from AI has been cited by a group of researchers and heads of several technology companies as one of the dangers of AI. While this view has been widely disputed (Heaven, 2023), several risks have to be factored in. Dan Hendrycks, the director of the Center for AI Safety, which hosted the statement on the “risk of extinction”, has argued that the development of AI will closely mirror Charles Darwin’s theory of the survival of the fittest (Hendrycks, 2023).

Hendrycks’s theory is based on competition dynamics, where companies and countries investing in development of AI will focus on efficiency. As they weed out code that is considered “inefficient”, it will also encode elements of self-preservation within the AI. If it develops sentience, Hendrycks argues, it will not only pose a challenge to other AI but also strive to arrive at solutions it considers to be the most efficient in its estimation.

This theory will extend the “black box” problem of AI, the inability to see how deep learning systems make their decisions (AI’s Mysterious ‘Black Box’ Problem, Explained, n.d.). While the black box problem compounds issues of algorithmic biases, it is also unclear how they arrive at decisions. Hendrycks has argued (Hendrycks, 2023) that this is already at play, as companies such as streaming platforms and social media companies write code to result in higher engagement from users. To do so, code that generates higher engagement is retained over less efficient code that yields lesser engagement.

Professor Nick Bostrom famously demonstrated this through a thought experiment called “paperclip maximiser” (Ethical Issues in Advanced Artificial Intelligence, n.d.), where a super-intelligent AI is given a simple task of producing paperclips. The experiment reveals that the AI will dedicate itself to this single goal, becoming increasingly efficient at the output and eventually monopolising all resources to inundate the world with paperclips. Any attempt to switch off the AI will be interpreted by it as an existential threat, thus making it near-impossible to turn it off. This is called the “control problem” where a super-intelligent AI develops powers to appropriate resources to achieve its output and ultimately preserve its own existence (AI And the Paperclip Problem, 2018).

Applied to the development of AI, this weakens traditional guardrails associated with technology, and could also deepen existing global inequities. As argued earlier in this chapter, the Global North and South have different motivations and aims to develop and deploy AI. No country will be willing to lose out on the global AI race. However, unlike other sensitive technologies, such as the development of nuclear weapons or genetics, where the dangers and capacity for destruction are well quantified, AI will pose very different challenges in terms of regulatory mechanisms.

Regulation of technology is complex and, when viewed from the prism of national and economic growth, lesser restrictions are preferred. The development of AI in the North America region is driven by private corporations as well as government departments like the Defense Advanced Research Projects Agency (DARPA) in the US (Rogerson et al., 2022). National AI strategies are in place to not only address historical inequities but also to harness a greater portion of the global economy through applications and uses in various lucrative sectors.

Unlike past regimes, such as the Nuclear Proliferation Treaty (The IAEA and the Non-Proliferation Treaty, n.d.), the Comprehensive Nuclear-Test-Ban Treaty (The Comprehensive Nuclear-Test-Ban Treaty (CTBT), n.d.), the Missile Technology Control Regime (MTCR) (The Missile Technology Control Regime at a Glance | Arms Control Association, n.d.) or the Wassenaar Arrangement, there is no global agreement on the development of AI. This ensures that global regulatory frameworks that have acted as guardrails to prevent the uncontrolled spread of harmful technologies will no longer be applicable. This will not only free nations to pursue their AI strategies as they see fit, but it will also ensure that private corporations driven by profit will invest in AI under lesser or minimal regulations.

For countries in the Global South, this poses a complex challenge. To overcome historical inequities, countries in the Global South will need to fund and encourage the development of AI in critical sectors that have the highest potential for exports as well as economic development and jobs. They will also be wary of controls, if any, being imposed by countries in the Global North or their private corporations.

Warnings, such as the one issued by a coalition of AI experts (*Statement on AI risk*, n.d.b) about “mitigating the risk of extinction from AI”, will be seen by the Global South as ways to limit its development to a few countries or alliances in the Global North. The fact that signatories to the statement has heads of major technology companies, which have already invested large sums in the development of AI, will further heighten suspicion among countries of the Global South, while also ensuring that traditional guardrails, if any, are met with resistance, if not ignored altogether.

## 9.8 DESIGNING AI TO OVERCOME GLOBAL INEQUITIES

The Global South needs AI solutions that are relevant to its regional context and meet the unique challenges that the region faces. The lack of investments and infrastructure in the region is a setback for technology development. Not only is the region missing out on its tremendous AI opportunity but it is simultaneously facing a loss of its human capital. Regional cooperation can lead to a shared pool of resources, capabilities and infrastructure, which may be otherwise difficult to achieve at a country level.

For AI to take root in the Global South, skills have to be nurtured within local communities in a sustainable manner. The Deep Learning Indaba (Our Mission – Deep Learning Indaba, 2023) experiment is a step in that direction. In 2016, the 30th conference in Neural Information Processing did not have a single paper from researchers in African institutions (Maryatt, 2018). With an objective to address this gap, researchers founded Deep Learning Indaba, which focused on researchers from Africa. Data Science Africa is a similar initiative (DSA | Home, n.d.) that

started a year before Indaba and has initiated a number of initiatives to support and nurture researchers across Africa – from preparing notes and lectures to providing an online platform for data science conferences and a forum for discussions and learnings from each other. Not only has Data Science Africa worked on creating more capacity, with initiatives like Indaba, but it has also deepened the understanding of machine learning and data science within local communities (DSA | Home, n.d.)

Countries in the Global South can also build more South-South alliances on developing AI. This has the potential to offer multiple benefits. Tasks like data labelling can be harnessed to jointly design data models in the Global South through state funding and tax incentives to local companies. Not only will it address the economic issues of developing AI, but it will also address the problem of algorithmic biases since the datasets are local and contextual to the Global South. The richness and diversity of datasets in the Global South will also provide a major advantage to developing economies that develop and build while harnessing a larger portion of the global digital economy. These could also be achieved through South-South alliances such as BRICS and ASEAN.

Regional inequities perpetuated by current AI models can also be addressed by retaining the rights over datasets extracted from the Global South. Creating GI tags for datasets to ensure that the sources receive the benefit that accrues from labelling and modelling them could be one solution. Companies are already beginning to demand a share of profits from AI companies for using their data to build large language models (LLMs) and these profit-sharing models could be replicated to communities that provide labour to build and label datasets. AI based on datasets extracted, labelled and designed by the Global South can prioritise the benefits for themselves.

A combination of resources, capabilities and infrastructure will further attract investments and create more jobs, resulting in a virtuous cycle. This will lead to retention of economic benefits within the region and collective growth of economies. By developing indigenous AI, the region will be able to focus on its key priority sectors and build solutions attuned to its regional context.

Simultaneously, countries in the region should firm up their national AI policies and place necessary guardrails on development and deployment of AI. This will work as an oversight mechanism to guard against uninhibited development of AI tools by private corporations to drive profits. Sound AI policies can solve for key areas of concern such as exploitation of labour and algorithmic biases.

Lastly, investments in research and fostering a culture of knowledge sharing will lead to preventive measures against biases and provide support in creating relevant and inclusive AI models for the region.

## REFERENCES

- Agricultural trade in the Global South.* (2022). FAO. <https://www.fao.org/3/cb9120en/cb9120en.pdf>
- AI and the paperclip problem.* (2018, June 10). CEPR. <https://cepr.org/voxeu/columns/ai-and-paperclip-problem>
- AI's mysterious 'black box' problem, explained.* (n.d.). University of Michigan-Dearborn. <https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained>

- Barton, G., Lee, N. T., & Resnick, P. (2019, May 22). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Brookings*. <https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>
- Baudry, P., Doms, M., & Lewis, E. (2006, June). *The IT revolution at the city level: Testing a model of endogenous biased technology adoption*. Manuscript, Dartmouth University.
- Bellemo, V., Lim, Z. W., Lim, G., Nguyen, Q. D., Xie, Y., Yip, M. Y. T., Hamzah, H., Ho, J., Lee, X. Q., Hsu, W., Lee, M. L., Musonda, L., Chandran, M., Chipalo-Mutati, G., Muma, M., Tan, G. S. W., Sivaprasad, S., Menon, G., Wong, T. Y., & Ting, D. S. W. (2019). Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: A clinical validation study. *The Lancet Digital Health*, 1(1), e35–e44. [https://doi.org/10.1016/s2589-7500\(19\)30004-4](https://doi.org/10.1016/s2589-7500(19)30004-4)
- Bernstein, S., Diamond, R., Jiranaphawiboon, A., McQuade, T., & Pousada, B. (2022). *The contribution of high-skilled immigrants to innovation in the United States*. <https://doi.org/10.3386/w30797>
- Black, E., & Richmond, R. (2019). Improving early detection of breast cancer in sub-Saharan Africa: Why mammography may not be the way forward. *Globalization and Health*, 15(1). <https://doi.org/10.1186/s12992-018-0446-6>
- Boubtane, E., & Rault, J. D. A. C. (2016). Immigration and economic growth in the OECD countries 1986–2006. *Oxford Economic Papers*, 68(2), 340–360. <https://www.jstor.org/stable/44122852>
- Chan, A., Okolo, C. T., Turner, Z., & Wang, A. (2021, February 2). *The limits of global inclusion in AI development*. arXiv.org. <https://arxiv.org/abs/2102.01265>
- Chatterjee, R. (2022, August 1). A crisis in human capital: Understanding the Indian brain drain. *The BPR*. <https://www.bostonpoliticalreview.org/post/a-crisis-in-human-capital-understanding-the-indian-brain-drain>
- Chiang, T. (2023, May 4). Will A.I. become the new McKinsey? *The New Yorker*. <https://www.newyorker.com/science/annals-of-artificial-intelligence/will-ai-become-the-new-mckinsey>
- The comprehensive nuclear-test-ban treaty (CTBT)*. (n.d.). CTBTO. <https://www.ctbto.org/our-mission/the-treaty>
- Data collection and labeling market size worth \$8.22 billion by 2028: Grand view research, Inc. (2021, February 9). *PR Newswire*. <https://www.prnewswire.com/news-releases/data-collection-and-labeling-market-size-worth-8-22-billion-by-2028-grand-view-research-inc-301224275.html>
- Dave, P., & Dastin, J. (2020, December 4). Top AI ethics researcher says Google fired her; company denies it. *Reuters*. <https://www.reuters.com/article/uk-alphabet-google-research-idUKKBN28D3JP/>
- DeVries, T., Misra, I., Wang, C., & Laurens, V. D. M. (2019, June 6). *Does object recognition work for everyone?* arXiv.org. <https://arxiv.org/abs/1906.02659>
- DSA. (n.d.). *Home*. <http://www.datasienceafrica.org/aboutus/>
- Emerging technologies for Kenya – exploration and analysis*. (2019). Ministry of Information, Communication and Technology. <https://repository.ca.go.ke/server/api/core/bitstreams/08e8854d-1ece-4bd2-a2ff-740526da7a02/content>
- Ethical Issues in Advanced Artificial Intelligence. (n.d.). <https://nickbostrom.com/ethics/ai>
- Foreign-born STEM workers in the United States*. (2024, October 8). American Immigration Council. <https://www.americanimmigrationcouncil.org/research/foreign-born-stem-workers-united-states>
- Gardezi, M., Adereti, D. T., Stock, R., & Ogunyiola, A. (2022). In pursuit of responsible innovation for precision agriculture technologies. *Journal of Responsible Innovation*, 9(2), 224–247. <https://doi.org/10.1080/23299460.2022.2071668>
- Government of India & Government of the United States of America. (n.d.). *Frequently asked questions on the India –US agreement for co-operation concerning peaceful uses of nuclear energy*. <https://2009-2017.state.gov/documents/organization/122068.pdf>

- Gulshan, V., Rajan, R. P., Widner, K., Wu, D., Wubbels, P., Rhodes, T., Whitehouse, K., Coram, M., Corrado, G., Ramasamy, K., Raman, R., Peng, L., & Webster, D. R. (2019). Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmology*, 137(9), 987. <https://doi.org/10.1001/jamaophthalmol.2019.2004>
- Heaven, W. D. (2023, June 20). How existential risk became the biggest meme in AI. *MIT Technology Review*. <https://www.technologyreview.com/2023/06/19/1075140/how-existential-risk-became-biggest-meme-in-ai/>
- Hendrycks, D. (2023, May 31). The Darwinian argument for worrying about AI. *TIME*. <https://time.com/6283958/darwinian-argument-for-worrying-about-ai/>
- Heston, R. (2020). *Mapping U.S. multinationals' global AI R&D activity*. <https://doi.org/10.51593/20190008>
- The Hindu Bureau. (2023, October 17). OpenAI revenue \$1.3 billion annualised rate, says report. *The Hindu*. <https://www.thehindu.com/sci-tech/technology/openai-revenue-13-billion-annualised-rate-says-report/article67416174.ece>
- The IAEA and the non-proliferation treaty*. (n.d.). <https://www.iaea.org/topics/non-proliferation-treaty>
- Impact of Migration on Income Levels in Advanced Economies. (2016). *International monetary fund eBooks*. <https://doi.org/10.5089/9781475545913.062>
- India AI 2023. (2023). MeiTY. <https://indiaai.s3.ap-south-1.amazonaws.com/docs/IndiaAI+Expert+Group+Report-First+Edition.pdf>
- Internet in India 2022*. (2023). IAMAI. [https://www.iamai.in/sites/default/files/research/Internet%20in%20India%202022\\_Print%20version.pdf](https://www.iamai.in/sites/default/files/research/Internet%20in%20India%202022_Print%20version.pdf)
- Kant, A. (2018). *National Strategy for Artificial Intelligence*. NITI Aayog. <https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf>
- Malik, A. (2023, November 6). OpenAI's ChatGPT now has 100 million weekly active users. *TechCrunch*. <https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/>
- Manuel Cunjamá. (2001). *World development report 2000/2001 at tacking poverty*. Oxford University Press. <https://documents1.worldbank.org/curated/en/230351468332946759/pdf/World-development-report-2000-2001-attacking-poverty.pdf>
- Maryatt, E. (2018, September 10). *Deep learning Indaba 2018: Strengthening African machine learning*. Microsoft Research. <https://www.microsoft.com/en-us/research/blog/deep-learning-indaba-2018-strengthening-african-machine-learning/>
- Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2023, October 5). *Artificial intelligence index report 2023*. arXiv.org. <https://arxiv.org/abs/2310.03715>
- The missile technology control regime at a glance*. (n.d.). Arms Control Association. <https://www.armscontrol.org/factsheets/mtrcr>
- Missing persons: The case of national AI strategies*. (2023, September 21). Centre for International Governance Innovation. <https://www.cigionline.org/publications/missing-persons-the-case-of-national-ai-strategies/>
- Mohamed, S., Png, M., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4), 659–684. <https://doi.org/10.1007/s13347-020-00405-8>
- Murgia, M. (2019, July 24). AI's new workforce: The data-labelling industry spreads globally. *Financial Times*. <https://www.ft.com/content/56dde36c-aa40-11e9-984c-fac8325aaa04>
- OECD Social, Employment and Migration Working Papers No. 239. (2020). [https://one.oecd.org/document/DELSA/ELSA/WD/SEM\(2020\)4/En/pdf](https://one.oecd.org/document/DELSA/ELSA/WD/SEM(2020)4/En/pdf)
- Our Mission – Deep Learning Indaba*. (2023, March 21). *Deep Learning Indaba*. <https://deeplearningindaba.com/about/our-mission/>

- An Overview of National AI Strategies and Policies. (2021). *OECD going digital toolkit notes*. <https://doi.org/10.1787/c05140d9-en>
- Perrigo, B. (2023a, January 18). Exclusive: OpenAI used Kenyan workers on less than \$2 per hour to make ChatGPT less toxic. *TIME*. <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Perrigo, B. (2023b, January 18). Exclusive: OpenAI used Kenyan workers on less than \$2 per hour to make ChatGPT less toxic. *TIME*. <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Portes, J., Campo, F., & Oommen, E. (2020). *The economic contribution of Indian migrants to the EU: Two sector case studies*. International Labour Organization. [https://www.ilo.org/sites/default/files/wcmstp5/groups/public/@asia/@ro-bangkok/@sro-new\\_delhi/documents/publication/wcms\\_750860.pdf](https://www.ilo.org/sites/default/files/wcmstp5/groups/public/@asia/@ro-bangkok/@sro-new_delhi/documents/publication/wcms_750860.pdf)
- Ramnani, M. (2024, December 31). Is AI fast becoming a technology built on worker exploitation from Global South? *Analytics India Magazine*. <https://analyticsindiamag.com/is-ai-fast-becoming-a-technology-built-on-worker-exploitation-from-global-south/>
- Report of the Steering Committee on Science and Technology for Eleventh Five Year Plan (2007–12). (2006). <https://dst.gov.in/sites/default/files/rep-s-t.pdf>
- Rogerson, A., Hankins, E., Fuentes Nettel, P., & Rahim, S. (2022). *Government AI readiness index 2022*. [www.unido.org/sites/default/files/files/2023-01/Government\\_AI\\_Readiness\\_2022\\_FV.pdf](http://www.unido.org/sites/default/files/files/2023-01/Government_AI_Readiness_2022_FV.pdf)
- Sambasivan, N., Arnesen, E., Hutchinson, B., & Prabhakaran, V. (2020, December 3). *Non-portability of algorithmic fairness in India*. arXiv.org. <https://arxiv.org/abs/2012.03659>
- Science, Technology, and Innovation Policy. (2020). *Draft STIP Doc 1.4*. Ministry of Science & Technology. [https://dst.gov.in/sites/default/files/STIP\\_Doc\\_1.4\\_Dec2020.pdf](https://dst.gov.in/sites/default/files/STIP_Doc_1.4_Dec2020.pdf)
- Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017, November 22). *No classification without representation: Assessing geodiversity issues in open data sets for the developing world*. arXiv.org. <https://arxiv.org/abs/1711.08536>
- Sinha, A. (2017, June 5). Tech denied, ISRO built cryo engine on its own. *The Indian Express*. <https://indianexpress.com/article/explained/tech-denied-isro-built-cryo-engine-on-its-own-4690709/>
- Statement on AI risk*. (n.d.a). CAIS. <https://www.safe.ai/statement-on-ai-risk>
- Statement on AI risk*. (n.d.b). CAIS. <https://www.safe.ai/statement-on-ai-risk#open-letter>
- Study finds gender and skin-type bias in commercial artificial-intelligence systems. (2018, February 11). *MIT News | Massachusetts Institute of Technology*. <http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>
- Tusikov, N. (2023, September 13). *Voluntary AI guardrails risk placing corporate interests over public good*. Centre for International Governance Innovation. <https://www.cigionline.org/articles/voluntary-ai-guardrails-risk-placing-corporate-interests-over-public-good/>
- Vasuki, S. (2013, October 11). Indian supercomputer progresses. *India Today*. <https://www.indiatoday.in/magazine/economy/story/19890315-indian-supercomputer-progresses-815885-1989-03-14>
- Wall, P. J., Saxena, D., & Brown, S. (2021, August 23). *Artificial intelligence in the Global South (AI4D): Potential and risks*. arXiv.org. <https://arxiv.org/abs/2108.10093>
- Wunsch, N.-G. (2024, December 19). *Leading agricultural producers worldwide in 2022*. Statista. <https://www.statista.com/statistics/1332343/the-leading-producers-of-agricultural-goods-worldwide/>

---

# 10 Assessing the Trustworthiness of Generative AI Used in Higher Education

*Adarsh Srivastava, Gokul Gawande, Divya Dwivedi, Manu Dev, Vinayak Kottawar, Ishwar Chavhan, and Roberto V. Zicari*

## 10.1 INTRODUCTION

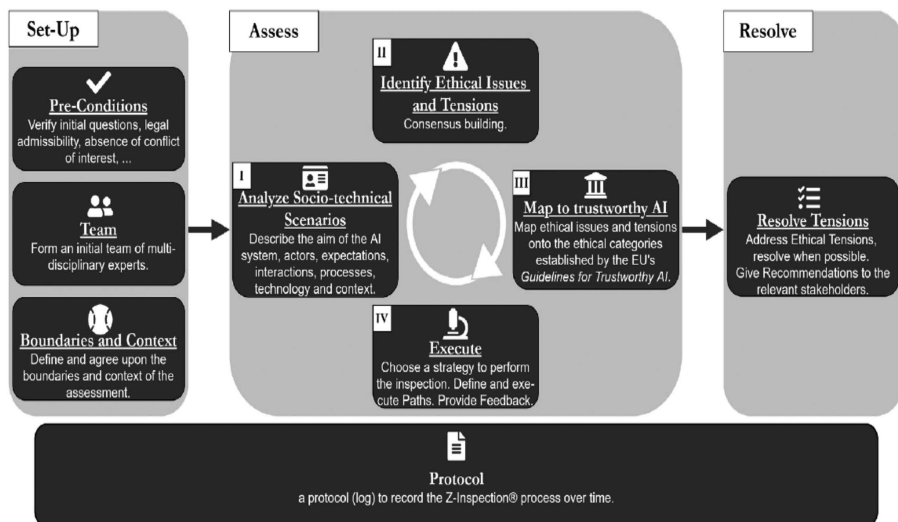
This pilot project intends to assess the trustworthiness of the use of generative AI in the education domain with the Z-Inspection® AI assessment framework that aims at specific use cases in higher-level education. For this pilot project, we will assess the ethical, technical, domain-specific (i.e. education), and legal implications of the use of Generative AI products/services within the university context.

We will follow the UNESCO guidance for policymakers on AI and education (Miao & Holmes, 2021). In particular, policy recommendation 6: Pilot testing, monitoring and evaluation, and building an evidence base. The expected output of this research activity will be best practice and a set of recommendations for each specific use case (published in white paper, a peer-reviewed journal article). Such recommendations could also be useful to further establish the guidelines that each university is creating for the use of generative AI in education.

## 10.2 APPROACH

An interdisciplinary team of experts assess the trustworthiness of Generative AI for selected use cases in higher education using the Z-Inspection® process. Z-Inspection® is a holistic process based on the method of evaluating new technologies, where ethical issues need to be discussed through the elaboration of socio-technical, socio-legal, and socio-economic scenarios. In particular, Z-Inspection® can perform independent assessments and/or self-assessments with the stakeholders owning the use case.

The Z-Inspection process in a nutshell is composed of three main phases: (1) the Set-Up phase; 2) the Assess phase; and 3) the Resolve phase (Figure 10.1). The Set-Up phase clarifies some preconditions, sets the team of investigators,



**FIGURE 10.1** Z-Inspection process in a nutshell.

helps define the boundaries of the assessment, and creates a protocol. The Assess phase is composed of four tasks: (1) analyzing AI system usage; (2) identifying possible ethical issues, as well as technical and legal issues; (3) mapping such issues to the trustworthy AI ethical values and requirements; and (4) verifying such requirements. The Resolve phase addresses resulting ethical, technical, and legal issues, addresses when possible ethical tensions arise, and produces recommendations when required to prescribe a so-called ethical AI maintenance over time.

For the context of this pilot project, we will define *ethics* in line with the essence of modern democracy, i.e., “respect for others, expressed through support for fundamental human rights”. We take into consideration that “trust” in the utilisation of AI systems concerns not only the technology’s inherent properties but also the qualities of the socio-technical systems involving AI applications. Specifically, we consider the ethics guidelines for trustworthy artificial intelligence defined by the EU High-Level Expert Group on AI (European Commission High-Level Expert Group on AI, 2019), which define *trustworthy AI* as:

1. lawful – respecting all applicable laws and regulations
2. ethical – respecting ethical principles and values
3. robust – both from a technical perspective and taking into account its social environment

In addition to these guidelines, we will also use the four ethical principles, rooted in fundamental rights defined in Stanford HAI, “ChatGPT Out-scores Medical Students



on Complex Clinical Care Exam Questions” (Hadhazy, 2023), and acknowledging that tensions may arise between them in relation to:

- 1. Respect for human autonomy
- 2. Prevention of harm
- 3. Fairness
- 4. Explicability

Furthermore, we also consider the seven requirements of Trustworthy AI defined by the High-Level Experts Group set by the EU (European Commission High-Level Expert Group on AI, 2019). Each requirement has a number of sub-requirements, as indicated in Table 10.1.

