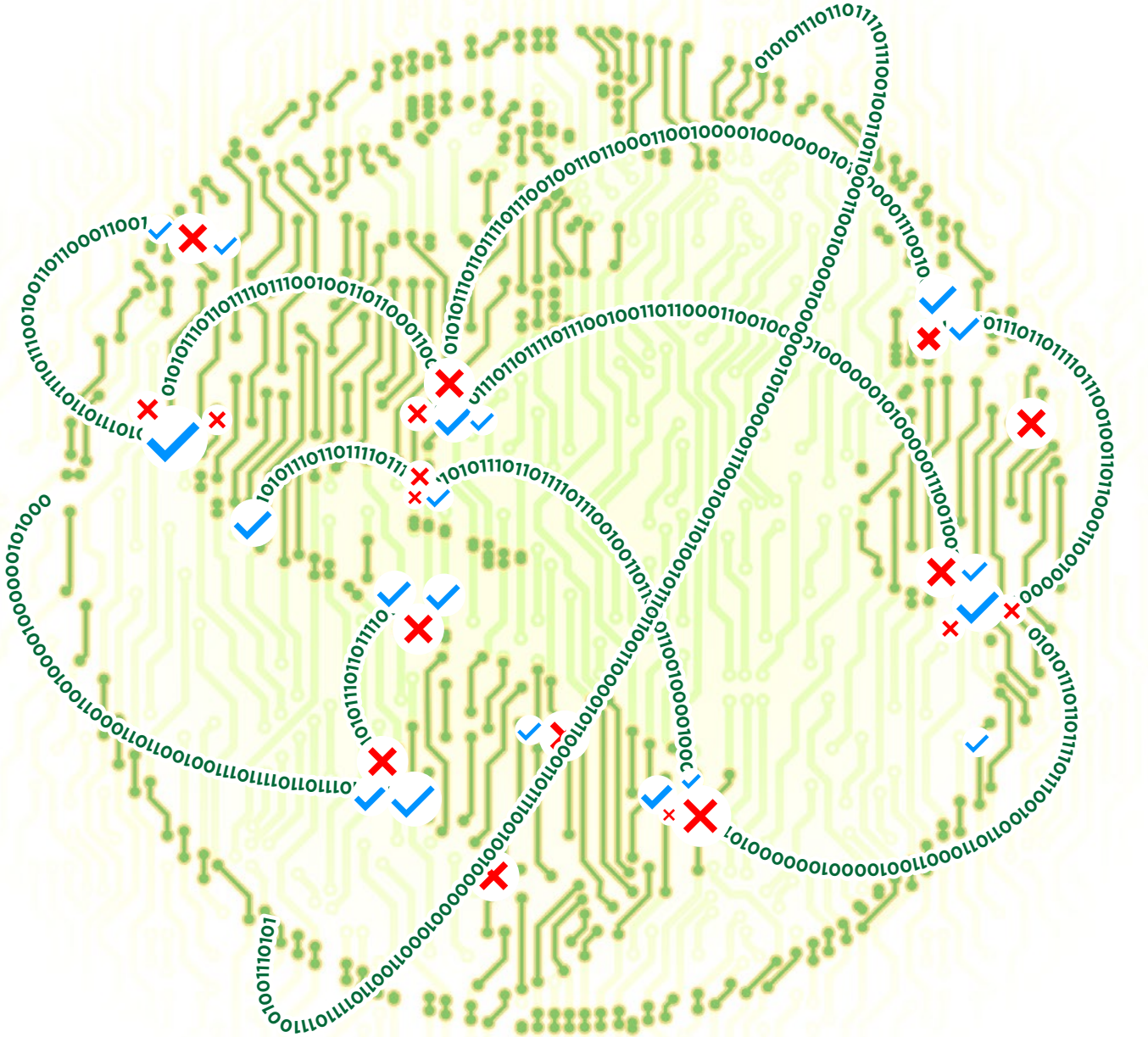
WORLD **PRIVACY** FORUM

Risky Analysis

Assessing and Improving AI Governance Tools

An international review of AI Governance Tools and suggestions for pathways forward



Kate Kaye Pam Dixon

December 2023

Brief Summary of Report

AI systems should not be deployed without simultaneously evaluating the potential adverse impacts of such systems and mitigating their risks. Most of the world agrees about the need to take precautions against the threats posed by AI systems. Tools and techniques exist to evaluate and measure AI systems for their inclusiveness, fairness, explainability, privacy, safety and other trustworthiness issues. These tools and techniques – called here collectively AI governance tools – can improve such issues. While some AI governance tools provide reassurance to the public and to regulators, the tools too often lack meaningful oversight and quality assessments. Incomplete or ineffective AI governance tools can create a false sense of confidence, cause unintended problems, and generally undermine the promise of AI systems. This report addresses the need for improved AI governance tools.

It is the goal of this research to help gather evidence that will assist in the building of a more reliable body of AI governance tools. This report analyses, investigates, and appraises AI governance tools, including *practical guidance*, s*elf assessment questionnaires*, *process frameworks*, *technical frameworks*, *technical code*, and *software* disseminated in Africa, Asia, North America, Europe, South America, Australia and New Zealand. The report also analyzes existing frameworks, such as data governance and privacy, and how they integrate into the AI ecosystem. In addition to an extensive survey of AI governance tools, the research presents use cases discussing the contours of specific risks. The research and analysis for this report connects many layers of the AI ecosystem, including policy, standards, scholarly and technical literature, government regulations, and best practices.

Our work found that AI governance tools used in most regions of the world for measuring and reducing risks and negative impacts of AI could introduce novel, unintended problems or create a false sense of confidence unless accompanied by evaluation and measurement of those tools and their effectiveness and accuracy.

In this report we suggest pathways for creating a healthy AI governance tools environment, and offer suggestions for governments, multilateral organizations, and others creating or publishing AI governance tools. These suggestions include best practices taken from existing AI and other quality assessment standards and practices already in widespread use. Appropriate procedural and administrative controls include: 1) providing AI governance tool documentation and contextualization, review