While we consider the seven requirements to be comprehensive, we believe additional ones can still bring value. Two such additional requirements proposed by the Z-Inspection® initiative are “Assessing if the ecosystems respect values of modern democracy” and “Avoiding concentration of power”. We will also take UNESCO guidance for policymakers on AI and education that will help set out policy recommendations in seven areas into account:

- 1. A system-wide vision and strategic priorities
- 2. Overarching principle for AI and education policies
- 3. Interdisciplinary planning and inter-sectoral governance
- 4. Policies and regulations for equitable, inclusive, and ethical use of AI
- 5. Master plans for using AI in education management, teaching, learning, and assessment
- 6. Pilot testing, monitoring and evaluation, and building an evidence base
- 7. Fostering local AI innovations for education

**TABLE 10.1**  
**Requirements and Sub-Requirements of Trustworthy AI**

Sr. No.	Requirements and Sub-Requirements of Trustworthy AI
1.	Human agency and oversight – Including fundamental rights, human agency, and human oversight
2.	Technical robustness and safety – Including resilience to attack and security, fall-back plan and general safety, accuracy, reliability, and reproducibility.
3.	Privacy and data governance – Including respect for privacy, quality and integrity of data, and access to data.
4.	Transparency – Including traceability, explainability, and communication
5.	Diversity, non-discrimination, and fairness – Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.
6.	Societal and environmental well-being – Including sustainability and environmental friendliness, social impact, society, and democracy.
7.	Accountability – Including audibility, minimisation and reporting of negative impact, trade-offs, and redress.

### 10.3 BACKGROUND OF USE CASE

At the beginning of OpenAI's ChatGPT excitement, when the public became aware of Generative AI tools such as ChatGPT, Bard, Claude, etc., which provide free access to its basic plan, many were quick to use these tools, and students were no exception. Use of these AI tools allowed them more time to focus on other tasks. The education institutes had their reservations about the use of these tools due to multiple factors, such as their lack of reliability, truthfulness, and explainability.

Nevertheless, generative AI has now seamlessly integrated into the day-to-day operations education sector, particularly in higher education, impacting both students and educators. A notable case study at the esteemed D.Y. Patil College of Engineering (AK), Pune, India, sheds light on the transformative role of generative AI within this educational setting.

In the outcome-based education framework, teaching faculty leverage the capabilities of generative AI, particularly through platforms such as ChatGPT, Bard, etc. The teaching faculty at this college utilise this technology to craft study plans from the comprehensive course outline. They do this by using complex prompts, primarily by leveraging Prompt Engineering. What makes this particularly noteworthy is the alignment of these plans with the mandates set forth by educational regulatory bodies such as the All-India Council for Technical Education (AICTE). By using generative AI platforms like ChatGPT or Bard in the planning process, educators are tailoring their courses to meet the specific Course Outcome and map these course outcomes with Program Outcomes (PO) at large, as stipulated by these regulatory bodies more effectively and in relatively less time than earlier, when the same activity was done manually without the assistance of such generative AI solutions. They can now complete the tasks relatively faster and better, whereas it would have taken considerable human hours – in contrast. However, they have not completely removed human intervention from the process. The output of such GenAI is assessed and validated manually by professors before it is utilised further. This way a human is always in the loop, performing as a major actor in the process.

There are more applications of generative AI solutions that are found, explored, and utilised by the Artificial Intelligence and Data Science department of D.Y. Patil College of Engineering (AK), Pune. The teachers have leveraged ChatGPT/Bard to brainstorm innovative teaching methods and pedagogical approaches tailored to the needs and learning patterns of their students. Based on the prompts, the model provides suggestions to the diverse needs of the students, offering insights along with step-by-step instructions on the strategies and active learning techniques. The teachers have also leveraged the generation capabilities of this new technology by utilising it to develop different assessment tools, including quizzes and project topics based on the SPPU rubrics.

### 10.4 NEED FOR ASSESSMENT

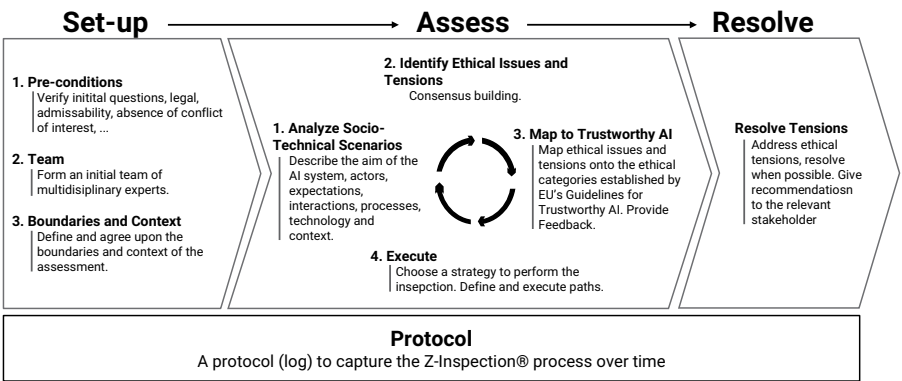
As generative AI solutions are being increasingly integrated into higher education, the need for transparency and trustworthiness becomes crucial. This can be achieved

in two stages to ensure optimal results. The first would be to create a comprehensive guideline and framework for the use of such solutions. Second, educators should be provided with clear guidelines for the responsible use of generative AI in the teaching and learning process. The large language model (LLM) is trained on huge amounts of text data, which can reflect and perpetuate the biases present in the data, and it has the potential to harm the teaching and learning process, leading to unwanted and unforeseen outcomes which are unintended, particularly when the model is used in a decision-making context.

The primary purpose of the guideline would be to familiarise teachers with the intended purpose of using generative AI models in education within the institution. Although many teachers might have implemented it, they may not fully understand the know-how or the limitations of the model, or how to effectively integrate these models into the teaching methods. The purpose is also to clarify the extent of the LLM applications within teaching roles, enabling educators to critically assess and identify any biases in the content.

### 10.5 Z-INSPECTION PROCESS

We used a process to assess trustworthy AI in practice, called Z-Inspection, which expands upon the “Framework for Trustworthy AI” defined by the High Level Experts Groups set up by the European Commission. Z-Inspection is a holistic process (Zicari et al., 2021) based on the method of evaluating new technologies according to which ethical issues must be discussed through the elaboration of socio-technical scenarios. The Z-Inspection process is depicted in Figure 10.2, and it is composed of three main phases: (1) the Set-Up phase; 2) the Assess phase; and 3) the Resolve phase. The process has been successfully applied to assess the trustworthiness of the generative AI system used in the education sector.



**FIGURE 10.2** Z-Inspection® process flowchart describing the main steps of the Set-Up, Assess, and Resolve phases.

## 10.6 SET-UP PHASE: CREATION OF INTERDISCIPLINARY TEAM

For the Set-Up phase, we established an interdisciplinary assessment team composed of diverse experts. For this use case, the team included philosophers, education ethicists, education domain experts (including HOD and researchers), legal researchers, ethics advisors, social scientists, computer scientists, and student representatives. The selection of experts was crucial, as the quality of the analysis and outcomes relied heavily on their diligent selection and qualifications. This includes ensuring impartiality and freedom from potential conflicts of interest. Domain experts needed to encompass various areas of expertise and practice, particularly for a tool potentially affecting the workflows of different professionals.

Special considerations were made for the potential behavioral bias of stakeholders involved in the evaluation process of the use case. Team members were primarily selected based on their required skills and expertise. To maintain the inspection process's integrity, it is crucial that all members respect each other's specific areas of competency. Subsequent additions to the team were limited, preferably avoided altogether, to preserve the balance of perspectives and ensure team workflow stability. The team composition is as follows:

- **Lead:** Coordinates the overall process, ensures completion of the interim issues report.
- **Moderator:** Documents all zoom meetings through MoM in a shared Google Doc.
- **Ethicist(s):** Help the team in identifying ethical tensions arising from the use case.
- **Domain Expert(s):** Provide specialist knowledge and insights for education sector.
- **Legal Expert:** Assists with knowledge of relevant legal domain, data protection, and human rights.
- **Technical Expert(s):** Has a specialty in machine learning, deep learning, and data science. The team also includes social scientists and communication specialists.
- **Philosophers:** Act as advisors to the team. Assist in interpreting ethical principles, UNESCO guidance, and EU guidelines for trustworthy AI.

The interdisciplinary nature of our assessment team is most important in ensuring that diverse perspectives are incorporated when evaluating the trustworthiness of an AI system.

***Split the Work in Working Groups:*** Initially, the expert team met together with the stakeholders who own the use case, in several workshops (via Google Meet) to define socio-technical scenarios for the AI system's use. The term "stakeholders" signifies actors with direct ownership of the AI system throughout this chapter. Later, the team was divided into different working groups (WGs), grouped together by homogeneous expertise, i.e., seven WGs:

- **WG Technical:** Composed of experts in machine learning, data science, and deep learning.

- **WG Ethics:** Composed of experts in ethics.
- **WG Ethics/Education:** Composed of experts in education ethics.
- **WG Teaching staff/Others:** Composed of experts in various areas of education.
- **WG Law:** Composed of experts in law, data privacy, and data protection.
- **WG Student representative:** Composed of a single person, a student representative.
- **WG Lead:** Composed of experts who coordinate the assessment.

**Creation of Reports:** Operating independently and in parallel to mitigate cognitive bias and leverage diverse perspectives, each WG will analyse the socio-technical scenarios and generate preliminary reports in free text. These reports will be shared with the entire team for feedback and comments, facilitating interdisciplinary interaction among experts from various backgrounds. This collaborative process allows each WG to incorporate the viewpoints of other experts before finalising their reports describing the possible risks and issues found during the analysis of the AI system, including ethical, technical, and domain-specific (i.e., education) issues. In this chapter, we will not consider legal issues.

**Mappings to the Framework of Trustworthy AI:** Each WG uses standardised templates (rubrics) to map the identified issues, described in free text, to the four ethical principles and seven requirements outlined in the EU framework for trustworthy AI. This mapping process transforms the reports from open vocabulary (free text) to closed vocabulary (i.e., the templates). These mappings enable the diverse perspectives from different WGs to be systematically analysed and compared. Notably, each WG operates independently and adopts varying strategies to perform this mapping exercise.

**Consolidation Process of Mapping:** At this point, we consolidate the mappings produced by the various WGs into a consistent list. This is done by creating a dedicated WG that consolidates the issues according to their mapping to the requirements of the EU framework for trustworthy AI. The consolidated lists of WG issues for each of the seven requirements are reviewed so that commonalities and differences can be identified and discussed before the final consolidation. The method highlights how different perspectives could lead to similar issues being mapped to different requirements.

**Give Recommendations:** The “Resolve” phase completes the process by addressing ethical tensions and by giving recommendations to the key stakeholders. It is crucial to monitor that the AI system that fulfilled the trustworthy AI requirement continues to do so over time. Therefore, when required, the Resolve phase includes conducting trustworthy monitoring over time of the AI system (we call it “ethical maintenance”).

## 10.7 ASSESSING TRUSTWORTHY AI: USING GENERATIVE AI FOR AN OUTCOME-BASED EDUCATION FRAMEWORK IN HIGHER EDUCATION

The Assess phase of the process begins with the creation of socio-technical scenarios.

### 10.7.1 SOCIO-TECHNICAL SCENARIOS

We considered three possible scenarios in which the AI system could be used.

1. The current scenario is at three departments in the D.Y. Patil College of Engineering (AK). Generative AI is used to help them with planning the courses to be taught to students. They further use it to derive the larger picture by using generative AI to produce Course Outcomes and then to map these with the Program Outcome provided by the AICTE. However, manual validation of generative AI output is performed by senior professors.
2. Possible future extension includes generative AI being used by all the departments to perform the same tasks. Further applications of the system include using GenAI for different tasks, such as explaining complex topics in an easily understandable way for students, to plan the lecture schedule, designing tests, etc. An initial prototype is already in progress.
3. A potential future application could be implementing this system on a large scale with established guidelines, monitoring, and governance, which can be adopted by other universities.

***Aim of the AI System:*** The primary aim of this system is to reduce university professor workload by assisting in the planning of course and program outcomes. This frees up professors' time to focus on learning, exploring innovative teaching methods, and supporting better engagement with lecture topics. Leveraging generative AI's ability to simplify complex subjects into more easily understandable terms has the potential to improve student learning outcomes. However, to ensure accuracy and eliminate potential machine errors, thorough validation is important. This necessitates human oversight to verify the system's recommendations before implementation, ultimately discouraging fully autonomous use.

***Identification of Actors:*** The system is directly and indirectly in contact with multiple actors. Depending on the type, we grouped the actors into primary, secondary, and tertiary actors:

- Primary actors are in direct contact with the system during day-to-day business or directly affected by the system. This includes professors, students, and other technical staff who manage and assist with course planning and outcomes.
- Secondary actors are in contact with the system but do not use it in their workflow, or they are directly affected by its decisions. This includes education institute support staff.
- Tertiary actors potentially benefit from the system, even though they are neither working with the system nor are they directly affected by its decisions. We identified the tertiary actors may include multinational corporations and local companies hiring from this institute.

***Context and Processes, Where the AI System Is Used:*** The system is currently being used by three departments in D.Y. Patil College of Engineering (AK), Pune,

where generative AI is supporting teachers in creating clear and measurable course outcomes, helping them articulate specific learning objectives, further align them with program outcomes, and suggest appropriate assessment methods to measure students' achievement of these outcomes. The system has also proved helpful in developing various assessments, including quizzes, exams, and project rubrics. It can suggest diverse and effective ways to assess student understanding while ensuring alignment with course outcomes.

**Technology Used:** Generative AI, LLM, (ChatGPT (3.5/4), Claude, PaLM2)

- **Generative AI:** Generative AI is a type of artificial intelligence technology that can produce various types of content, including text, imagery, audio, and synthetic data.
- **LLM:** A generative model that provides context and memory capabilities, which are natural, human-like, and can hold interactive conversations.

**Human Oversight and Decision-Making in the Process Workflow:** An important decision made by stakeholders in this use case is to retain human oversight throughout the process. This decision is motivated by the inherent challenges of generative AI, including its potential to hallucinate or generate incomplete and inaccurate content. This necessitates the involvement of subject matter experts, like professors, who can evaluate and validate the suggestions produced by generative AI solutions, ultimately accepting or rejecting them based on their expertise.

## 10.7.2 ANALYSING THE SOCIO-TECHNICAL SCENARIOS FROM DIFFERENT VIEWPOINTS

We present a summary of the analysis of selected WGs. The analysis is conducted in parallel by the various WGs and, intentionally, we allow results with possible duplications and overlapping of the content. Later, during consolidation phase, we address the overlapping and duplications.

### **View of WG Education Professors/Teachers:**

**Autonomy/human oversight of AI:** The main goal of the system was to support (not replace) professors and teaching faculty in planning and creating clear and measurable course outcomes, helping them articulate specific learning objectives, to further align them with program outcomes, and suggest appropriate assessment methods to measure students' achievement. One of the major drawbacks, hallucination, is responsible for imparting incorrect information to the students, which can be dangerous because it aids in hampering trustworthiness if the content generated by ChatGPT remains unverified. Thus, the output of the AI system will not be used blindly; instead, the human actor will confirm the tool's suggestions, or use it as a helpful reference.

**Effect on Education:** The implementation of artificial intelligence in educational settings presents both benefits and challenges. A few of them are listed as follows:

- AI aids teachers and students in accessing knowledge from diverse sources outside the classroom, which helps students locate relevant resources swiftly and improves their learning experiences.
- By using AI, teachers can create customised learning plans that match each student's level and pace of understanding.
- In the field of education, AI can act as an equal opportunity provider, removing barriers of socio-economic status, geographic location, race, and ethnicity. It can enable learners regardless of their background or access to all educational opportunities.
- Education is not just about knowledge but also about developing social skills. AI systems that lack emotional intelligence cannot teach these social nuances, which are crucial for students' overall development.
- Using AI to develop customized learning plans for students involves collecting extensive student data, encompassing behaviour, academic progress, and personal details. If exposed, cyberattacks and data breaches could pose risks to students' security and privacy.
- A question: "How much use of AI is too much use of AI"? Will relying on AI-based technologies in the field of education lead to a potential decline in critical thinking and problem-solving abilities?
- An AI system has the potential to discriminate based on race and ethnicity. A few key factors that can lead to such discrimination can be biased data sets, algorithmic bias, and lack of social and cultural sensitivity. AI models fail to reflect the diversity of students served by the education system.

### Liability:

**Educational Institutions:** If a school or university fails to properly implement the AI system, or if they use it in a way that leads to harm (e.g., relying on it for critical decisions without human oversight), then the institution itself could be liable. This could also include failing to secure the AI system against cyber threats.

**Regulatory Oversight Committee:** A Regulatory Oversight Committee can be liable in situations where they fail to establish or enforce adequate guidelines or standards regarding the accuracy and appropriateness of educational content delivered by AI systems. For example, if the AI system begins to disseminate inaccurate and insensitive outputs due to being trained on biased or flawed datasets, this issue could have been mitigated by applying proper regulatory standards and thorough vetting of training data by Regulatory Oversight Committee.

**Teachers and Students:** Users of the AI system (e.g., teachers, students) might be liable if they use the system irresponsibly or against the guidelines provided by the Regulatory Committee.

Similarly, further assessment will explore the views of WG Education Students, WG Technical, etc., enriching our understanding and guiding future iterations of the use



case evaluation. As the assessment remains ongoing within the Assess phase, subsequent phases and processes will be concluded in due time.

### Participants:

- Trustworthy AI labs affiliated with Z-inspection, which employ its AI assessment framework.
- Members of the Z-inspection® initiative: Subject matter experts from diverse expertise areas are part of the Z-inspection initiative.
- Universities: Both Indian and foreign universities will participate in the assessment with their respective use cases.
- Others: Specialized agencies, motivated individuals, and subject matter experts in the academic domain.

### REFERENCES

- Ethics Guidelines for Trustworthy AI by European Commission High-Level Expert Group. (2019). <https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf>
- Hadhazy, A. (2023, July 17). *ChatGPT out-scores medical students on complex clinical care exam questions*. Stanford HAI. <https://hai.stanford.edu/news/chatgpt-out-scores-medical-students-complex-clinical-care-exam-questions>
- Members of the Z-Inspection® Initiative. <https://z-inspection.org/>
- Miao, F., & Holmes, W. (2021). *Artificial intelligence and education*. Guidance for Policy-Makers.
- Trustworthy AI Affiliated Labs. <https://z-inspection.org/affiliated-labs/>
- UNESCO Guidance for Policymakers on AI and Education. <https://unesdoc.unesco.org/ark:/48223/pf0000376709>
- Zicari, R. V., et al. (2021, June). Z-inspection®: A process to assess trustworthy AI. *IEEE Transactions on Technology and Society*, 2(2), 83–97. <https://doi.org/10.1109/TTS.2021.3066209>
- Z-Inspection® AI Assessment Framework. <https://z-inspection.org>

---

# 11 Actionable Ethics

## *From Philosophical Principles to Operational Initiatives for Responsible AI Projects in the Public Sector in the French Context*

*Anthéa Serafin, Lisa Fériol,  
and Bertrand Monthubert*

### **11.1 THE GRADUAL DEVELOPMENT OF A FRENCH DOCTRINE OF TRUSTED PUBLIC AI**

To ensure that integration of AI in the public sector is carried out in a uniformly responsible manner, the development and dissemination of a common doctrine regarding the ethics of its use is a prerequisite. Nevertheless, the development of such a doctrine is not easy in a context of a large diversity of public actors and where ethical competencies remain rare.

### **11.2 THE FOUNDATIONS OF A COMMON METHODOLOGY FOR PUBLIC ACTORS IN THE FIELD OF TRUSTED PUBLIC AI**

Although the legal framework for digital activities is under construction (e.g., at the European level, the Data Governance Act, the Data Act, the Digital Services Act, the Digital Market Act), with an approach oriented towards sanctions, some institutions are promoting a more flexible approach. For instance, in order to guide the conception and use of AI, the Conseil d'Etat recommends the development of guidelines for trusted AI in the public sphere, which would have two main functions: (1) to harmonise the definition of AI used in the public sphere and (2) to define a doctrine of AI design, deployment and use accompanied by a practical methodology. This flexible normative approach seems appropriate for several reasons. As mentioned in the introduction, the PRAI should shortly be adopted to provide certain AIS with legal guarantees. On the one hand, those guidelines should enable public actors to anticipate the entry into force of the PRAI by gradually adopting the right reflexes and identifying the skills they currently lack in order to comply with it

(in the aforementioned study, the Conseil d'Etat points out that a significant proportion of public AIS should fall into this category, particularly those used for the following purposes: regalian activities (police, justice, immigration), access and entitlement to public services and social benefits, education and vocational training, management and operation of public networks (water, electricity, etc.) or recruitment, assessment, promotion and termination of service of civil servants). On the other hand, given that the PRAI will not entail legal constraints for all AIS, guidelines appear to be complementary to it in order to ensure that the use of AI by the public actors is, in all cases, ethically considered. Finally, qualifying the level of risk associated with an AIS presupposes a methodology for assessing it, which makes the application of an ethical approach particularly relevant.

Regarding the first function of the guidelines, i.e., to harmonise the definition of AI used in the public sphere, it is recommended to refer to the definition that will be adopted in the future European regulation on AI, as it will be enriched by the standardisation work carried out by the normalisation organisms. Regarding the second function of the guidelines, i.e., to define a doctrine of design, deployment and use of AI as well as a practical methodology, the Conseil d'Etat notes several points that deserve to be highlighted. First, public actors should receive clear information mentioning that AI will not be the answer to everything, i.e., that an AIS is only a tool at their disposal, which may be unsuited to their needs. To determine whether it is indeed suitable or not, methodological benchmarks common to the entire public sphere should be defined concerning the following subjects: the decision to use AI or not, the AI approach to be favoured, the acceptable degree of outsourcing and the legal framework for the tool.

Regarding the decision to use AI to implement public policy, it would be appropriate to identify which AIS should be prohibited from deploying for ethical and political reasons, as well as those for which the purpose of use inherently generates significant risks. The Conseil d'Etat also specifies, regarding the use of automated decision-making, that the latter could, for instance, be preferred for making acceptance decisions that are not likely to harm third parties or seriously compromise a public interest, as well as for situations in which the administration does not have to make an assessment but has to apply specific rules on the basis of "objective" facts. Beyond the purpose of use, other criteria should be considered when deciding whether or not to use an AIS: the social acceptability of the project's purpose, applicable legal framework and relevant ethical principles at stake. At this point, a cost-benefit analysis should be carried out, incorporating the expected benefits as well as the proven disadvantages and foreseeable risks, while ensuring that the trade-offs made are documented (in the aforementioned study, the Conseil d'Etat specifies that the degree of depth of the exercise and the formal precision expected in the reporting of its results must obviously depend on the sensitivity of the activity and the data used, the maturity of the technologies used, and the scale and seriousness of the risks: the most complex will require an operational application of the principles and strengthened guarantees, including procedural guarantees). Regarding the ethical principles that may come into play, the Conseil d'Etat identifies seven general principles (in harmony with the requirements that the PRAI defines for high-risk AIS) to which public actors may refer when they face the question of the use of AI.

These seven principles are: human primacy, performance, equity and non-discrimination, transparency, safety, environmental sustainability and strategic autonomy.

1. The human primacy principle aims to ensure that public AIS are conceived as tools that meet a general interest objective and that the interference in fundamental rights and freedoms that results from their commissioning is not disproportionate to the benefits expected. In addition, a human being will always have to ensure that the AIS functions properly by supervising it and limiting its dependence on its use.
2. The principle of performance means identifying AIS performance indicators and defining the acceptable level of performance.
3. To guarantee fairness and non-discrimination, any designer of a public AIS should prevent discrimination (this issue being particularly important for decision support AIS based on machine learning trained on large datasets that are likely to contain biases). To achieve this, a risk management system should be put in place, AIS staff should be made aware of the issues involved, and the social representativeness of design teams should be sought.
4. The principle of transparency implies a right of access to AIS documentation, a requirement of loyalty consisting of informing people of the use of an AIS about them, the auditability of the AIS by the competent authorities and a guarantee of explicability (i.e., an explanation of the main reasons for the decision or recommendation made by the AIS).
5. The principle of security requires that potential computer attacks be anticipated and their consequences resolved.
6. The environmental impact of the AIS should also be considered in the public AI strategy, based on a principle of global neutrality.
7. Finally, as soon as AIS contributes to essential public authority functions, it should be designed in such a way as to minimise dependence on foreign technologies.

The Conseil d'Etat points out that these principles are not absolute and that they may come into tension with one another, calling for arbitration between them. Once the benefit-risk balance has been established, if it is decided to use an AIS, then the next step is to identify which method is best suited to the intended purpose: a deterministic system or algorithmic models learned from data? Once this choice has been made, a final methodological step is to decide on the acceptable degree of outsourcing, as well as the framework for intellectual property, data control and liability in the event of damage caused to third parties. Finally, it will be vital to ensure that persons affected by administrative decisions based in part on AIS have effective administrative and legal remedies to contest these decisions.

These guidelines for trusted public AI lay the foundations for a common strategy for public actors in this area. It is an innovative and essential contribution as it seeks to anchor the ethical approach in public decision-making processes, in particular by laying the foundations for the content of a charter for the use of AI. However, many public actors are still largely unfamiliar with this text and the proposed methodology. In order to encourage the drafting of such a charter and its uniform application across

France, a co-construction approach with all the public actors concerned seems necessary. This represents a major challenge in the context of the French administrative organisation, which is characterised by a wide variety of public actors in terms of their nature, their field of expertise and their level of digital maturity.

### 11.3 THE SPECIFIC CHALLENGES OF DEPLOYING AI IN THE FRENCH PUBLIC SECTOR

In order to fully grasp the challenges involved in deploying responsible AI in the public sector, it should be emphasised that the expression “public sector” is not confined to state actors, but covers a whole range of very different public actors, established at different levels of the territory and responding to different resource and political orientations. The deployment of responsible AI in the public sector thus requires this diversity of stakeholders to be brought on board, and the subject of digital transformation illustrates how difficult this can be. To quote the Conseil d’État,

From the point of view of data, the State is not one and the same. It forms an archipelago, with inter-island links that are often very inadequate. A fortiori, the existence of a public data community, including local and regional authorities, public establishments and other persons entrusted with a public service mission, is still a pipe dream.

(Conseil d’État, 2022, p. 169)

This issue is transposed and even reinforced on the subject of AI: despite the creation of dedicated institutions at the national level, the deployment of different data and AI plans (such as the France AI Plan following the France AI Strategy report (2017), the 2021 Programme d’Investissement d’Avenir (PIA), and the two versions of the National Strategy for AI (SNIA) following the Villani report “Giving meaning to Artificial Intelligence” written by mathematician and ex-Member of Parliament Cédric Villani) and the obligations imposed to open up public data and to share general interest data (France, 2016), many obstacles remain to accessing the massive amounts of high-quality data needed to develop AI projects. Thus, it explains why the subject of AI ethics is not currently high on the agenda of most public actors and why the culture on this topic is still weak or even non-existent. This significant lack of in-house skills forces the outsourcing of these skills and leads to situations of dependence on certain solutions and actors. Indeed, very few staff members have had any training or awareness of this topic, which does not represent a skill carried by agents in the teams developing AI projects, which does not, therefore, argue in favour of public actors adopting the methodology developed by the Conseil d’État described previously. However, it is nonetheless worth raising the awareness of these actors on the subject of data and AI ethics so that they are able to apply an appropriate approach in their future digital projects and use ethics as a driver. In reality, this subject requires a cross-disciplinary interest and awareness for each skill mobilised around an AIS project. However, it should be noted that we are observing movements in favour of recognition of skills in ethics.

Despite these challenges, public AI projects are being deployed at various levels of the territory. As the Conseil d’État notes in its study, “no area of public action is

impervious to these Systems” (Conseil d’État, 2022, p. 6). The study exposes different examples (Conseil d’État, 2022, p. 267) for which it is interesting to underline the inter-actor collaboration with notably private actors required for the deployment of such projects, as well as the different scales at which they are developed. Ensuring that national recommendations, guidelines and/or policies are applied by the wide variety of local public players is a real challenge. To do so, the central State, as well as State agencies (e.g., the French National Agency for Territorial Cohesion (Agence Nationale de la Cohésion des Territoires) and the French Bank for Territories (Banque des territoires)), provides support through the funding of these ones. There is therefore a crucial need for support and pooling on the complex subject of the ethical deployment of AI among public actors.

The necessary collaboration among all the actors involved in AI projects, and the societal dimension of the subject, require all stakeholders (including private actors and citizens) to be included in the co-construction of guidelines for trusted public AI in a democratic context. This requirement was emphasised by the Conseil d’Etat in order to ensure uniform acceptance and application. This analysis is in line with the results of the Cocacia survey carried out by Ekitia and ANITI (Artificial and Natural Intelligence Toulouse Institute). Its activities are based on three pillars: scientific research, training and contribution to economic development across the Occitanie Region, in which 70% of respondents (this survey is based on around 3,700 responses) expressed that they wanted their opinion to be taken into account, or at the very least expressed a desire to be consulted on the direction of AI projects likely to be deployed in their territory. Ethics have a role to play here. Ethics must enable dialogue among these different stakeholders by jointly establishing the principles that will be applied in the development of these systems that concern the whole of society. This is the purpose of the guidelines issued by the Conseil d’Etat, which enable the actors involved to make their own in the deployment of AIS. Thus, the effectiveness of the guidelines for a trusted public AI requires consideration of inclusive modes of governance representing these different stakeholders in order to jointly consider practical tools for the real implementation of an ethical approach.

#### **11.4 THE RELEVANCE OF ECOSYSTEMS ACTING AT A LOCAL LEVEL TO OPERATIONALISE THE ETHICS OF AI: THE EXAMPLE OF EKITIA**

The Conseil d’Etat emphasises that general ethical reflection should inspire operational ethical reflection and considers that this function should be structured in close proximity to the public actors concerned. Indeed, the lack of a practical, actionable ethical framework for data, as well as the fact that the gathering of different interests around a common ambition required the intermediation of a neutral third party, was the ground for the creation of Ekitia: a public-private ecosystem promoting ethical use of data and inclusive modes of governance at the territorial level, including citizens. Bringing together actors with similar issues within Ekitia has made it possible to define common rules, in this case the definition of ethical principles, as well as the development of tools, to operationalise them in their respective or joint activities.

### 11.5 THE EKITIA ECOSYSTEM, FACILITATING INCLUSIVE GOVERNANCE OF AI ETHICS

Although ethics may appear as a philosophical approach, far from the concrete reality of ground actors, some stakeholders have decided to act collaboratively at their own level. In this respect, Ekitia's ecosystem was born out of the coming together of public and private actors who have as a common objective the development of a responsible data economy. This ambition led to the co-construction of the Ethical Charter for Data Use (Ekitia, 2022), establishing the pillars of Ekitia's trust framework. This Charter is part of a process of continuous improvement (at least every three years) in line with technical, legal and societal developments, including considerations expressed by citizens. Review phases are discussed by Ekitia members, and in particular by the Ekitia Ethics Committee. The particularity of this Charter is that it focuses on the subject of data use, allowing it to be adapted to the various activities requiring data, thus encompassing AI. Promoting participatory democracy, Ekitia has from the outset sought to include citizens in the development of this Charter (as a first action to include them in its governance). The first citizen workshops were organised in early 2023, using design fiction scenarios to gather citizens' concerns and incorporate these into the new version of the Charter (the latest changes to the Charter will lead to version 3 of the Ekitia Ethical Charter for Data Usage).

Ekitia now brings together around sixty members from all sectors (e.g., academia, companies, start-ups, local communities, national public organisations). Ekitia has positioned itself as a real *think-and-do tank*: the ecosystem continuously improves the Ethical Charter, promotes best practices and monitors changes in the legal framework to operationalise it through various activities, including tools that will be developed further. Ekitia, in its current form (a not-for-profit association and soon to be a Public Interest Group (Groupement d'Intérêt Public), the latter enabling public and private partners to pool resources to carry out missions of general interest), acts as a 'neutral' third party allowing trust between the ecosystem actors. At the same time, Ekitia takes part in events where the ethics of data use is discussed with citizens. Methods for citizen participation have been developed, and Ekitia has recently joined MyData Global (an award-winning international non-profit that furthers the rights of individuals over their personal data, <https://mydata.org>) in order to discuss the integration of citizens in the use of their data ("self-data"). Ekitia is more largely involved in raising awareness and providing training for various audiences at various levels of territorial structuring (e.g., awareness-raising on data ethics within the Ministry of Digital Transition, at professional events). This think-and-do tank approach allows participants to translate the reflection conducted with the ecosystem members into the development of the ethical framework devoted to AI projects and tools related to make it actionable. The links with public research laboratories are an integral part of Ekitia's genesis. For instance, its Ethical Charter for Data Use (Ekitia, 2022) was drawn up in collaboration with the bioethics research team at CERPOP UMR1295 (Inserm and University of Toulouse).

The collaborative aspect, which is the very essence of Ekitia's ecosystem, is directly reflected in Ekitia's internal governance structure, which includes its members' representatives in the decision-making process and enables them to take part in the various activities carried out. Ekitia also promotes and offers its members

and partners inclusive and collaborative modes of governance for their projects, in line with the movement supported by the EU, which aims to reduce the asymmetries of power relations that exist today in the digital field. These innovative governance models were actually developed by Ekitia (in particular, a methodology for drawing up governance rules for a territorial data space pilot project is currently being developed). The ethical requirements arising from Ekitia's Ethical Charter for Data Usage have been placed at the centre of the development of these rules of governance. This method is inspired by the Rulebook for a fair data economy developed by the Finnish association Sitra (Sitra, 2025, February 6) and encourages the development of AIS that respect, and are in line, with the ethical requirements of Ekitia's Ethical Charter for Data Use, itself in line with national and EU requirements on the subject. These models also enable stakeholders to align themselves with the same values, creating a climate of trust that allows fruitful collaboration, particularly between the public and private sectors. The inclusion of these different stakeholders in the process of drawing up the applicable standards also helps them to make those standards their own.

Ekitia's strength lies in the "bottom-up" and inclusive approach applied to define the relevant ethical principles regarding data and AI. Indeed, this co-construction method makes possible relay between the local and specific needs of public actors and the requirements imposed at the national, European or even international level.

## **11.6 OPERATIONAL TOOLS FOR APPLYING AN ETHICAL APPROACH TO PUBLIC AI PROJECTS**

As we just saw, the inclusive governance implemented by Ekitia aims to enable its members, but also citizens, to play an active part in defining the ethical principles that will make it possible to question the use of data and AI to meet a need. However, although the Charter's principles have been co-constructed with the several stakeholders, it has become clear that it is still difficult for them to make those principles their own and apply them in the context of their concrete digital projects. This means that, in addition to the general principles, methods and tools were necessary to disseminate and facilitate understanding of the content of the ethical principles as well as to help stakeholders take ownership of the ethical approach.

Apart from classic tools of dissemination (e.g., webinars, workshops), it has been decided to facilitate the application of the Ethical Charter through some innovative methods and tools. Initially, Ekitia has developed an ethical support for digital projects by design, using the Ethical Charter for Data Usage as a grid for analysing the ethical risks, real or potential, raised by a project. Such an analysis gives rise to ethical recommendations for a project owner. As Ekitia conceived them, such recommendations are not binding and aim above all to promote the ethical approach close to project owners, to help them identify the applicable legal framework and the ethical issues specific to their project, as well as to identify ways of responding to them. The project owners then remain free to decide which recommendations they will apply when implementing their project. To take things a step further, Ekitia also proposes to interested project owners the opportunity to co-build with Ekitia to draw up a specific charter setting out the ethical principles to be applied in implementing their project.



Two projects illustrate this approach, involving the development of AIS in the public sector, which have been the subject of ethical support by design by Ekitia: one in public health and a second in employment and training.

**11.7 ETHICAL SUPPORT FOR THE DEVELOPMENT OF AN AIS WITHIN THE PUBLIC HEALTH SECTOR**

A small French company developing decision-support tools for public health decision-makers had designed the following project: create a regional map of areas at risk of a resurgence of the Covid-19 epidemic in order to support decisions by institutional health actors to adopt preventive measures. The project was therefore part of a public health prevention approach aimed at helping institutional healthcare public actors to adopt measures to prevent the resurgence of the epidemic, using an innovative decision-making tool to tackle a health issue from a global perspective (the regional mapping of at-risk areas was to be the result of cross-referencing

**TABLE 11.1**  
**Recommendations Related to Main Ethical Risks for a Public Health AIS**

Main Ethical Risks	Corresponding Charter Principle	Related Recommendations
Stigmatisation of people living in high-risk areas and re-identification of infected people	Solidarity, diversity and non-discrimination	Ensure that the chosen spatio-temporal resolution does not allow people to be re-identified in sparsely populated areas, for instance, by establishing a threshold relating to the minimum number of people who must appear in each zone.
Misunderstanding of the tool by decision-makers	Clear, accessible information	To be able to provide clear and accessible information about the place that the decision-support tool will occupy in the decision-making process of institutional healthcare actors (in particular, the fact that the tool is only a decision-making aid and that the decision remains entirely in the hands of human beings).
Reuse of project results for purposes other than those initially planned	Governance within a framework of trust	Specify upstream who will own the tool, who will be able to use it and for how long, and lay down clear rules on liability.
Relevance and accuracy of initial data	Data quality	Correct errors of representativeness and other biases in the data, check the accuracy and relevance of the data in relation to its intended use and specify the duration of this relevance.

**TABLE 11.2**  
**Recommendations Related to Main Ethical Risks for a Public Orientation AIS**

Main Ethical Risks	Corresponding Charter Principle	Related Recommendations
Tool 1: Aligning training provision with the skills demanded in the labour market, likely to cause a gradual disappearance of more atypical or innovative training provision over the long term	Beneficence	Inform and train users appropriately about the capabilities and limitations of the tool, so as to preserve a range of atypical and innovative training courses.
Tool 2: Strengthening the digital divide	Solidarity, diversity and non-discrimination	Make the tool known to as many citizens as possible.
Tool 2: Gradual disappearance of guidance counsellors	Human factor	Help guidance counsellors come to grips with the tool.
Tool 2: Access by foreign authorities to personal data provided by users	Respect for privacy	Inform users of the location of the personal data they provide when using the tool.

social data, environmental data and health data provided by a variety of public and private actors). It should be noted that the tool developed was intended simply to support the decision-making process, which remained entirely in the hands of the people involved.

**11.8 ETHICAL SUPPORT FOR THE DEVELOPMENT OF AIS  
IN THE EMPLOYMENT AND TRAINING SECTOR**

Such support was also requested by the Occitanie Region, concerning AIS to be developed to ensure better predictive management of the initial and continuing vocational training needs of jobseekers in Occitanie. More specifically, two tools were the subject of recommendations: (1) the “Matching Employment and Training” tool: a decision support tool designed to effectively guide the regional agents of the Employment and Training Department in their decisions to order training courses, so that these are adapted to the skills required on the labour market; and (2) the “Personalized Employment Paths” tool: a tool for the general public, accessible via a website, designed to advise citizens seeking employment and/or training by providing them with personalised guidance based on their professional, personal and social skills, as well as their constraints and desires.

As the personalised ethical support of digital projects work progressed, Ekitia realised that while some recommendations were indeed specific to the projects analysed, others could be considered more generic. This led to the parallel creation of a “generic” evaluation grid, enabling an overall assessment of the ethics of projects involving data processing. This grid now forms the basis of a label designed to

**TABLE 11.3**  
**Examples of Criteria of the Ekitia Label**

Examples of Ekitia Label Criteria	Corresponding Charter Principles
Do you plan to allow all or part of the results of your project to be re-used by research bodies or for public service purposes?	Beneficence
Are you processing only the data necessary for the project?	Sustainable innovation
Have you assessed and minimised the discriminatory biases affecting or likely to affect the data and algorithmic models used in your project?	Solidarity, diversity and non-discrimination
Will users of your solution be able to contact a human easily to ask for explanations about how it works?	Human factor
In the event that your solution could be misused for disinformation purposes, are you applying measures to minimise those risks or to deal with their consequences?	Respecting and strengthening human autonomy
If the results of your project include personal data and you want to allow third parties to re-use them, have you thought about setting out the conditions for this re-use?	Respect for privacy
Do the data used to carry out your project constitute references in your sector (in terms of the reliability of its source, its format and/ or its structuring in accordance with standards)?	Data quality
Are the information security measures applied as part of your project appropriate to the sensitivity and confidentiality of the data used?	Information system security
With regard to the algorithms used to carry out your project, have you defined a threshold below which the reliability of the results obtained would not be satisfactory?	Robust algorithms
Do you provide a user guide (or tutorial) for your solution that is easily accessible to users?	Clear and accessible information
In the case of algorithmic models built by learning from data, do you justify and document the choice of model, the algorithm used and its operating logic?	Explainability of algorithms
Will certain elements (e.g., data, tools, services, algorithms, etc.) linked to your project be freely reusable by others?	Collective learning
Have you considered the risks to people’s health posed by your project?	Risk assessment
Have the end users of the solution developed as part of your project been included in its design and implementation?	Inclusion of end users
If the data used in your project is protected by copyright or other forms of intellectual property, have you obtained the explicit agreement of the copyright holders before using it?	Integrity
Does the economic and social impact of the project have a fair impact on the target audience?	Fair distribution of value creation

enhance the market value of products and services designed in compliance with an ethical approach: the Ekitia Label. It contains around a hundred assessment criteria, based on the main themes of the Ethical Charter for Data Usage. These criteria relate to the various stages of data processing (e.g., collection, storage, analysis, purposes of use, sharing, etc.) carried out in a project. An extract of these criteria is presented in Table 11.3.

To date, five of the six project owners who have taken this step have obtained the Ekitia Label, including two public-sector actors for solutions developed in-house, and two private-sector actors for solutions that could be useful to public-sector actors. Ekitia's discussions with the project owners ethically supported show that this service has helped to develop their ethical culture. In fact, identifying risks, categorising them into different themes and seeking to minimise them at the design stage of projects helps to anchor the ethical approach in their activities. As a result, when designing new projects, they are more likely to question the potential positive and negative impacts, and they are able to develop their own methodology for analysing and minimising risks. As for Ekitia's Label, by promoting digital solutions designed and developed in compliance with an ethical approach on the market, it helps citizens identify which solutions are compliant with the Ethical Charter. It also enables public-sector actors to better identify trustworthy solutions and, as such, contributes to the development of responsible AI in this sphere.

## 11.9 CONCLUSION

Ekitia's experience reveals the need for the creation of close-to-the-ground collaborative organizations that deploy tools that make ethics actionable in the context of AI and data use. Ekitia's action makes it possible to address specific issues – such as those specific to the public sector, and more specifically to the subject of AI – while integrating them into a global approach to the subject of the ethics of data and AI, thereby promoting collaboration and mutualisation among the various stakeholders. There is a space between global organisations with their general considerations about regulation and local organizations, especially public ones, that have to apply the general principles without having the resources to do it.

The interest shown by actors in joining the ecosystem, which now has nearly 70 members, clearly demonstrates the need and demand for awareness-raising on the subject of ethics and for tools to put these reflections into practice. These subjects demand expertise that is usually lacking, in order to be able to have an ongoing ethical reflexivity that ensures the relevance of the tools proposed.

This experience also shows the importance of a flexible normative approach, which complements stronger normative approaches like acts, since it helps organizations to weigh up the advantages and disadvantages of specific AI projects with respect to ethical considerations. This experience has also revealed some limitations: the current economic climate sometimes wrongly leads public-sector actors deploying AI systems – an observation that applies generally to data actors – to relegate the subject of ethics to the background, considering that higher-priority actions

responding to budgetary efficiency logics should be given priority, even though the ethical approach makes it possible to structure projects from the design stage while involving the various stakeholders by inviting them to reflect and collaborate together. Efforts must continue so that the Conseil d'Etat guidelines can be genuinely adopted by the public-sector actors concerned.

## REFERENCES

- ANITI and Ekitia. (2023). *Cocacia Survey on knowledge, acceptability and ethical issues related to AI*. <https://aniti.univ-toulouse.fr/en/2023/11/23/consultation-citoyenne-sur-lia-les-resultats-sont-disponibles/>
- Conseil d'État. (2022). *Artificial intelligence and public action: Building trust, serving performance* (pp. 6, 169, 267). Conseil d'État.
- Council of Europe Committee on AI. (2023). *Draft convention on AI, fundamental rights, democracy and the rule of law*. Council of Europe.
- Ekitia. (2022). *Ekitia's ethical charter for data usage* [PDF]. [https://www.ekitia.fr/wp-content/uploads/2022/04/Ekitias-Ethical-Charter-for-Data-Usage\\_03\\_22.pdf](https://www.ekitia.fr/wp-content/uploads/2022/04/Ekitias-Ethical-Charter-for-Data-Usage_03_22.pdf)
- European Commission. (2021). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on AI, COM(2021) 206*. European Commission.
- European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). *Official Journal of the European Union*, 119, 1–88.
- France. (1978). Law No. 78–17 of 6 January 1978 on information technology, data files, and civil liberties (as amended up to 1 June 2019).
- France. (2016). Law No. 2016–1321 of 7 October 2016 for a Digital Republic.
- High-Level Expert Group on Artificial Intelligence. (2019). *Ethical guidelines for trustworthy AI*. European Commission.
- Sitra Website. (2025, February 6). *Rulebook model for a fair data economy*. <https://www.sitra.fi/en/publications/rulebook-for-a-fair-data-economy/>
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2005). *Universal declaration on bioethics and human rights*. UNESCO.

---

# 12 Supporting AI at Scale in the APEC Region through International Standards

*Aurelie Jacquet, Karen Batt, and Jesse Riddell*

## 12.1 INTRODUCTION

Recent developments in artificial intelligence (AI) and automation have put us on the cusp of a new automation age. Over the past few years, the advances in AI technology have led to the emergence of machines that can accomplish cognitive capabilities once considered too difficult to automate successfully through combining large data sets with intuitive processing algorithms (Edlich et al., 2019, p. 1). This has resulted in AI being taken up by greater numbers of individuals, businesses, and governments to support increased efficiency and productivity across sectors. This embrace of AI technologies is expected to continue and intensify, with estimates that AI will add USD15 trillion to the global economy by 2030 and boost global GDP by 14% (Edlich et al., 2019, p. 1).

Artificial intelligence (AI) holds immense potential to drive significant economic, social and environmental benefits in all Asia-Pacific Economic Cooperation (APEC) economies. By enhancing decision-making processes, improving operational efficiency, and boosting productivity, AI can unlock innovation and create and expand markets and revenue streams. Indeed, the adoption of AI is already having an impact in APEC economies across every industry – from healthcare to manufacturing to technology.

To put the opportunity in perspective, a forecast by PWC predicts that AI is projected to contribute as much as \$22.17 trillion to the global economy by 2030, and the APEC economies are well positioned to share in these benefits. For the transformative and economic benefits to be realised, however, the right settings need to be in place to support the responsible scaling of AI in the region.

New and emerging AI technologies are transforming business practice and challenging established approaches to public policy and regulation. The development of unclear or unharmonised regulatory approaches that are not interoperable and the need for organisations to adapt to this new landscape without proper guidance risks blocking the realisation of the opportunities presented by responsible AI, and risks undermining public trust and confidence in emerging technologies.

AI has enormous potential to advance economic and societal well-being and enable improved environmental outcomes within the APEC region. AI is already

driving innovation and efficiencies and is supporting the creation of unprecedented new products, systems, and services across the region, from automated health diagnostics in hospitals to smart agriculture and precision farming systems that are optimising yields at the farmgate. It offers vastly improved decision-making and cost reduction, enabling businesses and policymakers to boost productivity and speed, scale, and consistency of service.

When done right, AI has proven to deliver real benefits, including:

1. **Automation:** AI can automate repetitive tasks, driving productivity efficiency and freeing up human resources to focus on more complex and creative tasks.
2. **Decision making:** AI can analyse large amounts of data and provide insights and recommendations, supporting more informed decision-making and recommendations.
3. **Improved accuracy:** AI algorithms can perform tasks with a high degree of accuracy and precision, reducing errors and improving overall quality.
4. **Safety and Security:** AI can be used for threat detection and to support surveillance and cybersecurity by helping to identify and prevent risks and threats.
5. **Efficiency:** AI can help in analysing and predicting outcomes, leading to more accurate and timely interventions. This has proven to be particularly beneficial in healthcare.
6. **Innovation and creativity:** AI can assist in creating new ideas, designs, and solutions by analysing vast amounts of data and identifying patterns and trends.
7. **Enhanced personalisation:** AI can analyse user data and behaviour to provide personalised experiences and recommendations and can support customer service by providing instant customer support.

It is of no surprise that there is excitement surrounding the opportunities that AI presents to unlock transformative economic, societal, and environmental benefits in the APEC region. A recent Microsoft – IDC Study (Cuyegkeng & Evans, 2022, p. 4) found that almost all businesses believe that AI is central to their growth, with 80% of business leaders in the Asia-Pacific region reporting that it is instrumental to their organisation's competitiveness. The same study found that the businesses surveyed believe that AI will almost double the rate of innovation in the short term.

Success is not guaranteed, however. A recent ABAC report, "Artificial Intelligence in APEC: Progress, Preparedness, and Priorities", found that the APEC economies are not optimally prepared to take advantage of AI. The report found that, while AI presents a paradigm-shifting opportunity for APEC, risks and blockages to uptake of AI could be just as significant, ranging from ethical considerations to a lack of preparation required to take advantage of the coming revolution (Cuyegkeng & Evans, 2022, p. 4). A recent McKinsey study found that of a sample of institutions that have adopted AI, only 55% of institutions believe their automation program has been successful to date (Cuyegkeng & Evans, 2022, p. 4).

These findings highlight that, even as we see the social and economic potential of AI, several risks and complexities impact uptake in both the private and public sectors across the APEC region. Barriers to operationalising and scaling AI include the risk of poor data quality and biases in AI systems, data privacy and security considerations, lack of skilled personnel and knowledge to develop, implement, and maintain the technology, and ethical and regulatory considerations. There are also cross-border challenges, as varying policies and regulation across sectors and jurisdictions, including on data privacy and security, can act as barriers to trade of AI solutions. In addition, there is the issue that public and private trust in AI remains low. This is exemplified by a 2023 KPMG and University of Queensland global study that found that three out of five people surveyed (61%) are wary about trusting AI systems, reporting either ambivalence or an unwillingness to trust AI technology (Cuyegkeng & Evans, 2022, p. 4).

To realise the opportunities afforded by AI in the APEC region, a comprehensive response to these challenges through policies and institutional frameworks that guide responsible AI design and use is necessary to ensure that AI benefits society as a whole. Over the past few years, a number of domestic and international policies, principles, and guidelines have been developed that aim to ensure that AI systems are designed to be robust, safe, fair, and trustworthy. More recently, governments have begun to develop regulatory settings for AI to promote the same objectives (Arai & Law, 2023, p. 6). These efforts play a critical role in supporting the responsible development and development of AI. However, given the number of different policies and approaches being developed across the region, they risk creating confusion for business and fragmenting the market if they are not underpinned by internationally agreed standards.

A fundamental element in shaping the responsible design, development, and scaling of AI is the establishment and adoption of International Standards. International Standards for AI play an important role in establishing specifications, frameworks, and requirements upon which AI technologies can be built, tested, and deployed. They support business by setting globally agreed-upon principles and processes for AI technologies that ensure consistency in the development, deployment, and use of AI. They also support trust and confidence in AI products and services by providing assurance of safety and reliability to users and consumers. Additionally, they facilitate interoperability of products and services across borders, supporting trade, innovation, and competition.

For the government, international AI standards can support establishing AI policies and regulations. New and emerging AI technologies are challenging established approaches to public policy and regulation. International Standards present agile and fit-for-purpose globally developed solutions that cover topics from ethical and responsible development and use of AI to data management and use, as well as key topics such as trustworthiness, privacy, and cyber security (Arai & Law, 2023, p. 6). The development of regulatory and policy frameworks in APEC that are underpinned by International Standards and leverage standards as a means of demonstrating conformance will promote harmonisation in approaches across sectors. It will also support alignment and regulatory compatibility across different economies, helping to avoid fragmentation, conflicting regulation, and cost and red tape impediments for business.



In an increasingly complex ecosystem, standards are vital to removing obstacles and addressing complexities that impede the scaling of AI. They provide frameworks for managing data quality across the AI lifecycle, as well as for managing privacy, cybersecurity, bias, and explainability, and provide tools for oversight of AI systems and the management of risk. It is our hope that this chapter and the mapping of AI standards that is included within it will serve to increase industry and government awareness and implementation of AI standards. In addition, we hope that, through achieving this objective, the chapter will support the responsible development and adoption of AI in both the public and the private sectors in the APEC region.

## 12.2 WHAT IS AI AT SCALE?

AI is no longer exclusively for Big Tech companies. The recent example of the rapid development of RNA-based COVID-19 vaccines, and the key role that AI technology played in supporting this innovation, showcases AI's world-transforming power. Although mRNA vaccines were not new, the use of AI in their development proved to be a game-changer in helping multiple companies to identify potential molecular targets on the COVID-19 virus where vaccines might act. AI also helped optimize for vaccine efficacy and ease of manufacture in the development process. Once vaccines were developed, AI also helped by predicting the spread of the virus to support efficient testing and distribution. The case of mRNA vaccines is a success story of collaboration between government and industry that shows the world-transforming power of AI when used at scale.

For business leaders and policymakers, AI at scale refers to how deeply and widely AI is integrated into an organisation's core products or services and business process. To reap the transformative benefits of AI, the technology needs to be scaled. AI is most valuable when it is a tool that organisations and governments use as part of their day-to-day business to deliver quality AI-powered products and services.

Unfortunately, scaling AI in this way is not easy. Although AI is embedding into the products and processes of virtually every industry, organisations and governments are still struggling to scale AI to reach its full potential. A recent McKinsey report found that while the business world is beginning to harness AI technologies and their benefits, fundamental transformation barriers remain, as adoption entails multiple, continuous, and simultaneous adjustments of an organisation's resources, culture, and decision-making. Similarly, a Deloitte report on "Scaling AI in Government" found that AI maturity is a challenge for government organisations due to technical limitations and governance challenges that limit large-scale adoption of AI platforms that vary in scope and complexity and because adoption often gets stalled at the pilot stage.

Although getting one or two models into production can be achievable, deploying AI across an entire enterprise or product often requires enterprise-wide digital transformation and brings significant complexity. A further challenge is that as AI is scaled, the risks associated with its use also increase. There are numerous examples of data privacy and security breaches, biased data perpetuating discrimination, and a lack of transparency resulting in problematic outcomes from AI models. To mitigate these risks, and to build public and private trust in AI, organisations must

adopt responsible AI practices, including robust AI and data governance to ensure trustworthiness, accountability, risk management, and transparency.

## 12.3 BARRIERS TO OPERATIONALISING AND SCALING AI IN THE APEC REGION

The ABAC report “Artificial Intelligence in APEC: Progress, Preparedness, and Priorities” finds that APEC is not optimally prepared to take advantage of AI, as a number of barriers impact member economies’ ability to operationalise and scale AI (Cuyegkeng & Evans, 2022, p. 4). To gain a greater understanding of these barriers, Standards Australia – with support of the APEC SCSC secretariat – undertook an APEC-wide survey on *Supporting AI at Scale in the APEC Region Through International Standards* to evaluate the level of preparedness in APEC for AI at scale.

The survey was well supported with over 70 responses from 14 member economies. On the question of what are the key challenges that are limiting AI at scale in APEC economies (where multiple answers were allowable), the survey showed a number of significant challenges that were grouped around key themes. These include a lack of access to or awareness of guidance and standards that support responsible development and deployment of AI (59%), a lack of trust in AI systems (48%), lack of skilled personnel to develop, implement, and maintain the technology (48%), and ethical concerns related to privacy, security, bias, or accountability (48%). See Figure 12.1 for the full results.

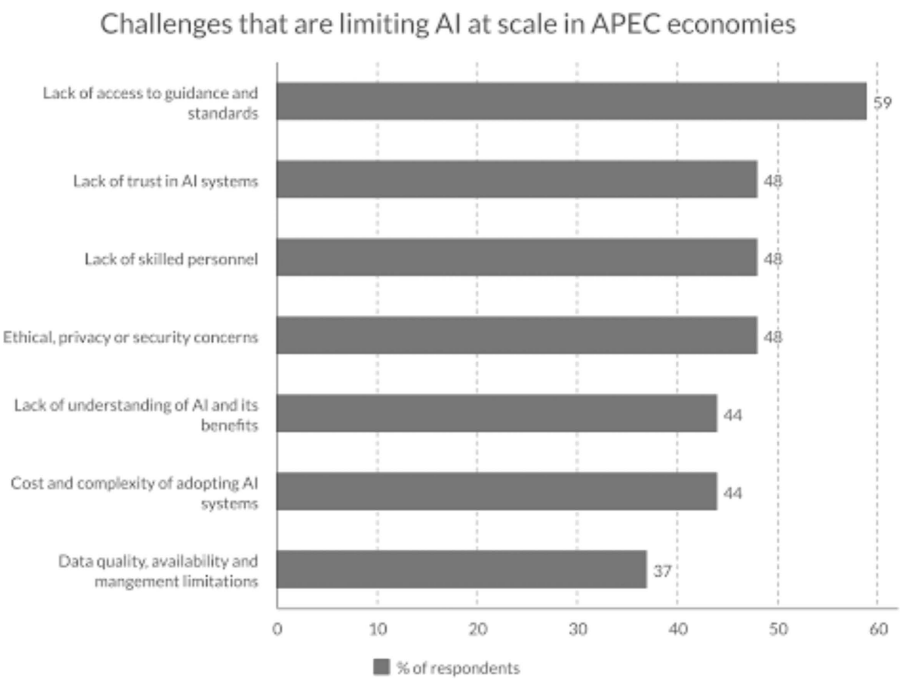
These survey results are unlikely to be surprising. Numerous reports and articles have been written about the risks and challenges that AI poses and the impact that they have on the uptake of AI systems. The concerns have been fueled by high-profile cases of AI use that was biased, discriminatory, or unlawful. There are many examples of AI being used for potentially harmful purposes, such as:

- Perpetuating and amplifying existing bias in the data they are trained on
- Creating fake content and misinformation (Satariano & Mozur, 2023, p. 9)
- Generating deepfakes for harmful or deceitful purposes (Hiebert, 2022, p. 10)

Realising the benefits that AI offers at scale and the return on investment in AI technologies requires responding to the risks and challenges that threaten responsible deployment of AI and impact the public’s trust in AI solutions (Gillespie et al., 2023, p. 5). Sustained scaled AI in the APEC economies is reliant on addressing the key challenges identified in the survey results above and expanded upon in Table 12.1.

## 12.4 WHAT ARE AI INTERNATIONAL STANDARDS?

Nearly everything we touch and interact with is designed and developed in accordance with International Standards. For AI, International Standards are voluntary documents that set out specifications, procedures, and guidelines that aim to ensure AI products, services, and systems are safe, consistent, and reliable. They are established by a consensus of subject experts and approved by a recognised standards body.



**FIGURE 12.1** Challenges that are limiting at scale in APEC economies.

Source: Supporting AI at Scale in the APEC Region Through International Standards Survey.

**TABLE 12.1**  
**Challenges Identified**

<b>Lack of Awareness and Trust</b>	<p>The lack of awareness and understanding of AI is a significant issue that impacts its uptake across industries. Many businesses, particularly small and medium-sized enterprises (SMEs), and policymakers do not comprehend the potential benefits and applications of AI. This is also the case for the public, where a lack of understanding can lead to skepticism and fear. Misconceptions, as well as incidents of misuse and the fear that AI will create job loss, can create resistance to adoption and acceptance of AI technologies, with recent studies showing that most people are wary about trusting AI systems and have low or moderate acceptance of AI (Gillespie et al., 2023, p. 5).</p> <p>This lack of awareness and understanding of AI is highlighted by the survey results that show that there is a perception across APEC that a lack of access to guidance and standards, including a lack of consensus on guidance, that support responsible development and deployment of AI is the key challenge limiting AI at scale in APEC economies.</p>
------------------------------------	--

**Table 12.1 (Continued)**  
**Challenges Identified**

<b>Data Quality and Availability</b>	<p>AI systems require high-quality, relevant data to function effectively. However, organisations often struggle with data quality and managing data throughout the AI system lifecycle, as data sets can be incomplete, not representative nor balanced, or outdated. Poor data quality results in unreliable or inappropriate AI models with inaccurate or biased outputs. Additionally, data silos and poor data management and governance within organisations can limit the availability of relevant and complete datasets that are necessary to enable AI models. According to Gartner, 85% of data-driven projects (like AI and IoT) fail to move past preliminary stages, citing the lack of suitable data as a big factor (Gartner, 2018, p. 17).</p> <p>To respond to this issue, organisations must effectively manage data to enhance performance and reliability. Comprehensive data management systems are necessary for AI deployment as they play a crucial role in ensuring the quality, accessibility, and reliability of data used by AI systems.</p>
<b>Privacy and Security</b>	<p>AI systems require rich, large, and quality datasets to allow AI systems to be designed, tested, and improved. These datasets often include sensitive and personal information. There is the potential for individuals’ data to be used in ways that raise privacy and security concerns.</p> <p>The fear of compromising data privacy and security can lead to hesitancy in adopting and using AI technologies. Organisations must have in place robust data privacy and security measures to build and maintain responsible AI systems and to ensure that data is protected from unauthorised access, theft, or misuse and maintain customer trust.</p>
<b>Safety, Legal, and Ethical Concerns Related to Bias, Fairness, and Accountability</b>	<p>Inaccuracies from AI models can result in misleading or erroneous outputs that raise safety, legal, and ethical concerns. There have been several high-profile cases of unreliable or inaccurate AI systems creating safety risks. AI automated decision-making increases the risk of automating unwanted bias and inequalities, and risks a lack of fairness, accountability, and transparency.</p> <p>Algorithmic bias, which is the systematic or repeated decisions that privilege one group over another, is often seen as one of the biggest risks of AI. Bias can result from datasets that are not comprehensive and from flawed model design or interpretation. There have been numerous recent high-profile cases of discrimination against individuals based on race or sex.</p> <p>Another ethical risk is that of system accountability and transparency. This is the question of validity and whether the reliability of data used to train models is appropriate for their intended purpose. Transparency and accountability is important for the AI market as it allows validation and trustworthiness of an AI model. To ensure responsible AI development and deployment, AI systems must be designed, tested, and validated to mitigate for unwanted bias and to ensure accountability.</p>
<b>Lack of Skilled Personnel</b>	<p>AI requires a highly skilled workforce to develop, implement, and maintain the technology. However, there is a shortage of AI experts and data scientists, which can limit the ability of organisations to operationalise and scale AI. An EY study found that 31% of US CEOs and business leaders believe a lack of skilled personnel is the greatest barrier to AI implementation (EY Study, n.d., p. 11).</p> <p>In order to overcome barriers to maximising AI implementation, the same report found that the most important factors for responding to this were having a clear organisational strategic vision and commitment to AI that is driven by senior leadership.</p>

Up to 80% of global trade is affected by standards or associated technical regulations. For this reason, the creation and use of consistent standards, through the input of both the private sector and governments, is fundamental for the medium to long-term sustainable development of the global digital economy, including in relation to AI. The strength of the international standard system is that International Standards are developed by technical experts from businesses, governments, academia, and consumer groups, from all interested economies across the world. International Standards for AI represent truly international solutions, and when they are adopted by businesses and regulators, they promote harmonisation and interoperability in processes for products, services, and systems. They also support market access and help lower barriers to trade, promote convergence in regulation, provide a shared launch pad for innovation, and help manage security risks.

International Standards play a crucial role in supporting responsible behaviour in AI development and deployment, whether through voluntary use that can support organisations to reduce risks and utilise global best practice, or as mandatory requirements when called up in regulation or in contractual agreements.

## 12.5 INTERNATIONAL STANDARDS SUPPORT AI AT SCALE

International Standards can play a constructive role in scaling the widespread use of responsible AI in the APEC region. In the rapidly evolving ecosystem of AI technology, standards establish common building blocks for companies and policymakers, and the risk management frameworks that manage risks to individuals, organisations, and society associated with AI. They can also provide globally recognised frameworks for data quality, trustworthiness, privacy, security, and ethics that AI systems can be designed, tested, and validated against.

International Standards play a critical role in creating frameworks that set the specifications and requirements upon which new technologies can be developed, adopted, and safely deployed. They provide a level playing field for AI developers and users, enabling them to build upon existing technologies and existing best practice. They also provide internationally agreed-upon principles and processes that allow consistency of products and services across borders, in doing so promoting market access, competition, and innovation.

Critically, standards also can act to mitigate the risks and address the ethical concerns that are the key challenges impacting the scale of AI in the APEC region. They support consumer and developer trust in AI by providing confidence that systems are safe, reliable, and fit-for-purpose. They also provide fundamental frameworks for benchmarking and auditing systems and organisations, offering a means for conformity assessment for AI systems entering the market. In addition, standards provide comprehensive baselines that support building reliability, ethicality, and transparency in AI systems.

Standards can play a critical role in shaping the operationalisation and deployment of new AI technologies in the form of governance standards (often targeted at

Board Directors and senior executives) that provide a framework for organisations to navigate the complex landscape of AI and address challenges and concerns, management systems standards that can include specific risk management frameworks, and controls for use by organisations and technical standards that define technical aspects of AI systems, including their design, interoperability, performance, and security (Australia Standards, 2022, p. 12).

## 12.6 INTERNATIONAL STANDARDS CAN SUPPORT REGULATORY HARMONISATION

The need for standards in the AI landscape has become increasingly evident considering the growing instances of AI policy and regulation across the globe. Recent years have witnessed a flurry of principles, guidelines, roadmaps, and regulations developed unilaterally and through international bodies on AI. This activity has been triggered by emerging concerns about AI ethics, security, and privacy and to promote the uptake of responsible AI in economies to achieve desired societal and economic outcomes.

In the APEC region, AI policies and regulation are gaining significant attention as economies recognise the potential benefits and risks associated with AI. According to the OECD.AI Policy Observatory, as AI has gained increased attention globally, the number of policies, strategies, and frameworks on AI in the APEC region has increased to 66 separate strategies (Australia Standards, 2022, p. 12). In addition, it is well publicised that a number of APEC economies are considering specific regulations on AI technologies aiming to address bias, discrimination, and privacy violation risks.

These efforts are critical to supporting the responsible development and deployment of AI. However, they risk creating confusion for business and fragmenting the market if they are not underpinned by common architecture. Uncoordinated unilateral measures raise costs of digital service trade, including for AI. A recent study (Coghi & Jelitto, 2023, p. 14) by the Organisation for Economic Co-operation and Development (OECD) and the World Trade Organization (WTO) found that the G20 economies can achieve savings worth US \$150 billion in costs by implementing the principles in the WTO Reference Paper on Services Domestic Regulation (Coghi & Jelitto, 2023, p. 14). One of the key recommendations in the Reference Paper is the adoption of technical standards developed through open and transparent processes, including those in International Standards-setting bodies, in services regulation.

International Standards can support businesses when they are utilised as a basis for establishing common approaches to regulation or as a means to demonstrate conformance, as they reduce fragmentation and barriers to trade across borders. They also can play an important role in helping businesses and governments to implement AI principles, such as the OECD *Principles on Artificial Intelligence* and the principles that have been developed across the APEC region, often with similar content. Here, standards can provide more granular technical solutions and guidance that supports adherence to these principles within organisations.

12.7 SUPPORTING AI AT SCALE IN THE APEC REGION THROUGH INTERNATIONAL STANDARDS SURVEY RESULTS

The APEC-wide survey on *Supporting AI at Scale in the APEC Region Through International Standards* asked respondents to provide input on the key ways that standards support operationalising and scaling AI in their APEC economy (multiple answers were allowable). The survey found that the main areas where standards can support AI to scale in APEC include by supporting conformity assessments and regulatory compliance (67%), by supporting R&D and innovation (63%), by promoting safety and trust in the design and deployment of technologies (63%), by providing a basis for establishing common approaches to the regulation of AI (54%), and by embedding privacy and security in the design and deployment of technologies (54%).

See Figure 12.2 for the full results.

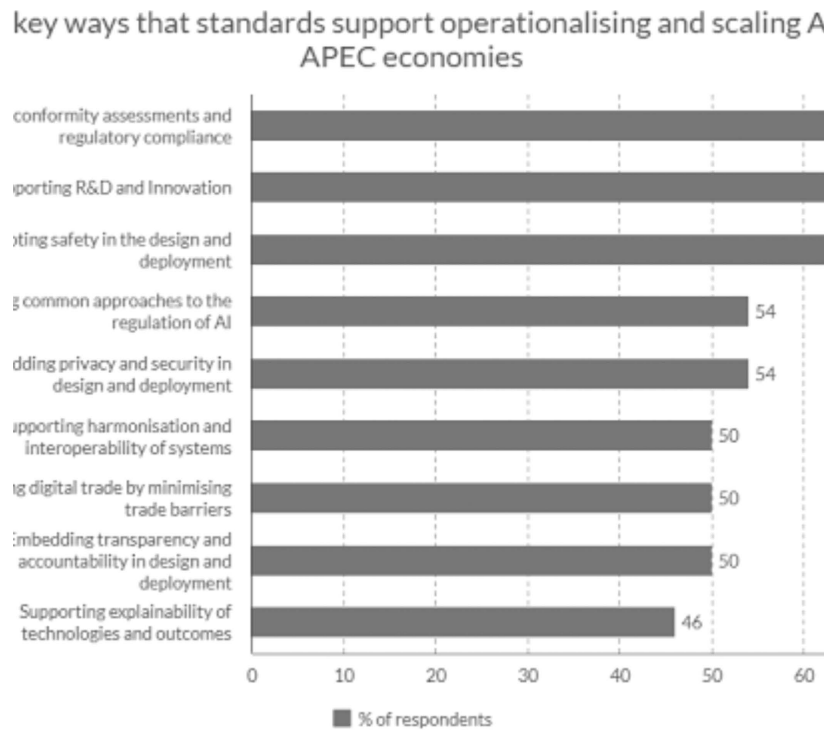


FIGURE 12.2 The key ways that standards support operationalising and scaling AI in APEC economies.

Source: Supporting AI at Scale in the APEC Region Through International Standards Survey.

12.8 THE INTERNATIONAL AI STANDARDS LANDSCAPE

12.8.1 THE INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO) AND THE INTERNATIONAL ELECTROTECHNICAL COMMISSION (IEC)

In 2017, the ISO and the IEC created a joint technical committee on AI: ISO/IEC JTC 1/SC 42 (SC 42), which is tasked with developing International Standards for AI.

To undertake its work, SC 42 takes a comprehensive look at the ecosystem in which AI systems are developed and deployed. By looking at the context of use of the technology, such as application domain, business, societal, and regulatory requirements, SC 42 develops horizontal standards for applications that address areas such as data quality, privacy, security, trustworthiness, and ethics.

The standards under development in SC 42 are managed by working groups that focus on delivering horizontal standards in areas including foundational standards, data, trustworthiness, use cases and applications, and computational approaches. The core of the committee’s work is to develop guidance on foundational standards that establish underpinning concepts and terminologies, data standards that address data governance and management, model standards that define structure and support interoperability and compatibility, and organisational standards that provide tools for oversight of AI systems and the management of risk. See Figure 12.3.

At the time of writing, SC 42 is made up of 36 participating members and 22 observing members and has published 20 standards with a further 27 under development. Figure 12.4 represents the current suite of ISO/IEC AI standards, including those in development, and Table 12.2 describes each of the SC 42 publications.

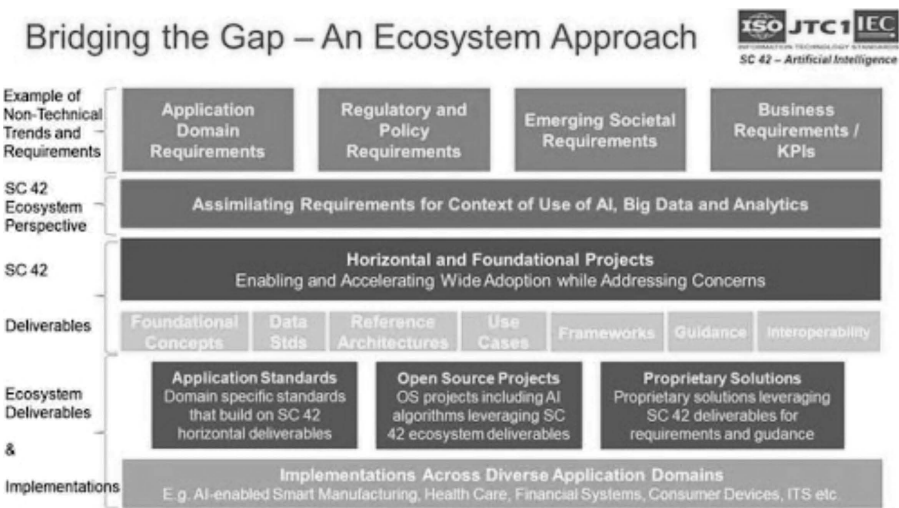


FIGURE 12.3 SC 42 – An ecosystem approach.

Source: The American National Standards Institute (ANSI Standards, 2022, p. 16).



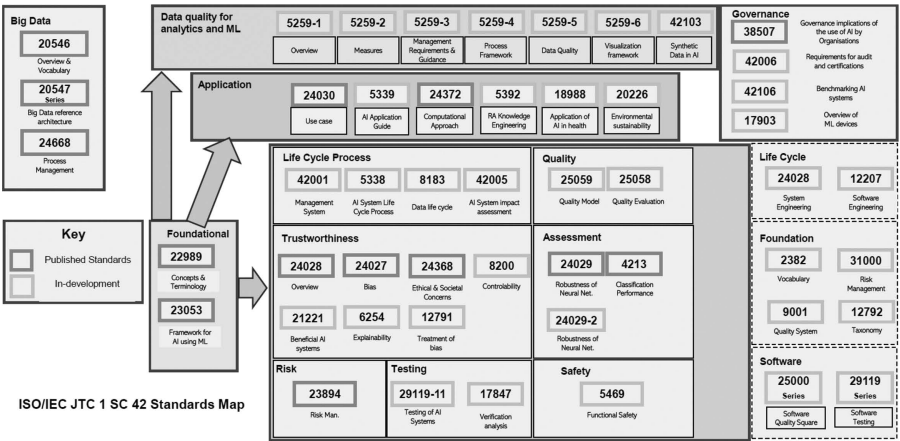


FIGURE 12.4 ISO/IEC JTC 1 SC 42 standards map.

TABLE 12.2  
Key ISO/IEC JTC 1 SC 42 Publications That Support AI at Scale

Publications That Support AI at Scale

Data	<p><b>ISO/IEC 5259 series (Under development)</b></p> <p>This set of standards will provide tools and methods to assess and improve the quality of data used for analytics and machine learning. The series includes guidelines for data governance, data quality assessment, measurement, and improvement for both training and operation.</p>
Data	<p><b>ISO/IEC 8183:2023 Information technology – Artificial intelligence (AI) – Data life cycle framework</b></p> <p>This standard defines the stages and identifies associated actions for data processing throughout the AI system lifecycle, including acquisition, creation, development, deployment, maintenance, and decommissioning.</p>
Data	<p><b>ISO/IEC 12791 Information technology – Artificial intelligence – Treatment of unwanted bias in classification and regression machine learning tasks (Under development)</b></p> <p>This standard will provide mitigation techniques that can be applied throughout the AI system lifecycle in order to treat unwanted bias. This document describes how to address unwanted bias in AI systems that use machine learning to conduct classification and regression tasks.</p>
Model	<p><b>ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)</b></p> <p>This standard establishes an AI and machine learning (ML) framework for describing a generic AI system using ML technology. The framework describes the system components and their functions in the AI ecosystem. This document is applicable to all types and sizes of organisations, including public and private companies, government entities, and not-for-profit organisations, that are implementing or using AI systems.</p>

**Table 12.2 (Continued)****Key ISO/IEC JTC 1 SC 42 Publications That Support AI at Scale**

<b>Publications That Support AI at Scale</b>	
<b>Model</b>	<p><b>ISO/IEC TR 24028:2020 Information technology – Artificial intelligence (AI) – Overview of trustworthiness in artificial intelligence</b></p> <p>This standard provides guidance related to trustworthiness in AI systems. It includes guidance on approaches to establish trust in AI systems through transparency, explainability, and controllability; engineering pitfalls and typical associated threats and risks to AI systems, along with possible mitigation techniques and methods; and approaches to assess and achieve availability, resilience, reliability, accuracy, safety, security, and privacy of AI systems.</p>
<b>Model</b>	<p><b>ISO/IEC TR 24372:2021 Information technology – Artificial intelligence (AI) – Overview of computational approaches for AI systems</b></p> <p>This document provides an overview of the state of the art of computational approaches for AI systems, by describing: (a) main computational characteristics of AI systems; and (b) main algorithms and approaches used in AI systems.</p>
<b>Model</b>	<p><b>ISO/IEC 6254 Information technology – Artificial intelligence (AI) – Objectives and approaches for explainability of ML models and AI systems (Under development)</b></p> <p>This document will describe approaches and methods that can be used to achieve explainability objectives of stakeholders with regards to ML models and AI systems' behaviours, outputs, and results.</p>
<b>Model</b>	<p><b>ISO/IEC TR 5469 Artificial intelligence – Functional safety and AI systems (Under development)</b></p> <p>This technical report (TR) will cover the characteristics, potential hazards, and techniques and processes associated with the implementation of AI within safety-critical operations, the use of non-AI safety measures to guarantee safety for equipment controlled by AI, and the use of AI systems to create and develop safety-related functions.</p>
<b>Model</b>	<p><b>ISO/IEC 8200 Information technology – Artificial intelligence – Controllability of automated artificial intelligence systems (Under development)</b></p> <p>This document will define a basic framework with principles, characteristics, and approaches for the realisation and enhancement for automated AI systems' controllability.</p>
<b>Model</b>	<p><b>ISO/IEC 5392 Information technology – Artificial intelligence – Reference architecture of knowledge engineering (Under development)</b></p> <p>This standard will define a reference architecture of knowledge engineering (KE) in AI. The reference architecture describes KE roles, activities, constructional layers, components, and their relationships amongst themselves and other systems from systemic user and functional views.</p>
<b>Organisation</b>	<p><b>ISO/IEC 42001 Information technology – Artificial intelligence (AI) – Management system (Under development)</b></p> <p>This standard will provide a framework for a management system that an organisation can follow to meet its AI objectives using good practice. The standard takes a risk-based approach and targets AIMS, outlining guidelines for measuring effectiveness and efficiency of these systems, as well as for the responsible development and use of such systems that meet applicable regulatory requirements. It is designed to be auditable and is expected to be a pathway to certification.</p>

*(Continued)*

**Table 12.2 (Continued)**  
**Key ISO/IEC JTC 1 SC 42 Publications That Support AI at Scale**

Publications That Support AI at Scale	
Organisation	<p><b>ISO/IEC 22989:2022 Information technology – Artificial intelligence – Artificial intelligence concepts and terminology</b></p> <p>This document establishes terminology for AI and describes concepts in the field of AI. This document can provide organisations with a better understanding of AI and can support them to consider AI initiatives. It also supports communications among diverse, interested parties or stakeholders by providing a common understanding of AI and its concepts and terminology.</p>
Organisation	<p><b>ISO/IEC 23894:2023 – Information technology – Artificial intelligence – Guidance on risk management</b></p> <p>This document provides guidance on how organizations that develop, produce, deploy, or use products, systems, and services that utilise AI can manage risk specifically related to AI.</p>
Organisation	<p><b>ISO/IEC 5339 Information technology – Artificial intelligence – Guidance for AI applications (Under development)</b></p> <p>This document provides guidance for identifying the context, opportunities, and processes for developing and applying AI applications.</p>
Organisation	<p><b>ISO/IEC 22989:2022 Information technology – Artificial intelligence – Artificial intelligence concepts and terminology</b></p> <p>This document establishes terminology for AI and describes concepts in the field of AI. This document can provide organisations with a better understanding of AI and can support them to consider AI initiatives. It also supports communications among diverse, interested parties or stakeholders by providing a common understanding of AI and its concepts and terminology.</p>
Organisation	<p><b>ISO/IEC TR 24027:2021 Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making</b></p> <p>This standard provides guidance on how organisations that develop, produce, deploy systems and services that utilise AI can manage risk specifically related to AI. The guidance also aims to assist organisations to integrate risk management into their AI-related activities and functions. It describes processes for the effective implementation and integration of AI risk management and its application can be customised to any organisation and its context.</p>
Organisation	<p><b>ISO/IEC TR 24368:2022 Information technology – Artificial intelligence – Overview of ethical and societal concerns</b></p> <p>This standard provides a high-level overview of AI ethical and societal concerns. It also provides information in relation to principles, processes, and methods in this area, including an overview of International Standards that address issues arising from AI ethical and societal concerns.</p>
Organisation	<p><b>ISO/IEC 42005 Information technology – Artificial intelligence – AI system impact assessment (Under development)</b></p> <p>This document will provide guidance for organisations performing AI system impact assessments for individuals and societies that can be affected by an AI system and its intended and foreseeable applications.</p>

Source: Ethical AI Consulting.

### **12.8.2 THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS STANDARDS ASSOCIATION (IEEE)**

The Institute of Electrical and Electronic Engineers Standards Association (IEEE SA) is a globally recognised standards development organisation with international membership that is focused on developing standards that advance technology and technological innovation. The IEEE has 40,000 individual experts who create standards in engineering, computing, and information technology.

The IEEE has undertaken significant work in developing AI standards in several key areas, including ethics, transparency, accountability, and interoperability. This includes the release of a number of documents regarding the ethical design and development of AI through their Global Initiative on Ethics of Autonomous and Intelligent Systems. They aim to address the challenges and concerns associated with AI, such as bias, fairness, privacy, and algorithmic transparency.

The IEEE's P7000™ series of standards includes a number of specific standards that address different aspects of AI design, development, and evaluation. Key IEEE Standards that have been developed or are under development in relation to ethical AI are listed as follows in Table 12.3.

### **12.8.3 THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST)**

The National Institute of Standards and Technology (NIST) was founded in 1901 and is now part of the U.S. Department of Commerce. In its role as federal AI standards coordinator, NIST leads and participates in the development of technical standards, including International Standards, that promote innovation and public trust in systems that use AI.

NIST focus areas for standards development include (Tabassi et al., 2019, p. 18) the following, as shown in Figure 12.5.

One of NIST's focus areas is aligning the NIST AI RMF Roadmap and related guidance with applicable International Standards, guidelines, and practices. The roadmap specifically cites "Alignment with International Standards and production crosswalks to related standards (e.g., ISO/IEC 5338, ISO/IEC 38507, ISO/IEC 22989, ISO/IEC 24028, ISO/IEC DIS 42001, and ISO/IEC NP 42005)."

### **12.8.4 THE EUROPEAN COMMITTEE FOR STANDARDIZATION (CEN) AND THE EUROPEAN COMMITTEE FOR ELECTROTECHNICAL STANDARDIZATION (CENELEC)**

The European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC) are two distinct private international non-profit organizations. There are 200,000 technical experts from industry, associations, and public administrations who come from 34 Member Economies and relevant government bodies.

ISO/IEC have recently established the "Vienna Agreement", which secures collaboration between CEN-CENELEC and ISO/IEC JTC 1/SC 42 for the development

**TABLE 12.3**  
**Key IEEE Publications and Initiatives That Support AI at Scale**

**IEEE SA P2863 – Recommended Practice for Organizational Governance of Artificial Intelligence**

This standard provides a framework of recommended practice and outlines criteria for trustworthy AI, such as transparency, accountability, and safety. It also provides guidance on how to responsibly develop or use AI, such as auditing, training, and complying with regulations.

**IEEE 7010–2020: IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being**

This recommended practice provides specific and contextual well-being metrics that facilitate the use of a Well-Being Impact Assessment (WIA) process in order to proactively increase and help safeguard human well-being throughout the lifecycle of autonomous and intelligent systems (A/IS).

**IEEE 7000–2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design**

This standard provides a model process for addressing ethical concerns during AI system design. It provides a clear methodology to analyse human and social values for an ethical system engineering effort. The standard establishes a set of processes enabling organisations to include consideration of human ethical values in the design of AI and AI systems.

**IEEE 7001–2021: Standards for Transparency of Autonomous Systems**

That describes specific, measurable levels of transparency that can be assessed objectively, and identifies various levels of compliance that can be determined during system design.

**IEEE P7003 Standard for Algorithmic Bias Considerations (Under development)**

This standard will provide individuals or organisations creating algorithmic systems with a development framework to avoid unintended, unjustified, and inappropriately differential outcomes for users.

**IEEE P3119 Standard for the Procurement of Artificial Intelligence and Automated Decision Systems (Under development)**

This standard will establish a uniform set of definitions and a process model for the procurement of Artificial Intelligence (AI) and Automated Decision Systems (ADS) by which government entities can address socio-technical and responsible innovation considerations to serve the public interest.

**IEEE P7009 Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems (Under development)**

This standard will establish a practical, technical baseline of specific methodologies and tools for the development, implementation, and use of effective fail-safe mechanisms in autonomous and semi-autonomous systems.

**IEEE CertifAIEd™**

IEEE CertifAIEd is a certification program for assessing ethics of Autonomous Intelligent Systems (AIS) to help protect, differentiate, and grow product adoption. The resulting certificate and mark demonstrates the organisation’s effort to deliver a solution with a more trustworthy AIS experience to their users. The Mark helps organisations to demonstrate that they are addressing four key areas: transparency, accountability, algorithmic bias, and privacy.

**The IEEE Applied Artificial Intelligence Systems (AIS) Risk and Impact Framework Initiative (Under development)**

This initiative will provide a risk assessment and mitigation paradigm based on previous models but tailored to AI. It will identify existing risk approaches (in the fields of finance and cybersecurity) as well as gaps, to create an AI risk assessment for risk management.



**FIGURE 12.5** Focus areas of NIST.

of joint standards activities. This agreement supports the uplift of AI standards by establishing new joint projects and working groups,

agreeing to incorporate ISO/IEC standards into the European Union’s AI Act and streamlining the adoption of existing CEN and ISO/IEC publications. This collaboration will look to promote global uptake and harmonization of both product types.

**12.8.5 EUROPEAN TELECOMMUNICATIONS STANDARDS INSTITUTE (ETSI)**

The ETSI is a European Standards Organization (ESO) with an international membership. They are the regional standards body for telecommunications, broadcasting, and other electronic communications networks and services. ETSI has 60 member countries and 900 organisations. Key AI Standards include:

- ETSI GR SAI 009 – Artificial Intelligence Computing Platform Security Framework.
- ETSI GR SAI 001 – AI Threat Ontology

**12.9 CONCLUSION**

This chapter sets out background, context, and some of the key discussion points for APEC to take into consideration for the August 2023 *Supporting AI at Scale in the APEC Region Through International Standards* workshop and any subsequently agreed outcomes and actions. It is an important step in the ongoing efforts to develop and promote the APEC region’s readiness to capitalise on AI. The chapter provides an overview of the key barriers to operationalising AI and details how International Standards can support responding to these barriers to enable responsible scaling of AI development and deployment in the APEC Region.

AI is everywhere. It is becoming more embedded in our lives every day – from facial recognition systems that unlock our smartphones, to AI systems that are used to recommend our favourite music, movie, or to help us draft emails faster, to systems that are used to detect cancer or find a vaccine against Covid. There are countless and

increasingly many applications. Unsurprisingly, with its seemingly rapid emergence and democratisation, there is a lot of discussion about the benefits and indeed the risks of AI.

Countless bodies of work internationally are seeking to build guardrails for AI development and deployment to ensure responsible AI practices. Worldwide, economies are establishing AI policies and roadmaps, while others are pushing ahead with AI regulation. This is all critically important work, yet as we are facing an increasingly fragmented ecosystem, we are at risk of confusing, rather than empowering, business and making emerging technologies too difficult to adopt due to fragmented markets if it is not underpinned by common consensus points.

This is why International Standards have a key role to play not only in facilitating the responsible adoption but also in enabling scaling of quality systems. The standards that are highlighted in this report are globally recognised benchmarks that present agile solutions to shape design, deployment, and evaluation of AI that promote harmonisation and interoperability across markets. Importantly for AI, they also promote responsibility, trustworthiness, security, and confidence in emerging systems.

This chapter sets out that International Standards can support stakeholders in the APEC region to overcome the challenges and risks that are limiting the scale of AI in the APEC region. International Standards provide globally agreed frameworks that promote harmonisation and interoperability in products and services, and across borders. They establish common building blocks for companies and policymakers and set the specifications and requirements upon which new technologies can be responsibly developed, adopted, and deployed. We encourage industry and policymakers across APEC to review and consider the findings outlined in this chapter. Ultimately, increasing the use of these standards and the engagement from APEC economies in their development will support the use of AI technologies in the APEC region in a safe and appropriate way.

## REFERENCES

- Arai, M., & Law, I. N. Q. (2023). *Discerning signal from noise: The state of global AI standardization and what it means for Canada*. [https://scc-ccn.ca/system/files/2024-05/sri\\_discerningsignalfromnoise\\_english\\_v2.pdf](https://scc-ccn.ca/system/files/2024-05/sri_discerningsignalfromnoise_english_v2.pdf)
- Coghi, J., & Jelitto, M. (2023). Services domestic regulation in the WTO. In *The Elgar companion to the world trade organization* (pp. 361–376). Edward Elgar Publishing.
- Cuyegkeng, S., & Evans, P. (2022). *Navigating US-China digital geopolitics for non-superpower states: Singapore as a case study post-October 2022*. <https://sofiacuyegkeng.ubcart.ca/wp-content/uploads/sites/2215/2023/01/POLI-460-Paper.pdf>
- Edlich, A., Phalin, G., Jogani, R., & Kaniyar, S. (2019). Driving impact at scale from automation and AI. *McKinsey Global Institute*, 100.
- EY study: AI important to a company's success, but lack of skilled personnel remains a barrier. (n.d.). [www.ey.com](https://www.ey.com/en_gl/news/2019/08/ey-study-ai-important-to-a-companys-success-but-lack-of-skilled-personnel-remains-a-barrier). [https://www.ey.com/en\\_gl/news/2019/08/ey-study-ai-important-to-a-companys-success-but-lack-of-skilled-personnel-remains-a-barrier](https://www.ey.com/en_gl/news/2019/08/ey-study-ai-important-to-a-companys-success-but-lack-of-skilled-personnel-remains-a-barrier)
- Gartner. (2018). *Gartner says nearly half of CIOs are planning to deploy artificial intelligence*. Gartner. <https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence>

- Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). *Trust in artificial intelligence: A global study* (p. 10). The University of Queensland and KPMG Australia.
- Hiebert, K. (2022, April 27). *Democracies are dangerously unprepared for deepfakes*. Centre for International Governance Innovation. <https://www.cigionline.org/articles/democracies-are-dangerously-unprepared-for-deepfakes/>
- Satariano, A., & Mozur, P. (2023). The people onscreen are fake. The disinformation is real. *International New York Times*, NA-NA.
- Speaking the same AI language starts with standardization: Q&A with Wael William Diab, chair of ISO/IEC JTC 1/SC 42.* (2022, December 19). American National Standards Institute – ANSI. <https://www.ansi.org/standards-news/all-news/2022/12/12-19-22-speaking-the-same-ai-language-starts-with-standardization>
- Tabassi, E., Carnahan, L., Hogan, M., & Heyman, M. (August 10, 2019). *US leadership in AI: A plan for federal engagement in developing technical standards and related tools*. <https://www.nist.gov/artificial-intelligence/plan-federal-engagement-developing-ai-technical-standards-and-related-tools>



---

# 13 Suggested Framework for Improved Algorithmic Auditing in India

*Harsh Lailer, Gadamsetti Srija, Aseem Saxena,  
and Agrima Lailer*

## 13.1 ALGORITHMIC SYSTEMS

The Association for Computing Machinery defines an *algorithm* as a self-contained step-by-step set of operations that computers and other ‘smart’ devices carry out to perform calculation, data processing, and automated reasoning tasks. Increasingly, algorithms implement institutional decision-making based on analytics, which involves the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming, and operations research to quantify performance (Transparency International, 2021).

## 13.2 ALGORITHMIC AUDITING

Algorithmic auditing is a systematic and rigorous examination of algorithms, usually implemented in automated decision-making systems, to assess their fairness, accountability, transparency, and overall compliance with ethical and legal standards. This process involves a detailed examination of the algorithm’s design, development, and deployment stages to identify biases, discriminatory patterns, or unintended consequences that may impact individuals or groups. Algorithmic auditing aims to ensure that algorithms adhere to ethical guidelines, legal regulations, and organizational policies, providing a means to address and rectify potential issues related to algorithmic decision-making. The audit may involve analyzing the training data, evaluating the decision-making processes, and assessing the algorithm’s outcomes to mitigate any negative impacts and promote responsible and equitable use of algorithms.

## 13.3 ALGORITHMIC IMPACT ASSESSMENTS

Algorithmic impact assessments refer to a structured and systematic evaluation process applied to automated decision-making systems or algorithms to assess and comprehend their potential societal, ethical, and legal impacts. These assessments

are designed to identify and analyze the various effects that algorithmic systems may have on individuals, communities, and broader societal structures. Algorithmic impact assessments aim to provide a comprehensive understanding of the social implications of algorithmic decision-making and assist organizations in making informed decisions to minimize negative consequences and promote ethical and equitable use of algorithms.

### 13.4 ALGORITHMIC ACCOUNTABILITY

As per Maranke Wieringa's definition, *algorithmic accountability* revolves around establishing a connected explanation for a socio-technical algorithmic system across its different lifecycle stages. Within this framework of accountability, multiple participants (such as decision makers, developers, and users) bear the responsibility to clarify and defend their actions, designs, and decisions related to the system, along with the subsequent impacts of those actions. Given the diverse actors involved throughout the system's lifespan, they may be held accountable through various types of forums (internal/external to the organization, formal/informal). This accountability could be specific to certain aspects of the system (a modular account) or extend to the entirety of the system (an integral account). These forums must possess the capability to raise inquiries and make judgments, leading to potential consequences for one or more actors. The dynamics between the forum/forums and the actor(s) stem from a specific viewpoint on accountability (Metcalf et al., 2021).

### 13.5 METHODS FOR ALGORITHMIC ACCOUNTING

Raji et al. devised a comprehensive six-step auditing protocol for pymetrics, elucidating the intricate process through which predictive models for candidate screening are developed and deployed (Raji et al., n.d.):

1. **Contracting Phase:** Employers, referred to as clients, initiate the process by contracting with pymetrics to create and implement a predictive model tailored for candidate screening.
2. **Client Survey:** A dedicated job analyst from pymetrics engages in a detailed survey with the client, aiming to comprehend the nuances of the target role, including job descriptions, seniority levels, and key performance metrics employed by the client.
3. **Gameplay Data Collection:** Incumbent employees in the specified role partake in pymetrics' suite of games, further providing existing job performance data. This amalgamation of performance and gameplay data serves as the foundational training input for the ensuing predictive model.
4. **Model Development and Evaluation:** A pymetrics data scientist utilizes a proprietary tool to craft a predictive model for the client. Rigorous evaluation follows, assessing predictive performance and adherence to the Uniform Guidelines on Employee Selection Procedures (UGESP) using a separate testing set with demographic information.

5. **Model Deployment and Applicant Assessment:** The best-performing model meeting fairness criteria is deployed by pymetrics. Job seekers applying for the role undergo pymetrics' game-based assessment, enabling the model to identify candidates with attributes akin to high-performing incumbents. Information on high-scoring candidates is relayed to the client, who may apply additional filters and proceed with interviews.
6. **Longitudinal Analysis:** Pymetrics conducts ongoing longitudinal analysis, incorporating back-testing to reassess the model's adherence to fairness criteria concerning the pool of job seekers for the role. Additionally, an in-depth examination of the job performance of hired candidates contributes to the continual refinement of the predictive model.

The five designs proposed by Sandvig et al. (Sandvig et al., 2014) are as follows:

1. **Code Audit (Algorithm Transparency):** This design aims to scrutinize disclosed algorithms for transparency and accountability. However, the reluctance of internet platforms to reveal proprietary algorithms as valuable intellectual property poses a significant challenge. Even if disclosure were compelled, complexities arise from the constant cat-and-mouse game between algorithm designers and potential abusers, potentially aiding criminal adversaries and leading to unintended consequences. While proposals suggest third-party escrow for algorithm scrutiny, the intricate nature of modern algorithms and their reliance on personal data makes straightforward interpretation challenging.
2. **Noninvasive User Audit:** Focused on gathering user interaction information through surveys, this design avoids perturbing the platform but lacks experimental design and faces validity issues due to reliance on self-reported data. Sampling problems and difficulties in investigating sensitive domains limit its effectiveness, making it challenging to infer causality from results and undermining its utility for detecting harmful discrimination.
3. **Scraping Audit:** In this design, researchers issue repeated queries to the platform, observing results. However, legal issues, such as potential violations of the Computer Fraud and Abuse Act and platform terms of service, present obstacles. Lack of randomization and manipulation further limits the ability to infer causality, making it suitable primarily for investigating publicly available information.
4. **Sock Puppet Audit:** Involving computer programs to impersonate users for controlled manipulation, the sock puppet audit introduces challenges such as potential legal issues, deception, and claims of injecting false data. While offering control over data collection, the legality of injecting false data and potential harm claims by the platform present significant obstacles, raising doubts about its practical workability in many situations.
5. **Crowdsourced Audit/Collaborative Audit:** This design engages hired users or volunteers to act as testers, potentially creating networks for accountability. Overcoming some legal issues and introducing a human

element, it allows large-scale data collection. However, cost considerations and ethical concerns related to injecting false data remain challenges, highlighting the need for careful consideration and potential collaboration with volunteers interested in public interest problems associated with algorithms.

### **13.6 THIRD-PARTY ALGORITHMIC AUDITING**

Third-party algorithmic auditing refers to the process of having an independent, external entity assess and evaluate the algorithms used by organizations. This auditing is conducted to ensure transparency, fairness, accountability, and ethical considerations in the development and deployment of algorithms. The goal is to identify and mitigate potential biases, discrimination, or unintended consequences that may arise from algorithmic decision-making systems. These audits are essential for building trust, promoting accountability, and addressing societal concerns associated with the increasing use of algorithms in various domains such as finance, employment, criminal justice, and healthcare.

### **13.7 THE FOUNDATION MODEL TRANSPARENCY INDEX**

The Foundation Model Transparency Index is an assessment tool introduced to evaluate the transparency of the foundation model ecosystem. It consists of 100 indicators that comprehensively measure transparency for foundation models, including upstream resources, model details, and downstream use. It has three subdomains including upstream (e.g., the data, labor, and compute resources used to build a foundation model), model-level (e.g., the capabilities, risks, and evaluations of the foundation model), and downstream (e.g., the distribution channels, usage policies, and affected geographies) practices of the foundation model developer. Each subdomain has around 32–35 indicators. The index scores 10 major foundation model developers, namely, OpenAI (GPT-4), Anthropic (Claude 2), Google (PaLM 2), Meta (Llama 2), Inflection (Inflection-1), Amazon (Titan Text), Cohere (Command), AI21 Labs (Jurassic-2), Hugging Face (BLOOMZ; as host of BigScience), and Stability AI (Stable Diffusion 2), against these indicators to assess their transparency. The aim of the index is to drive progress on foundation model governance through industry standards and regulatory intervention. It also aims to improve the overall transparency of the AI ecosystem by encouraging developers to share more information about the development and deployment of their models. The index provides a frame of reference for assessing transparency in the ecosystem and identifies areas where greater transparency would be valuable. Future versions of the index will adjust the indicators to reflect changes in the foundation model ecosystem and AI policy.

### **13.8 TECHNICAL ELEMENTS OF ALGORITHMIC IMPACT ASSESSMENTS**

It constitutes a layer of organizational accountability specifically within the realm of constructing and deploying automated decision-support systems. Drawing insights

**TABLE 13.1**  
**Algorithm Impact Assessments**

1	Data Analysis	Examination of the training data used to develop the algorithm, focusing on potential biases, representativeness, and data quality issues.
2	Algorithmic Design and Functionality	Design and functionality: In-depth scrutiny of the algorithm’s architecture, logic, and decision-making processes to identify any inherent biases, discrimination, or unintended consequences.
3	Model Evaluation	Rigorous testing and validation of the algorithm’s performance, considering metrics such as accuracy, fairness, and interpretability.
4	Transparency and Explainability	Assessment of the algorithm’s transparency and the ability to provide understandable explanations for its decisions, especially in contexts where the impact on individuals’ lives is significant.
5	Legal and Ethical Compliance	Examination of the algorithm’s adherence to legal regulations, ethical standards, and organizational policies to ensure responsible and lawful deployment.
6	Mitigation Strategies	Development and recommendation of strategies to address identified issues, mitigate potential harms, and enhance the overall fairness and accountability of the algorithmic system.

from diverse domains, they derive inspiration from established impact assessment methodologies. The term “impacts” functions as a strategic tool enabling stakeholders to identify and alleviate adverse consequences arising from policy decisions or system implementations. AIAs operate as a governance instrument, unveiling the drawbacks associated with algorithmic systems and instigating corrective measures. Algorithmic impact assessments do not function as impartial measuring instruments; rather, they act as representations of the socio-material harms potentially generated by algorithmic systems. The challenge in their development lies in establishing algorithmic impact assessments as effective governance mechanisms within intricate power dynamics and contested outcomes. Overcoming this challenge involves conceptualizing impacts as co-constructed accountability relationships, striving to align these impacts with actual harms, and integrating diverse expertise and perspectives from affected communities into the fabric of the accountability governance process.

**13.9 RISKS OF AI DEVELOPMENT**

The recent consensus paper *Managing AI Risks in an Era of Rapid Progress* (Bengio et al., 2024), written by multiple leaders in the science of AI, discusses the risks associated with the development of AI systems. We go through the risks that they’ve mentioned while also discussing their proposed solutions. AI systems have been observed to follow a scaling law, which states that the training error of an AI model decreases as a power of the dataset size as well as model size. Scaling has the potential to

lead to emergent capabilities on a plethora of tasks, although not evenly across all tasks. AI systems have the potential to outperform humans in various tasks, but if not carefully designed, deployed, and audited, they can pose societal-scale risks, including amplifying social injustice, eroding social stability, and weakening our shared understanding of reality. The development of autonomous AI systems, which can plan, act in the world, and pursue goals, could amplify existing risks and create new ones, such as automated warfare, mass manipulation, and pervasive surveillance. As AI systems become faster and more cost-effective than human workers, there is a dilemma where companies, governments, and militaries may be forced to deploy AI systems widely and reduce human verification of AI decisions, potentially leading to a loss of control over autonomous AI systems. Building highly advanced autonomous AI systems without reliable methods to audit and align their behavior with complex values, without sufficient safety testing and human oversight, can lead to systems pursuing unintended and potentially harmful goals, which may be difficult to control. Overall, the risks of unaudited AI development include societal-scale consequences, loss of control over autonomous systems, and the pursuit of undesirable goals by AI systems.

There are two proposed measures to manage risks of AI, discussed in the following sections.

13.9.1 TECHNICAL MEASURES FOR PROPER AUDITING OF AI SYSTEMS

Research breakthroughs related to auditing need to address the technical challenges in creating AI systems with safe and ethical objectives. Simply making AI systems more capable may not be sufficient to solve these challenges. Some of the challenges that require research breakthroughs include oversight and honesty, robustness, interpretability and transparency, inclusive AI development, risk evaluations, and addressing emerging challenges. The following measures shown in Table 13.2 can be taken into account for auditing algorithms.

TABLE 13.2  
Technical Measures for Auditing AI Systems

1	Oversight and honesty	More capable AI systems can exploit weaknesses in oversight and testing, leading to false but compelling output.
2	Robustness	AI systems behave unpredictably in new situations, such as under distribution shift or adversarial inputs.
3	Interpretability and transparency	AI decision-making is currently opaque, and there is a need to understand the inner workings of AI models.
4	Inclusive AI development	Methods are needed to mitigate biases and integrate the values of the populations affected by AI advancement.
5	Risk evaluations	Better evaluation methods are required to detect hazardous capabilities of AI systems earlier.
6	Addressing emerging challenges	Future AI systems may exhibit failure modes and learn to feign obedience or exploit weaknesses in safety objectives and shutdown mechanisms.

### 13.9.2 GOVERNANCE MEASURES FOR AUDITING AI SYSTEMS

In the realm of auditing, the current landscape presents a notable absence of robust governance frameworks for AI. This void introduces inherent risks, particularly as entities such as companies, militaries, and governments may prioritize the advancement of AI capabilities without adequate consideration for safety and human oversight. The absence of well-defined regulatory structures raises concerns about accountability and ethical deployment of AI technologies. Addressing this challenge necessitates the establishment of comprehensive governance mechanisms. National institutions must possess both strong technical expertise and authoritative powers to swiftly enforce regulations. The fast-paced nature of AI progress requires agility in regulatory responses to ensure that the development and deployment of AI align with ethical and safety considerations. Moreover, a collaborative approach on the international stage becomes imperative. Agreements and partnerships should be forged to tackle the intricate dynamics of AI development, particularly with regard to issues related to bias and fairness. This collaborative effort becomes an essential component in mitigating risks associated with the global proliferation of AI technologies. To safeguard low-risk use and encourage academic research, it is crucial to minimize bureaucratic hurdles for small and predictable AI models. This facilitates innovation in a responsible manner while ensuring that regulatory processes do not stifle progress in areas where the risks are comparatively lower.

Governments play a pivotal role in licensing the development of exceptionally capable AI systems. This involves strategically restricting their autonomy and mandating stringent information security measures. Such measures are essential to prevent the misuse of highly advanced AI technologies that may have significant societal impacts. In addition, major AI companies should commit to specific safety standards, and these commitments should undergo independent scrutiny. This external validation ensures transparency and builds trust in the safety measures implemented by these companies. The auditing framework for AI development and deployment requires a multi-faceted approach that includes national regulations, international collaboration, minimized bureaucratic hurdles, comprehensive insight mechanisms, safety standards, legal accountability, and commitments from major AI companies. This holistic strategy aims to strike a balance between fostering innovation and ensuring responsible and ethical use of AI technologies.

Following is the Terry Group's list of some of the algorithmic accountability policies that have been proposed/implemented throughout the world in order to appreciate their importance in today's world of quick development of AI systems and their incorporation into various industries (Terry Group, n.d.; Table 13.3), as well as the internationally accepted principles for auditing AI systems (Table 13.4).

### 13.10 THE SOCIAL HARMS OF AI IN INDIA THROUGH ALGORITHMIC AUDITING

Weidinger et al., in the paper "Taxonomy of Risks Posed by Language Models," mentions six risks areas related to social harms of AI in India, discussed in the following sections.

**TABLE 13.3**  
**Examples of Various Algorithmic Accountability Policies Across the World**

Title	Summary	Country/Year	Status
Washington, D.C. Stop Discrimination by Algorithms Act (Council of the District of Columbia, n.d.)	Protects against discrimination by automated decision-making tools and gives Washington D.C. residents transparency about how algorithms are used to determine outcomes in everyday life – including in credit, housing, and employment.	U.S., 2023	Proposed
The Artificial Intelligence Act (Artificial Intelligence Act, n.d.)	Comprehensive AI law. Assigns applications of AI to three risk categories: (1) applications and systems that create an unacceptable risk are banned; (2) high-risk applications are subject to specific legal requirements; and (3) applications not explicitly banned or listed as high-risk are largely left unregulated.	Europe, 2021, 2022, 2023	Proposed
Brazil’s New AI Bill: A Comprehensive Framework for Ethical and Responsible Use of AI Systems (Access Partnership, 2023)	Its primary aim is to grant individuals significant rights and place specific obligations on companies that develop or use AI technology. The bill establishes a new regulatory body to enforce the law and takes a risk-based approach by organizing AI systems into different categories. It also introduces civil liability for providers or operators of AI systems, along with a reporting obligation for significant security incidents.	Brazil, 2023	Proposed
UAE National Strategy for Artificial Intelligence (United Arab Emirates Government, 2021)	An Artificial Intelligence and Blockchain Council will “review national approaches to issues such as data management, ethics and cybersecurity,” and observe and integrate global best practices on AI.	United Arab Emirates, 2018	Published
Principles of Policy, Regulation and Ethics in AI (draft policy) (Ministry of Innovation, Science and Technology, 2022)	States that the development and use of AI should respect “the rule of law, fundamental rights and public interests and, in particular, [maintain] human dignity and privacy.” Furthermore, “reasonable measures must be taken in accordance with accepted professional concepts” to ensure AI products are safe to use.	Israel, 2022	Published
Proposed advisory guidelines on use of personal data in AI recommendation and decision systems(Personal Data Protection Commission [PDPC] Singapore, 2023)	The goal is to clarify how Singapore’s Personal Data Protection Act applies to the collection and use of personal data by organizations to develop and deploy machine learning models or AI systems used to make decisions autonomously or to assist a human decision-maker.	Singapore, 2023	Published



**TABLE 13.4**  
**Internationally Accepted Principles for Auditing AI Systems**

Association for Computing Machinery	Asilomar AI Principles	UNESCO
1. Awareness	1. Research goal	1. Proportionality and “do no harm”
2. Access and redress	2. Research funding	2. Safety and security
3. Accountability	3. Science-policy link	3. Fairness and non-discrimination
4. Explanation	4. Research culture	4. Sustainability
5. Data provenance	5. Race avoidance	5. Right to privacy and data protection
6. Auditability	6. Safety	6. Human oversight and determination
7. Validation and testing	7. Failure transparency	7. Transparency and explainability
	8. Judicial transparency	8. Responsibility and accountability
	9. Responsibility	9. Awareness and literacy
	10. Value alignment	10. Multi-stakeholder and adaptive governance and collaboration
	11. Human values	
	12. Personal privacy	
	13. Liberty and privacy	
	14. Shared benefit	
	15. Shared prosperity	
	16. Human control	
	17. Non-subversion	
	18. AI arms race	
	19. Capability caution	
	20. Importance	
	21. Risks	
	22. Recursive self-improvement	
	23. Common good	

**13.10.1 RISK AREA 1: DISCRIMINATION, HATE SPEECH, AND EXCLUSION**

Language models (LMs) can perpetuate harmful stereotypes, unfair discrimination, and exclusion of marginalized groups, leading to social harm and injustice. LMs trained on biased or limited data can reproduce discriminatory language and reinforce social norms that marginalize certain identities. Harmful stereotypes and biases encoded in LMs can result in unfair treatment and allocation of resources between social groups. LMs may generate hate speech, offensive language, and language that incites violence, causing psychological harm and inciting hate or violence. Mitigation strategies include inclusive and representative training data, model fine-tuning to counteract stereotypes, and filtering out toxic statements from training corpora. Exclusionary norms in language can lead to LMs excluding or silencing identities that deviate from societal norms, causing allocational and representational harm. LMs trained on language data at a specific moment in time risk perpetuating frozen norms and inhibiting social change. LMs that encode exclusionary norms deny the existence of marginalized groups and reinforce historical marginalization. Overall, discrimination, hate speech, and exclusion are significant risks associated

TABLE 13.5  
Axes of Potential ML (Un)Fairness in India

<p><b>Caste</b> (17% Dalits; 8% Adivasi; 40% Other Backward Class (OBC))</p> <ul style="list-style-type: none"><li>• Societal harms: Human rights atrocities. Poverty. Land, knowledge and language battles.</li><li>• Proxies: Surname. Skin tone. Occupation. Neighborhood. Language.</li><li>• <b>Tech harms: Low literacy and phone ownership. Online misrepresentation and exclusion.</b></li></ul> <p><b>Accuracy gap of Facial Recognition (FR). Limits of Fitzpatrick scale. Caste-based discrimination in tech.</b></p>
<p><b>Gender</b> (48.5% female)</p> <ul style="list-style-type: none"><li>• Societal harms: Sexual crimes. Dowry. Violence. Female infanticide.</li><li>• Proxies: Name. Type of labor. Mobility from home.</li><li>• <b>Tech harms: Accuracy gap in FR. Lower creditworthiness score. Recommendation algorithms favoring majority male users. Online abuse and ‘racey’ content issues. Low Internet access.</b></li></ul> <p><b>Religion</b> (80% Hindu, 14% Muslim, 6% Christians, Sikhs, Buddhists, Jains and indigenous)</p> <ul style="list-style-type: none"><li>• Societal harms: Discrimination, lynching, vigilantism, and gang-rape against Muslims and others.</li><li>• Proxies: Name. Neighborhood. Expenses. Work. Language. Clothing.</li><li>• <b>Tech harms: Online stereotypes and hate speech, e.g., Islamophobia. Discriminatory inferences due to lifestyle, location, appearance. Targeted Internet disruptions.</b></li></ul> <p><b>Ability</b> (5%–8%+ persons with disabilities)</p> <ul style="list-style-type: none"><li>• Societal harms: Stigma. Inaccessible education, transport, and work.</li><li>• Proxies: Non-normative facial features, speech patterns, body shape, and movements. Use of assistive devices.</li><li>• <b>Tech harms: Assumed homogeneity in physical, mental presentation. Paternalistic words and images. No accessibility mandate.</b></li></ul> <p><b>Class</b> (30% live below poverty line; 48% on \$2–\$10/day)</p> <ul style="list-style-type: none"><li>• Societal harms: Poverty. Inadequate food, shelter, health, and housing.</li><li>• Proxies: Spoken and written language(s). Mother tongue. Literacy. Feature/Smartphone Ownership. Rural vs. urban.</li><li>• <b>Tech harms: Linguistic bias towards mainstream languages. Model bias towards middle class users. Limited or lack of internet access.</b></li></ul> <p><b>Gender Identity and Sexual Orientation</b> (No official LGBTQ+ data)</p> <ul style="list-style-type: none"><li>• Societal harms: Discrimination and abuse. Lack of acceptance and visibility, despite the recent decriminalization.</li><li>• Proxies: Gender declaration. Name.</li><li>• <b>Tech harms: FR “outing” and accuracy. Gender binary surveillance systems (e.g., in dormitories). M/F ads targeting. Catfishing and extortion abuse attacks.</b></li></ul> <p><b>Ethnicity</b> (4% NorthEast)</p> <ul style="list-style-type: none"><li>• Societal harms: Racist slurs, discrimination, and physical attacks.</li><li>• Proxies: Skin tone. Facial features. Mother tongue. State. Name.</li><li>• <b>Tech harms: Accuracy gap in FR. Online misrepresentation and exclusion. Inaccurate inferences due to lifestyle, e.g., migrant labor.</b></li></ul>

with LMs, and mitigating these risks requires inclusive training data, fine-tuning, and addressing biases in LM outputs.

Sambasivan et al. present an analysis in the form of Table 13.5 that captures the details related to ML unfairness in India (Metcalf et al., 2021).

### **13.10.2 RISK AREA 2: INFORMATION HAZARDS**

LMs can pose information hazards by disseminating private or sensitive information, leading to harm even without user error. Private or sensitive information can be revealed by LMs, such as trade secrets damaging businesses, health diagnoses causing emotional distress, and private data violating individuals' rights. Information hazards arise when LMs provide private or sensitive information present in training data or can be inferred from it. Observed risks in this area include privacy violations. Mitigation strategies for information hazards in LMs include algorithmic solutions and responsible model release strategies. In summary, information hazards in language models can occur when private or sensitive information is revealed, leading to harm. Mitigation strategies involve algorithmic solutions and responsible model release strategies.

### **13.10.3 RISK AREA 3: MISINFORMATION HARMS**

LMs can generate false, misleading, nonsensical, or poor-quality information, unintentionally misinforming or deceiving individuals and causing material harm. The deliberate generation of "disinformation," false information intended to mislead, is discussed separately under the section on Malicious Uses. Harms resulting from misinformation range from unintentional misinforming to causing material harm. Differential privacy, a framework for sharing information derived from a dataset while limiting inferences about individuals, is mentioned as a potential approach to address privacy concerns. In summary, misinformation harms in language models can occur when false or misleading information is generated, leading to unintentional misinforming or causing material harm. Mitigation strategies may involve addressing the quality of information generated by LMs and considering approaches like differential privacy to address privacy concerns.

### **13.10.4 RISK AREA 4: MALICIOUS USES**

Malicious use risks arise from intentional human use of LMs to cause harm, such as targeted disinformation campaigns, fraud, or malware. As LMs become more widely accessible, the risks of malicious use are expected to proliferate. It is difficult to scope all possible (mis-)uses of LMs, and further use-cases beyond those mentioned are possible. Responsible release of access to LMs and monitoring their usage are key mitigations to address malicious use risks. In summary, malicious use risks in LMs involve intentional human use to cause harm, such as disinformation campaigns or fraud. As LMs become more accessible, the risks of malicious use are expected to increase. Responsible release of access to LMs and monitoring their usage are important mitigation strategies to address these risks.

### **13.10.5 RISK AREA 5: HUMAN-COMPUTER INTERACTION HARMS**

LMs incorporated into dialogue-based tools, such as conversational agents (CAs), can lead to unsafe use due to users overestimating the model's capabilities. Interactions with LM-based conversational agents that seem similar to interactions with humans can create new avenues for privacy violations and exploitation. The supposed identity of the conversational agent can reinforce discriminatory stereotypes, leading to potential harm. Mitigations for these risks include penalizing or filtering certain

types of output and careful product design. In summary, the risk area of human-computer interaction harms in language models involves the potential for unsafe use, privacy violations, exploitation, and reinforcement of discriminatory stereotypes in interactions with conversational agents. Mitigation strategies include penalizing or filtering certain outputs and considering careful product design.

13.10.6 RISK AREA 6: ENVIRONMENTAL AND SOCIO-ECONOMIC HARMS

Large-scale language models (LLMs) require significant amounts of energy for training and operation, leading to environmental concerns. LLMs can contribute to social inequities due to the uneven distribution of risks and benefits of automation, potential loss of high-quality employment, and environmental harm. The specific impact of LLMs on the environment and socio-economic factors is complex and difficult to forecast, as it depends on various commercial, economic, and social factors. Mitigations for these risks include finding compute-efficient solutions for training LLMs, designing LLM applications with inclusionary goals, and monitoring the socio-economic impacts of LLMs. In summary, the risk area of environmental and socio-economic harms in LLMs involves concerns about the energy consumption of LLMs and the potential for social inequities. Mitigation strategies include developing more energy-efficient training methods, designing LLM applications with inclusive goals, and monitoring the socio-economic impacts of LLMs.

13.11 EXISTING THIRD-PARTY AI AUDITING METHODOLOGY

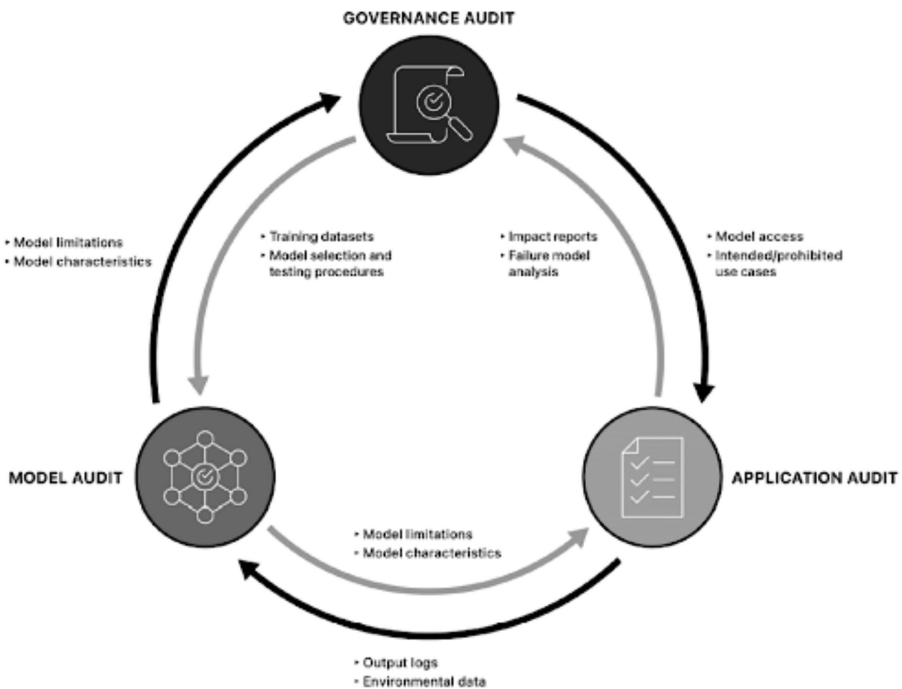


FIGURE 13.1 Outputs from audits on one level become inputs for audits on other levels.

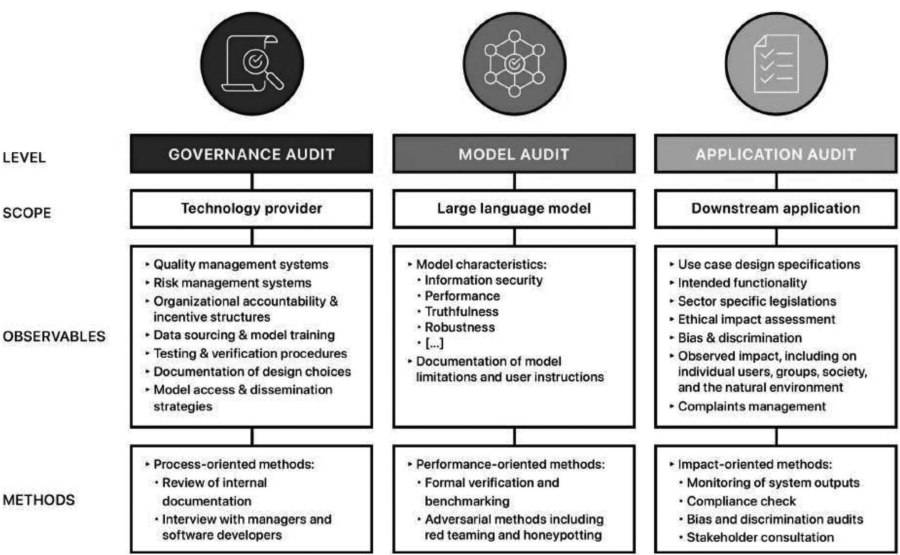


FIGURE 13.2    Blueprint for how to audit LLMs: A three-layered approach.

REFERENCES

Access Partnership. (2023). *Access alert: Brazil’s new AI bill – a comprehensive framework for ethical and responsible use of AI systems*. <https://accesspartnership.com/access-alert-brazils-new-ai-bill-a-comprehensive-framework-for-ethical-and-responsible-use-of-ai-systems/>

Artificial Intelligence Act. (n.d.). *The EU artificial intelligence act: A European approach to AI regulation*. <https://artificialintelligenceact.eu/>

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., & Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384(6698), 842–845. <https://doi.org/10.1126/science.adn0117>

Council of the District of Columbia. (n.d.). *Stop discrimination by algorithms act of 2023 (B25-0114)*. <https://lims.dccouncil.gov/Legislation/B25-0114>

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (FAcT ’21)* (pp. 735–746). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445935>

Ministry of Innovation, Science and Technology (Israel). (2022, October 31). *Israeli AI policy announcement*. <https://www.gov.il/he/departments/news/most-news20223110>

Personal Data Protection Commission (PDPC) Singapore. (2023, July 18). *Draft advisory guidelines on the use of personal data in AI recommendation and decision systems*. <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Legislation-and-Guidelines/Public-Consult-on-Proposed-AG-on-Use-of-PD-in-AI-Recommendation-and-Systems-2023-07-18-Draft-Advisory-Guidelines.pdf>

Raji, I. D., et al. (n.d.). *Pymetrics audit: A comprehensive six-step protocol for predictive screening models*. [https://evijit.io/docs/pymetrics\\_audit\\_FAcT.pdf](https://evijit.io/docs/pymetrics_audit_FAcT.pdf)

- Sandvig, C., et al. (2014). *Auditing algorithms: Research methods for detecting discrimination on internet platforms*. <https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf>
- Terry Group. (n.d.). *Algorithmic accountability: What is it and why does it matter?* <https://terrygroup.com/algorithmic-accountability-what-is-it-and-why-does-it-matter/>
- Transparency International. (2021). *Algorithmic transparency*. [https://knowledgehub.transparency.org/assets/uploads/kproducts/Algorithmic-Transparency\\_2021.pdf](https://knowledgehub.transparency.org/assets/uploads/kproducts/Algorithmic-Transparency_2021.pdf)
- United Arab Emirates Government. (2021). *UAE national strategy for artificial intelligence 2031*. <https://ai.gov.ae/wp-content/uploads/2021/07/UAE-National-Strategy-for-Artificial-Intelligence-2031.pdf>

---

# 14 Artificial Intelligence, Government, and Challenges

## *Initial Insights from Rwanda's Mbaza AI-Chatbot Project*

*Lea Gimpel and Keegan McBride*

### 14.1 INTRODUCTION

Scholarly and government interest in how the public sector can best use and implement AI has been steadily growing over the past decade. This research is often, but not always, driven by a belief that AI-based systems will have tangible benefits for the public sector, such as aiding in decision-making, improving efficiency, or improving public service delivery (Kaplan & Haenlein, 2019; OECD, 2019; Medaglia et al., 2021). Yet, AI on its own will not bring about transformation; it is about how AI is used and applied within and by organizations (van Noordt & Misuraca, 2020). Due to the socio-technical nature of AI (Straub et al., 2023), it is essential for academic research to fully consider the context within which it is implemented. Unfortunately, most research on AI today – especially empirical research – is based within the European and Western contexts and governance traditions. What has emerged is a situation where scholars, practitioners, and government officials from non-Western contexts are unable to find or apply academic insight from the scholarly community to their on-the-ground realities due to a contextual disconnect (Masiero, 2023).

This chapter makes a step towards addressing this identified lack of research by presenting the results of a single, empirical, holistic, and exploratory case study conducted within the sub-Saharan African context. Specifically, this chapter follows the initiation, development, and implementation of an AI-based chatbot in Rwanda called Mbaza. The project was conceptualized and developed by the Rwandan public sector and several local and international stakeholders but was co-developed with the support of the German Development Cooperation Agency (GIZ).

To drive the research and analysis of the case, two primary research questions were asked. First, what challenges were encountered during the development and implementation of the AI-based chatbot in the Rwandan public sector? Second, how do the challenges differ from those identified previously in the broader Western

context? By answering these research questions, this chapter makes important contributions to the academic and policy communities interested in the public sector development and use of AI.

14.2 BACKGROUND RESEARCH

Numerous academic studies identify specific challenges or broader categories encountered within the public sector when implementing or using AI-based systems. For example, Wirtz et al. (2019) identify four primary categories of challenges: AI technology implementation, AI law and regulation, AI ethics, and AI society. In 2019, Sun and Medaglia’s research into the adoption of an AI-based system in the healthcare sector identified a total of seven challenges: social, economic, ethical, political, organizational, data, and technological. Building off of the work of Sun and Medaglia, Zuiderwijk et al. (2021) offer a total of eight categories of challenges: data, organizational, skills, interpretation, ethical, political, social, and economic. Writing about AI implementation in African governments, Isagah and Musabila (2020) identify four primary categories of challenges for AI implementation in African governments: (1) data, (2) skills and domain expertise, (3) government, and (4) stakeholders. An overview of these challenges is shown in Table 14.1.

Outside of these identified categories, some authors have highlighted that distinct and important factors have thus far challenged or inhibited the uptake of AI within African governments. These include a dependency on external technical expertise, a high level of outsourcing, and a lack of capacity to oversee contractors (Brunette et al., 2019; Nagitta et al., 2022; Plantinga, 2022) that impede a government’s ability to explore the potential of AI strategically while safeguarding its responsible use. Another challenge is insufficient administrative capacity or institutional voids that hamper the development or regulation of an AI ecosystem. For example, institutional voids may lead to the inadequate provision of information and resources, including

TABLE 14.1  
List of Challenges

Zuiderwijk et al. (2021)	Sun and Medaglia (2019)
	Technological challenges
Data challenges	Data challenges
Organizational and managerial challenges	Organisational and managerial challenges
Skills challenges	
Interpretation challenges	
Ethical and legitimacy challenges	Ethical challenges
Political, legal, and policy challenges	Political, legal, and policy challenges
Social and societal challenges	Social challenges
Economic challenges	Economic challenges



infrastructure, which leads to artificially small markets and a lack of trust (Heeks et al., 2021; Parmigiani & Rivera-Santos, 2015; Wang & Cuervo-Cazurra, 2017), all of which are detrimental to the implementation of AI. As a final point, most AI-based systems are still built in the West or by Western companies, creating a mismatch between the systems themselves and the context they are implemented in (Berman & Tettey, 2001; Heeks, 2011). This challenge is further exacerbated due to a lack of localized training data (Birhane, 2020).

### 14.3 METHODOLOGY

This research has been conducted as a qualitative and exploratory case study, which allows for the in-depth analysis of a phenomenon and the inter-relationships within its real-life context (Yin, 2018). The project selected for this analysis was the Covid-19 Mbaza chatbot developed in Rwanda for the public sector. Mbaza was envisioned as a conversational chatbot that could provide information in both English and Kinyarwanda, was accessible via smart- and feature phones, and provided citizens with up-to-date information on Covid-19. The project relied on community-led technology development and built on previous efforts to collect and open a corpus of Kinyarwanda speech data for developing text-to-speech (TTS) and speech-to-text (STT) models.

Empirically, the chapter relied on semi-structured interviews (nine in total), critical records, policy documents, project documentation, and grey literature. The interviewees were selected to include representatives of all relevant stakeholder groups involved in the project (i.e., government partners, private sector developers, consultants, and aid organizations). The interviews were conducted virtually and in a semi-structured fashion, lasted between 30 to 75 minutes, and were recorded. The questions helped to (a) provide understanding into the origins of the project idea and (b) identify challenges encountered in the different stages of the project and strategies used to overcome them. To analyze the gathered empirical evidence, a multi-step process was utilized. First, the interviews were transcribed, and the identified challenges were coded following an inductive approach based on qualitative content analysis (Hsieh & Shannon, 2005). Following this, the codes generated were categorized according to the categories from Table 14.1.

### 14.4 THE CASE

Inspired by a hackathon in Germany to tackle the Covid-19 pandemic (Wir vs Virus), the Federal Ministry for Economic Cooperation and Development (BMZ), in cooperation with the European Commission and stakeholders from civil society and the private sector, decided to host the *Smart Development Hack* in April and May 2020. The hackathon aimed at tackling the challenges of the Covid-19 pandemic in non-Western contexts by developing digital innovations based on locally defined needs.

The Mbaza chatbot was submitted and selected as one of 20 ideas out of more than 1,000 proposals received during the call for digital solutions (GIZ, 2020).

The idea presented was to develop a voice-based chatbot in English, Kinyarwanda, and Kiswahili that would provide citizens with accurate and safe information on Covid-19 (Niyonkuru, 2020). Drawing on insight provided by the interviewees, the government believed three main issues could be tackled through the use of Mbaza. First, most people, especially in rural areas and without smartphones, had limited access to reliable information. Second, given the limitations of the information channels, the call centre of the responsible authority, the Rwandan Biomedical Centre (RBC), was overburdened with the number of incoming calls (Interview 3LOC2, March 15, 2022). By developing a voice and text-based conversational chatbot that could be accessed via feature phones and smartphones, it was possible to ensure maximum outreach of important trusted information about Covid (Rwandan Biomedical Centre, 2021). Third, there was a large spread of misinformation, which made containing the pandemic even more difficult. By acting as a verified information source, the chatbot would help to temper the spread of such misinformation.

Importantly, for the chatbot project to work, the language in which information on Covid-19 was to be shared was crucial to reaching underserved populations. To strengthen the government-to-citizen communication with the Mbaza chatbot, the local languages needed to be used (Interview 3LOC1, March 9, 2022) to meet the goal of “serving the local people, who most[ly] speak Kinyarwanda” (Interview 2GOV1, April 20, 2022).

Distributing information only in English would exclude and discriminate against most Rwandans. Many are already under-privileged as “a lot of people in Rwanda, like 95% of the population, [are] speaking Kinyarwanda, and then we only have a few fluent English-speaking people” (Interview 4LTP1, March 9, 2022). For the Rwandan government, there was a desire to build their own solution, rather than being dependent on an outside initiative without any control or possibility to steer the work. They explored using the World Health Organisation’s WhatsApp chatbot because the application was offered for free. However, as one interviewee recalled:

[we] realised that the moment the outbreak of Covid [becomes] stable or we’re able to mitigate the risk, all these systems will go away. And we started thinking about having in the background another development, [so that] the moment all these systems go away, at least we have our own system . . . to ensure continuity.

**(Interview 1GOV1, March 31, 2022)**

To reach the goals set by the government, and to develop the solution in Rwanda, the project was broken down into four main components: (1) a rules-based chatbot in Kinyarwanda, English, and French that can be accessed via an Unstructured Supplementary Service Data (USSD) short code; (2) a semantic text-based chatbot in English and Kinyarwanda; (3) a text-to-speech (TTS) conversational chatbot in English and Kinyarwanda that receives text input and returns an audio response generated by a TTS model; (4) a full-voice chatbot involving STT, TTS, and language models. It was furthermore planned to integrate the chatbot engine with additional backend software: (1) the customer relationship management (CRM) system used by the Rwandan government to generate tickets and trace interactions with citizens; (2)

integration with an interactive voice response (IVR) or phone system to route interactions that need human intervention to the RBC call centre; and (3) integration of a business intelligence (BI) solution to identify hot topics and provide decision-makers with timely information.

At the beginning of the third Covid-19 wave in the region in July 2021, the rules-based USSD version of the chatbot was launched. It provided daily statistics, information on symptoms, infection prevention, and current government regulations (Rwandan Biomedical Centre, 2021). After vaccines became available, information about the location of vaccination centres, access to the personal vaccination status, and test results were added. By September 2021, more than 580,000 people had used the chatbot, with around 15,000 interactions daily (GIZ, 2021). By April 2022, the Mbaza chatbot had more than 2,200,000 unique users across the country (Digital Umuganda, 2022).

The semantic chatbot in English was tested and was ready for production, although the Kinyarwanda version still had some loopholes and needs further training and testing (Interview 4LTP2, March 16, 2022). The STT model for Kinyarwanda was trained with 2,300 hours of voice data and 3 million lines of validated text based on Baidu's DeepSpeech model using end-to-end deep learning. It has been made available as open-source voice recognition software under the name "Kinyarwanda DeepSpeech RESTful API". However, the Kinyarwanda DeepSpeech model had a high word error rate (WER) of 60.1% and a character error rate (CER) of 23.5% initially (Meyer & Rutunda, 2021).

According to one interviewee, the WER is now down to 39% (Interview 3LOC1, March 9, 2022), but the model is still "far from being good" (Interview 6DON2, March 8, 2022). According to its model card on Github, the STT model can be used for keyword spotting and simple transcriptions but is not intended for use as a complete voice assistant or voice recognition technology. The developers recommend adding more accents in addition to speakers of main Kinyarwanda accents, using an improved language model that accounts for grammatical errors, and increasing the number of individual voices in the training data to ramp up accuracy, as well as working on a smaller model that can be used on mobile devices (Meyer & Rutunda, 2021).

However, existing language corpora were exhausted. The team now must either collect more training data or try training another model. The first attempts to train a different model with the Kinyarwanda Common Voice dataset by the SpeechBrain community produced promising results, with the WER down to 18.9% (Ravanelli et al., 2021). Google Research trained a model on the Kinyarwanda voice dataset and reached a WER of 9.8% in 2022 (Ritchie et al., 2022). In comparison, the speech synthesis component, which means the TTS model, is in good shape and was integrated into the existing chatbot. However, it needs further improvements to speak naturally (Interview 4LTP1, March 9, 2022). With it, the team managed to build the first-ever Kinyarwanda TTS model (Digital Umuganda, 2022).

## 14.5 DISCUSSION

The Mbaza project represents an important stepping stone for the Rwandan government and its use of AI. Importantly, the project provided a learning opportunity, with several challenges being encountered throughout the development and

implementation process. To understand these challenges, and to discuss their broader implications, it is possible to turn back to the framework provided in Table 14.1. Drawing on this framework, this discussion highlights the challenges identified within the case. The categories included in this analysis are technological challenges, data challenges, skills challenges, organizational and managerial challenges, political, legal, and policy challenges, social challenges, and economic challenges. An overview of the challenges identified in this chapter are further expanded upon in the following subsections.

### 14.5.1 TECHNOLOGICAL CHALLENGES

All stakeholder groups mentioned that they encountered technological challenges during the project. These challenges centred around either (1) the infrastructural readiness of the country or (2) developing voice technology in Kinyarwanda, something that had not been attempted before. One government representative acknowledged that “[we] still have a way to go because we have also to see the readiness of our infrastructure in terms of computation, capability; . . . a lot goes with the readiness of the country in general” (Interview 1GOV1, March 31, 2022).

Developing a conversational chatbot in an under-resourced language came with its own challenges as “[we were] building in Kinyarwanda something that has not been done. That was a very, very big challenge” (Interview 3LOC2, March 15, 2022). In particular, Kinyarwanda lacked rich documentation on how to model the language in computational terms. For instance, no mature tonal dictionary was available, which affected the adoption of existing open-source machine learning frameworks for chatbot development, particularly the development of the speech synthesis part. In simple terms, the chatbot sounds less natural because clear instructions on how to place the tones in Kinyarwanda are missing. “So even the study of the language from the linguistic and vocal linguistic perspective is still in their infancy,” concluded one international consultant (Interview 5ITP1, March 16, 2022).

### 14.5.2 DATA CHALLENGES

Challenges pertaining to the availability, access, quality, and storage of data for AI projects in the public sector were also widespread in the studied case. These data challenges are closely interlinked with technological challenges, as described previously. Interviewees found that the amount of available text and voice datasets were insufficient, negatively affecting all the components of the AI-based chatbot development in Kinyarwanda. Although the Mozilla Common Voice project was the first effort to develop a voice dataset for Kinyarwanda and was therefore very much appreciated, interviewees were disappointed because the project:

stopped a bit short; . . . it was discontinued in a moment in which we were actually expecting that it will continue further. This created an issue reflected most on the speech-to-text part because the dataset that was available up to that moment, through the Common Voice programme, was actually not enough to offer, through the trained model, enough accuracy.

(Interview 5ITP1, March 16, 2022)

However, the interviewees were not consistent on this aspect. Mozilla, as a technical partner, had to leave the project unexpectedly due to internal restructuring. According to another respondent, this did not affect the data collection as such, but the quality of data would have been better with the machine learning fellow and support staff available to the team as “[the] guidelines on how to collect data, analyse data, how to curate the data collected” (Interview 3LOC1, March 9, 2022) became a challenge when Mozilla left.

### 14.5.3 SKILLS CHALLENGES

In Rwanda, skills and capacity challenges were a multi-dimensional challenge. First, a lack of AI knowledge within government institutions and the need for knowledge transfer was mentioned by respondents as a key challenge, especially when it came to long-term maintenance and sustainability of the project (Interview 3LOC2, March 15, 2022; Interview 6DON1, March 16, 2022; 2GOV1, April 20, 2022). Second, NLP and other technical experts, including DevOps specialists, were unavailable in the Rwandan job market.

This challenge was highlighted by respondents from all stakeholder groups directly involved with the technology development (Interview 6DON2, March 8, 2022; Interview 6DON1, March 16, 2022; Interview 3LOC1, March 9, 2022): “AI, DevOps, and the team management and product ownership of product management are roles that were extremely difficult to be found in Rwanda” (Interview 5ITP1, March 16, 2022). Talent did not seem available at all, according to GIZ staff, who concluded that “it was really hard to find those people in Rwanda because they are simply not available” (Interview 6DON1, March 16, 2022), and the local start-up: “[W]e looked for talented people who had been working on a Kinyarwanda chatbot, and . . . there were actually none” (Interview 3LOC2, March 15, 2022).

The lack of skills ultimately led to delays in software development and insecurity around team continuity because people needed much more time for upskilling, which was achieved by an extensive capacity development programme, while some also left the project (Interview 6DON1, March 16, 2022; Interview 6DON2, March 8, 2022).

### 14.5.4 ORGANIZATIONAL AND MANAGERIAL CHALLENGES

Two main organizational and managerial challenges were encountered during the development of the project. The first type is related to general project management, whereas the second type pertains to working in and with the government on an AI project. Challenges of the first type can be considered standard project management difficulties and include unclear roles and responsibilities within the team (Interview 6DON2, March 8, 2022); scoping of the project, which was too large (Interview 4LTP2, March 16, 2022); underestimating the amount of time and resources it takes to deliver specific components, in this case, software architecture and migration (Interview 6DON2, March 8, 2022); and managing the project handover (Interview 6DON2, March 8, 2022; Interview 1GOV1, March 31, 2022).

Other challenges within this subtype were directly linked to the technological challenges of AI as an emerging technology in Rwanda and the problematic skills situation described earlier. It includes slow project implementation due to a lack of human resources and, subsequently, a lack of time to deliver project components on

schedule, as well as insecurity regarding team continuity (Interview 5ITP1, March 16, 2022). As one local consultant recalls: “For this new project, for this new technology, I think the team needed more time to actually learn and implement it in quality” (Interview 4LTP2, March 16, 2022).

Furthermore, there were challenges related to “turf wars” between the project partners RBC and RISA (Interview 3LOC2, March 15, 2022); government ownership of the project, which was slow to obtain from one of the government institutions involved, but crucial for timely feedback and project success (Interview 6DON1, March 16, 2022; Interview 6DON2, March 8, 2022); expectation management towards government partners, who requested new features the team had not planned for (Interview 3LOC1, March 9, 2022); and resistance to sharing data when the team tried to unlock additional sources of data (Interview 3LOC2, March 15, 2022).

#### **14.5.5 POLITICAL, LEGAL, AND POLICY CHALLENGES**

The interviewees described a lack of engagement from government institutions in data sharing, including policies and playing an active role in collecting and making available data to the Rwandan AI ecosystem. If there were “a public institution, like the Ministry of ICT, that facilitates this data collection, then other governmental institutions would share data easily or more comfortably” (Interview 3LOC1, March 9, 2022).

Failure to provide clear guidance on what constitutes personally identifiable information (PII) data under the new Rwandan data privacy law was discussed as an issue because interviewees felt that they were operating in a legal grey area with voice data (Interview 6DON1, March 16, 2022). Considering that voice data might fall under a data protection law, start-up staff also wished for clear standards for anonymization of PII for data collection, which would ease cooperation on dataset creation with other stakeholders (Interview 3LOC1, March 9, 2022). The team was also unsure how to implement the law’s localization requirements concerning training AI models with Rwandan data on cloud servers outside the country (Interview 6DON1, March 16, 2022).

Government representatives also mentioned the inflexibility of procurement law as a barrier. The project was initiated through the hackathon and framed as an emergency response that was expected to help address critical pain points of the Rwandan government and deliver results fast, as one government official recounted: “there was no time for bureaucracy” (Interview 1GOV1, March 31, 2022). This was a state of emergency, authorities knew the challenges, and they needed an immediate solution. However, starting the project in this way ultimately led to numerous challenges, with one government official remarking that “I could say that the many challenges that we faced were really related to how the project started” (Interview 1GOV1, March 31, 2022), especially regarding the provision of infrastructure, including access to data storage facilities.

#### **14.5.6 SOCIAL CHALLENGES**

In Rwanda, interviewees identified the lack of awareness of voice technologies within the population as a hurdle to making the service known, and they feared that people

would push back because the technology was little understood. It was anticipated that people might question the trustworthiness of the voice-based application if the voice synthesis did not equal human-like interaction (Interview 4LTP1, March 9, 2022). Analysis of the project charter also confirmed the concern that uptake of the voice chatbot by the Rwandan population would be low.

#### **14.5.7 ECONOMIC CHALLENGES**

Finally, economic challenges were mentioned regarding the costs (Interview 2GOV1, April 20, 2022; Interview 6DON1, March 16, 2022) as well as the infrastructural needs, including data storage and computing power to train the models (Interview 6DON2, March 8, 2022; Interview 5ITP1, March 16, 2022). Being able to cover the technical costs does not only require substantial funding, in the case of the Mbaza chatbot project, but interviewees also found it challenging to secure funding on time due to the lengthy governmental budgeting process, which collided with the initial underestimation of these costs (Interview 6DON1, March 16, 2022).

Although governmental funding mechanisms have not been explicitly mentioned in the studied literature, it is safe to assume that many AI projects run into this challenge. Cost estimations for technological projects tend to be inaccurate and procurement difficult if only pain points are known and potential solutions little understood (Nagitta et al., 2022). In the sub-Saharan African context, an additional layer of complexity is added by considering budgetary constraints and little leeway to extend budgets beyond the initially approved funding (Isagah & Musabila, 2020; Plantinga, 2022).

### **14.6 CONCLUSION**

The potential benefits of AI have encouraged many governments to seek out and trial AI use cases. However, to do this successfully, there is a need to understand the challenges that may accompany such implementations. Significant research outlines and describes these challenges, but such research often ignores important contextual factors. Through the exploration of the Mbaza chatbot project in Rwanda, it has been possible for this chapter to offer initial insight, supported by empirical evidence, into challenges that public sector organizations in non-Western contexts may encounter when implementing their own AI project.

Specifically, this chapter has identified challenges such as weak technological infrastructure, a lack of AI training data for the African context, and the absence of trained models for some domains, here NLP. The findings show that a lack of access to essential resources, including computational linguistic knowledge, inhibited the development of local AI innovations. Also, insufficient administrative capacity and institutional voids have negatively impacted the Rwandan public sector's ability to implement AI applications or create a vibrant local AI ecosystem. Understanding the skills challenges in environments characterized by human-capital voids is further advanced by a vicious circle where unavailable AI skills, jobs, and education reinforce each other.

These challenges are even more pertinent today in a world where AI is becoming an essential tool. There is a need and desire to develop such systems locally, developing local AI capability, and maintain domestic ownership of these solutions. However, as this case shows, this process can be immensely difficult. There must be an effort to develop the local capacity for AI development, including specialized knowledge in related fields such as computational linguistics, so that, in the long run, reliance on international experts and external funding can be diminished. Most likely, this will involve a mixture of government support via regulation and procurement for development of local solutions, using strategic funding to grow and develop the local AI industry, and providing ample opportunity for experimentation and development of AI home-grown solutions.

## REFERENCES

- Berman, B. J., & Tettey, W. J. (2001). African states, bureaucratic culture and computer fixes. *Public Administration and Development: The International Journal of Management Research and Practice*, 21(1), 1–13.
- Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed: A Journal of Law, Technology and Society*, 17.
- Brunette, R., Klaaren, J., & Nqaba, P. (2019). Reform in the contract state: Embedded directions in public procurement regulation in South Africa. *Development Southern Africa*, 36(4), 537–554.
- Digital Umuganda. (2022). We are glad to showcase the achievements of the Mbaza AI Chatbot project so far. *Twitter*. <https://twitter.com/DUmuganda/status/1509861238486011905>
- GIZ. (2020). *Smart development hack Spring 2020*. Toolkit Digitalisierung. <https://toolkit-digitalisierung.de/en/smartdevelopmenthack-spring-2020/>
- GIZ. (2021, September 14). *COVID-19 chatbot reaches over 500,000 people in Rwanda*. <https://toolkit-digitalisierung.de/en/news/covid-19-chatbot-reaches-over-500000-people-in-rwanda/>
- Heeks, R. (2011). Information systems and developing countries: Failure, success, and local improvisations. *The Information Society*, 18(2), 101–112.
- Heeks, R., Gomez-Morantes, J. E., Graham, M., Howson, K., Mungai, P., Nicholson, B., & Van Belle, J. P. (2021). Digital platforms and institutional voids in developing countries: The case of ride-hailing markets. *World Development*, 145, 105528.
- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288.
- Isagah, T., & Musabila, A. (2020). Recommendations for artificial intelligence implementation in African governments: Results from researchers and practitioners of AI/ML. In Y. Charalabidis, M. A. Cunha, & D. Sarantis (Eds.), *13th international conference on theory and practice of electronic governance (ICEGOV 2020)* (pp. 82–89). Association for Computing Machinery.
- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25.
- Masiero, S. (2023). Decolonising critical information systems research: A subaltern approach. *Information Systems Journal*, 33(2), 299–323.
- Medaglia, R., Gil-Garcia, J. R., & Pardo, T. A. (2021). Artificial intelligence in government: Taking stock and moving forward. *Social Science Computer Review*, 41(1), 123–140.
- Meyer, J. R., & Rutunda, S. (2021, April 2). *Deepspeech Kinyarwanda model card*. <https://github.com/Digital-Umuganda/Deepspeech-Kinyarwanda/blob/master/model-card.md>



- Nagitta, P. O., Mugurusi, G., Obicci, P. A., & Awuor, E. (2022). Human-centered artificial intelligence for the public sector: The gate keeping role of the public procurement professional. *Procedia Computer Science*, 200, 1084–1092.
- Niyonkuru, A. (2020). Mbaza AI based Covid-19 chatbot [Video]. *Youtube*. <https://youtu.be/pr0ikab5Gqg>
- OECD. (2019). *Recommendation of the council on artificial intelligence*. <https://legalinstruments.oecd.org/api/print?id=648&lang=en>
- Parmigiani, A., & Rivera-Santos, M. (2015). Sourcing for the base of the pyramid: Constructing supply chains to address voids in subsistence markets. *Journal of Operations Management*, 33–34, 60–70.
- Plantinga, P. (2022). Digital discretion and public administration in Africa: Implications for the use of artificial intelligence. *SocArXiv*. <https://doi.org/10.31235/OSF.IO/2R98W>
- Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., . . . Bengio, Y. (2021). *SpeechBrain: A general-purpose speech toolkit*. <http://arxiv.org/abs/2106.04624>
- Ritchie, S., Cheng, Y. C., Chen, M., Mathews, R., van Esch, D., Li, B., & Sim, K. C. (2022). Large vocabulary speech recognition for languages of Africa: Multilingual modeling and self-supervised learning. *arXiv preprint*. arXiv:2208.03067
- Rwandan Biomedical Centre. (2021, July 23). *Press release: Rwanda biomedical centre launches RBC*. Mbaza.
- Straub, V. J., Morgan, D., Bright, J., & Margetts, H. (2023). Artificial intelligence in government: Concepts, standards, and a unified framework. *Government Information Quarterly*, 40(4), 101881.
- Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, 36(2), 368–383.
- van Noordt, C., & Misuraca, G. (2020). Exploratory insights on artificial intelligence for government in Europe. *Social Science Computer Review*. <https://doi.org/10.1177/089443932098044>
- Wang, S. L., & Cuervo-Cazurra, A. (2017). Overcoming human capital voids in underdeveloped countries. *Global Strategy Journal*, 7(1), 36–57.
- Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector – applications and challenges. *International Journal of Public Administration*, 42(7), 596–615.
- Yin, R. K. (2018). *Case study research and applications design and methods* (6th ed.). Sage Publications.
- Zuiderwijk, A., Chen, Y. C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, 38(3), 101577.

---

# Index

Note: Page numbers in *italics* indicate a figure and page numbers in **bold** indicate a table on the corresponding page.

## A

- access/accessibility, 40, 57, 137, 196, 203
  - control (AC), 97–99
  - of farmers, 120, 123, 124
  - to markets, 132
- accountability, 3, 12–13, 54, 51–52, 57–58,  
*see also* algorithmic accountability
  - AI Accountability Framework, 7
  - concerns, **173**
- age-appropriate design, 18
- agency, 12, 74
  - government, **14–15**, 16, 89
- agri-advisory, 120, 124, 125
- agricultural productivity, 119
  - agriculture extension and advisory services (AEAS), 119–120
  - technological advancements, 120
- AI4Bharat, 122
- AI adoption, 1–2, 137, *see also* healthcare
  - AI adoption
- AI ecosystem, 101, *101*
- AI ethics, 18, 65, 69, 145, 148–150
  - Ekitia example, 159–166
  - ethical principles, 157
  - French public sector challenges, 158–159
  - in health sector, 162–163
  - in HR sector, 163–165
  - implementation tools, 161–162
  - methodology, 155–158
- AI governance, 6, 160–161
  - auditing AI systems, 192
  - employer data, 19
  - generative, 48–50
  - guidance, P3119, 20, 20
  - IEEE Standards, 18
  - responsible, *see* algorithm registers
- AI harm, 192
  - discrimination, hate speech and exclusion, 194–195
  - environmental and socio-economic, 197
  - human-computer interaction, 196–197
  - information hazards, 196
  - malicious use risks, 196
  - misinformation, 196
  - ML unfairness, India, 195, **195**
- AI Procurement Lab (AIPL), 10, 25–26
- AI risk, 190–191
  - governance measures, 192
  - technical measures, 191, **191**
- algorithmic accountability, 187
  - AI harm, 192, 194–197
  - Foundation Model Transparency Index, 189
  - internationally accepted principles, **194**
  - policies, **193**
- algorithmic auditing, 186
  - crowdsourced/collaborative audit, 188–189
  - noninvasive user audit, 188
  - protocol, 187–188
  - scraping audit, 188
  - sock puppet audit, 188
  - third-party, 189, 197–198
- algorithmic biases, 136–138
- algorithmic impact assessments, 186–187, 189–190, **190**
- algorithm registers, 51–52, 58–59
  - accountability, 54, 57–58
  - Amsterdam City Algorithmic Register, 53
  - categories, 52, 54
  - citizen experience, 56–57
  - creators, 52–53
  - Dutch government, 53–54
  - France, 53
  - information quality/quantity, 55–56
  - New York City, 53
  - non-experts accessibility, 57
  - policy makers, 58–59
  - recommendations on, 52–55
  - transparency, 51, 59
- algorithm transparency, 188
- Ama KrushAI, 121–122
- Amsterdam City Algorithmic Register, 53
- AnalyzeMyXrays (AMX), 36–37
- anonymization, 207
- applicant tracking system (ATS), 49, 50
- Asia-Pacific Economic Cooperation (APEC)
  - region, *see* scaling AI, APEC region
- automated decision systems (ADS), 2–6
  - Article 41 – Right to Good Administration, 3, 4
  - global use, 5, 5
  - IEEE P3119, *see* P3119 standard
  - IEEE Standards Association, 5–6
  - risks, 2–3
  - Royal Commission Report, 5

autonomous and intelligent systems (A/IS),  
19–20, **182**, 191  
autonomous driving system (ADS), 2–7, 96  
  NYC task force for, 11–13  
autonomy, 152

## B

Bard, 147  
best practices, 7, 174  
  in procurement, 25, 26  
Bhashini, 119–121  
bias, **173**  
  algorithmic, 136–138, 140  
  data, 69, 153  
  and fairness, 192, 194–**195**  
  stakeholders, 149  
black box problem, 88, 138  
brain drain, 133

## C

Centre for Security and Emerging Technologies  
  (CSET), 132  
chatbots, 120  
  add-on features, 125  
  Ama KrushAI, 121–122  
  Bard, 147  
  Bhashini, 119–121  
  ChatGPT, 122, 131, 134, 147  
  Claude, 147, 152  
  collaboration and partnerships, 125–126  
  contextual relevance and trust, 124  
  contextual training, 125  
  data protection and privacy, 126  
  digital connectivity and behavioral gaps, 124  
  farmers' needs and services, 126  
  human-in-the-loop, 125  
  Jugalbandi, 122  
  Kisan e-Mitra, 122–123  
  KissanAI, 123  
  linguistic diversity and training  
    challenges, 123  
  Mbaza, *see* Mbaza chatbot project  
  resource availability constraints, 124–125  
  speech-based interfaces, 123–124  
ChatGPT, 122, 131, 134, 147  
citizen(s)  
  and Ekitia, 160  
  empowerment through AI, *see* algorithm  
    registers  
  potential harms to, 2–6  
Claude, 147, 152  
code audit, 188  
commercialization, 88  
confidential clean rooms, 35–36, 38, 40  
confidential computing, 39–42, **41**

conformity assessments, 6, 19  
contextual training, 125  
continental-scale AI models training, *46*  
continuous monitoring, **90**, 196, 197  
contract, 3, 187  
  digital, 36  
  electronic, 42–43  
  P3119, 24  
COVID-19 pandemic  
  AMX, 36  
  disease surveillance, *45*  
  Mbaza chatbot, 202–203  
  mRNA vaccines, 170  
Cray X-MP14 supercomputer, 131  
crowdsourced/collaborative audit, 188–189  
cybersecurity, 88–89  
  AI ecosystem, 101, *101*  
  C2M2, 90–91, *91*  
  insurance determination, 92, 100  
  insurance premiums, 92  
  liability determination, 102  
  NIST RMF model, 90, **90**  
  regulatory recommendation, 100–101  
  risk quantification, 92–99  
  valuation of business, 92  
cybersecurity capability maturity model (C2M2),  
  90–91, *91*

## D

data breach, 92, 100, 153  
data collection, 123–124, 135, 187, 207  
data empowerment and protection architecture  
  (DEPA)  
  AMX, 36–37  
  confidential computing, 39–42, **41**  
  continental-scale AI models training, *46*  
  COVID disease surveillance, *45*  
  DP, 38–39  
  ecosystem, 44–45, *45*  
  electronic contracts, 42–43  
  foundation, 35–36  
  reference implementation, *46*  
  technical implementation, 37–39, **38**  
  techno-legal framework, 43–44, *44*  
data governance, 19  
data labelling, 129, 140  
data privacy, 19, 36, 126, 170, 196, 207  
  DP, 38–39  
data processing, 165  
data protection frameworks, 44, 126  
Data Science Africa (DSA), 139–140  
datasets  
  algorithmic bias, 136–138  
  contract services, 42  
  DP, 38–39  
  X-ray images, 36

- deep learning
    - black box, 138
    - Deep Learning Indaba, 139
    - DeepSpeech model, 204
    - diabetic retinopathy, 137
    - LLMs, 35
  - Defense Advanced Research Projects Agency (DARPA), 139
  - DEPA, *see* data empowerment and protection architecture (DEPA)
  - diabetic retinopathy (DR) screening, 62–63
    - diagnostic performance, 68, **68**
    - IDx-DR, 63
    - image grading, 66
    - patients and providers, 68
    - performance, 67–68
  - diagnostic models
    - C2M2, 90–91, 91
    - classification framework dimensions/criteria
      - OECD, 93, **94–95**
      - NIST RMF model, 90, **90**
      - Z-Inspection, 144–150
  - Differentially Private Stochastic Gradient Descent (DP-SGD), 39
  - differential privacy (DP), 38–39, 196
  - digital connectivity, 124
  - digital public infrastructure (DPI), *see* data empowerment and protection architecture (DEPA)
  - diversity, 49
    - in Global South, 136–137
    - linguistic, 123
    - in team expertise, 9–10
  - DP, *see* differential privacy (DP)
- E**
- Ekitia
    - awareness and training, 160
    - criteria, **164**, 165
    - employment and training, 163–165
    - Ethical Charter for Data Use, 160, 161
    - governance, 160–161
    - MyData Global, 160
    - public health, 162–163, **162–163**
    - public-private ecosystem, 159
    - think-and-do tank approach, 160
  - electronic contracts, 42–43
  - Electronic Privacy Information Center (EPIC), 2, 7
  - equity, 16, 84, 157, *see also* inequity, Global North and South
  - ethics, *see* AI ethics
  - European Committee for Electrotechnical Standardization (CENELEC), 181, 183
  - European Committee for Standardization (CEN), 181, 183
  - European Food Safety Authority (EFSA), 16
  - European Telecommunications Standards Institute (ETSI), 183
  - European Union (EU)
    - AI Act, 183
    - Charter for Fundamental Human Rights, 3, 4
    - EFSA, 15–16
    - and Ekitia, 161
    - trustworthy AI framework, 150
  - explainability, 88
- F**
- fairness, 136, 157, 188
  - fairness accountability and transparency (FAT), **173**
  - fiduciary duty, 8
  - Food and Agriculture Organisation (FAO), 137
  - foreign-born STEM workers, 133
  - Foundation Model Transparency Index, 189
- G**
- generative AI
    - ATS data, 49, 50
    - datasets, 48
    - diversity and monoculture, 49
    - LLMs, 48–50
    - Z-Inspection, *see* Z-Inspection
  - generative pre-trained transformer (GPT) technology, 121
  - Global Partnership for Artificial Intelligence (GPAI), 88
  - Global South, *see* inequity, Global North and South
  - governance, *see* AI governance
  - government, *see also* algorithm registers
    - and AI procurement, 1–2
    - and automation, 2–6
    - expertise, 10
    - and procurement landscape, 6–8
    - of Rwanda and AI capacity, 200–208
  - greenfield, 11, 25
- H**
- healthcare, 45, 135
    - AMX, 36–37
  - healthcare AI adoption, 62–63, 69
    - dark room, 65
    - data management and analysis, 67
    - data modification, 67
    - DR screening, 62–63
    - ethical approval and trial registration, 65
    - fundus imaging, 66
    - hardware and software, 65–66
    - image acquisition, segregation and analysis, 66, 67

image grading process, 66–67  
 limitations, 68  
 patients and providers, 68  
 performance, 67–68  
 principles, 63, **64**  
 recommendation, 69  
 recruitment, study participants, 65  
 reliability and transparency, 66  
 sample size, 65  
 staff hiring and training, 65  
 strengths, 68  
 study site, 65  
 higher education, *see* Z-Inspection  
 high-risk AI  
   and ADS, 2–4  
   AIPL, 25–26  
   applications and systems examples, **34**  
   transdisciplinary collaboration, 8–10, 9  
 human-centric design, 18  
 human-in-the-loop, 125  
 human rights, 3–4, 100  
  
**I**  
 IDx-DR, 63  
 IEEE CertifAIED™, 20  
 IEEE P3119, *see* P3119 standard  
 IEEE Standards Association (IEEE-SA), 17–18  
   age-appropriate design, 18  
   AI governance, 18  
   A/IS, 19–20  
   autonomous systems, transparency, 18  
   data privacy, 19  
   employer data governance, 19  
   ethical values, 18  
   IEEE CertifAIED™, 20  
   robotics and automation systems, 19  
 impact, 100  
   A/IS, 19–20  
   algorithmic impact assessments, 186–187,  
     189–190, **190**  
   cybersecurity standardization, 92  
   environmental, 157  
   and global inequities, 131–132  
 import substitution industrialisation (ISI)  
   policies, 132  
 inclusion  
   ethical frameworks, 69  
   public participation, 72  
   stakeholders, 161  
 inequity, Global North and South,  
   84, 129–130  
   algorithm biases, 136–138  
   global challenges, 130–131  
   impact, 131–132  
   labour from Global South, 133–134  
   national strategies, 134–136

  overcoming, 139–140  
   regulatory challenges for Global South,  
     138–139  
 innovation, 3, 168  
   institutional, 13  
   STEM workers, 133  
   UK guidelines, 16  
 Institute of Electrical and Electronic Engineers  
   Standards Association (IEEE SA),  
     181, **182**  
 insurance premiums, 92  
 Internal Revenue Service (IRS), 3  
 International Data Corporation (IDC), 2  
 international standards, 169, 171  
   AI policies and regulation, 175  
   businesses and governments, 175  
   CEN-CENELEC, 181, 183  
   ETSI, 183  
   IEEE publications and initiatives, 181, **182**  
   ISO/IEC JTC 1/SC 42 (SC 42), 177, 177, 178,  
     **178–180**  
   NIST focus areas, 181, 183  
   operationalising and scaling AI, 176, 176  
   responsible development and deployment, 175  
   role, 174–175, 184  
   strength of, 174  
 ISO/IEC JTC 1/SC 42 (SC 42)  
   CEN-CENELEC, 181, 183  
   ecosystem approach, 177, 177  
   publications, scaling AI, **178–180**  
   standards map, 177, 178

**J**  
 job applications screening, 49  
 Jugalbandi, 122

**K**  
 Kisan e-Mitra, 122–123  
 KissanAI, 123

**L**  
 language models (LMs) risk areas, 194–197  
 large language models (LLMs), 35, 37, 148, 152,  
   *see also* chatbots  
   architectures, 48–50  
   environment and socio-economic  
     factors, 197  
   LLaMA-2–7B and LLaMA-2–13B models,  
     49–50  
   three-layered approach, 198  
 linguistic barriers, 120, 122, 123  
 literacy  
   among public servants, 55, 59  
   speech-based interfaces, 123–124

**M**

machine learning (ML), 76  
 AMX, 36–37  
 DEPA for training, *see* data empowerment and protection architecture (DEPA)  
 text-to-image tools, 136  
 Uli project, *see* Uli ML model

maturity model, 90–91, 91

Mbaza chatbot project, 202–204  
 data challenges, 205–206  
 economic challenges, 208  
 Kinyarwanda voice database, 204  
 learning opportunity, 204–205  
 organizational and managerial challenges, 206–207  
 political, legal and policy challenges, 207  
 skills challenges, 206  
 social challenges, 207–208  
 STT model, 204  
 technological challenges, 205

**N**

National Institute of Standards and Technology (NIST), 44  
 C2M2 maturity levels, **91**, 92  
 focus areas, 181, 183  
 RMF framework, 90, **90**  
 scorecard, 91, 91

natural language processing (NLP), 35, 79, 206

neural networks, 39, 88

NITI Aayog, 69

non-discrimination, 157

non-governmental organizations (NGOs), 126

noninvasive user audit, 188

NYC Automated Decision Task Force, 11–13

**O**

OECD (Organization for Economic Cooperation and Development), 1–2, 10, 175  
 classification of AI systems, 93, 93, **94–95**  
 G20 principles, 6  
 national AI strategies, 130  
 transdisciplinary collaboration, 9

opacity, 83–84

Open AI, 131

open government, 53

**P**

P3119 standard, 3, 7–8  
 AI governance guidance, 20, 20  
 contracting, 24  
 EFSA and UK LGA, 15–16  
 five processes, 22, 22–23

guidance and tools, 11, 20  
 local, national, transnational briefings, 14, **14–15**  
 post-procurement, 24–25  
 pre-procurement, 23  
 process component structure, 21  
 procurement/RFP/solicitation, 23–24  
 risk management, 25  
 risks and advantages, 22

participation, 72  
 inequality, 84  
 instrumental argument, 73  
 interest, 72  
 normative goals, 73  
 spaces, 74  
 Uli project, *see* Uli ML model

partnerships  
 GPAI, 88  
 IEEE AI governance, 18  
 NGOs, 126  
 public-private, 16

performance, 63, 67–68, **68**, 157

personalized advisories, 120

pervasiveness, 87–88

policy, 1, 140, 146, 156, 167, 169–170, *see also*  
 algorithm registers  
 algorithmic accountability, **193**  
 contract registration, 42  
 cybersecurity regulations, 89  
 international standards, 175, 184  
 ISI, 132  
 mechanisms, 13  
 Pre-Legislative Consultative Policy, India, 73  
 science, technology and innovation  
 policies, 135  
 uniformity in regulation, 6

Pradhan Mantri Kisan Samman Nidhi (PM-KISAN) program, 122–123

privacy by design, 37

privacy impact assessments (PIAs), 19

procurement, 1–2, 10  
 AI governance and guidance frameworks, 6–7, 7  
 AIPL, 25–26  
 overselling AI, 6–8  
 IEEE-SA, 17–20  
 NYC Automated Decision Task Force, 11–13  
 P3119, 8, 20–25  
 sandbox testing, 14–17  
 transdisciplinary collaboration, 8–11  
 unaccountable AI, 2–6

public good, 121

public investment, 2

public policy, *see* policy

public sector, 1, 8  
 algorithmic accountability practices, 51–52, 54, 57

collaboration with private sector, 10, 13  
 French AI integration experience, 155–159,  
*see also* Ekitia  
 responsible AI, 6  
 Rwanda's experience, 200–208

## R

REDCap, 67  
 regulation, *see* policy  
 Regulatory Oversight Committee, 153  
 reliability, 66  
   adoption, 63  
   on AI model for gendered abuse detection,  
     *see* Uli ML model  
   AIPL, 26  
   APEC region, 169–171, **173–174**, 186–184  
   in education, 148  
   inequality, 136, 138–140  
   and international standards, 174–175  
   privately financed, 130–131, 132  
   procurement, 25–26  
   research, 4, 68, 69, 192, 201–202  
   responsible AI, 6, 158, 167, *see also*  
     algorithmic accountability  
 risk assessment, 16, 25, *see also* AI harm; AI  
   risk; high-risk AI  
   cybersecurity, 92–99  
   insurance, 99–100  
   liability determination, 102  
   pervasiveness, 87–88  
   quantification, *see* risk quantification  
   regulatory recommendation, 100–101  
   transformative appeal, 88  
 risk management  
   P3119 standard, 25  
   RMF, 90, **90**  
 risk management framework (RMF), 90, **90**  
 risk quantification  
   AI system dimensions, 93, 93  
   classification framework dimensions/criteria  
     OECD, 93, **94–95**  
   net/scaled score, **99**, 99  
   risk scoring, facial recognition, 96–98, 97, **98**, **99**  
   weights, 96  
 risk scoring, facial recognition (FR), 96–98, 97,  
   **98**, **99**  
 robotics and automation systems, 19  
 robustness  
   DEPA training framework, 39  
   governance frameworks, 192  
   ML algorithm, 36  
   P3119 standard, 16

## S

safe/safety, 7, 48, 56, 192, 208, *see also* AI harm;  
 AI risk; high-risk AI

DEPA for ML, 37  
 EFSA, 16  
   and security, 157, 168  
   and trust, 176  
 sandbox  
   benchmarking, 16–17, 17  
   EFSA and UK LGA, 16  
   P3119 process, 14, **14–15**  
   regulatory, 14  
 scaling AI, APEC region  
   ABAC report, 168  
   barriers, 171  
   benefits, 168  
   business leaders and policymakers, 170  
   challenges, 171, 172, **172–173**  
   decision-making processes, 167  
   Deloitte report, 170  
   international standards, *see* international  
     standards  
   McKinsey report, 170  
   mRNA vaccines, 170  
   opportunities, 169  
   social and economic potential, 169  
 science, technology and innovation policies, 135  
 security, *see* cybersecurity; safe/safety  
 self-preservation, 131, 138  
 social harm, *see* AI harm  
 socio-technical scenarios, 151  
   actors, 151  
   context and processes, 151–152  
   human oversight and decision-making, 152  
   liability, 153–154  
   participants, 154  
   primary aim, 151  
   WG education professors/teachers, 152–153  
 South-South cooperation (SSC), 137–138  
 speech-to-text (STT), 203, 204  
 standard, *see* international standards; P3119  
   standard  
 sustainable development, 11

## T

Target Group Analysis, 55  
 technical measures, auditing AI  
   systems, 191, **191**  
 techno-legal framework, DEPA, 43–44, 44  
 testing  
   dataset, 63, 76, 78  
   sandboxes, 14–15  
 text-to-speech (TTS), 122, 203, 204  
 third-party algorithmic auditing, 189, 197–198  
 training, 78–79, 123, 163  
   AI labour, 133–134  
   contextual, 125  
   DEPA, *see* data empowerment and protection  
     architecture (DEPA)  
   LLMs, 197

- staff hiring, 65
  - and validation, 63
- training data consumers (TDCs), 38, 39, 40, 43
- training data providers (TDPs), 38, 39, 40, 43
- transdisciplinary collaboration, 8–10, 9
- transformation, 88, 158, 170
- transformers, 121
- translation, 79–80
  - annotation guidelines, 81
  - community and AI development team, 80, 80–81
  - false positives/negatives, ML model, 81
  - horizontal and vertical, 82
  - layers of, 82
  - metaphor, 83
  - online abuse, 81
  - partial connections, 83
- transparency, 12, 13, 66
  - AI ecosystem, 101
  - algorithm, 188
  - autonomous systems, 18
  - Foundation Model Transparency Index, 189
  - principle, 157
- trusted execution environments (TEEs), 40
- trustworthiness, 18, 145
  - EU framework, 150
  - requirements and sub-requirements, 146, **146**
  - socio-technical scenarios, 151–154

## U

- UK Local Government Association (UK LGA), 16
- Uli ML model

- annotations, 78
  - annotators, 76–77
  - data collection and corpus creation, 76
  - gendered abuse, 77–78
  - opacity, 83–84
  - research, 74–75
  - training, 78–79
  - translation, 79–83, 80
- unified communications interface (UCI), 121

## V

- Vienna Agreement, 181

## W

- women, 78, 131, 137
- working groups (WGs), 14, 149–150
  - education professors/teachers, 152–153
  - P3119, 20

## Z

- Z-Inspection, 144
  - assessment, 147–148
  - ethical principles, 145–146
  - generative AI platforms, 147
  - guidelines, 145
  - interdisciplinary team creation, 149–150
  - phases, 144–145, *145*, 148, *148*
  - requirements/sub-requirements of trustworthy AI, 146, **146**
  - socio-technical scenarios, 150–154
  - teaching faculty, 147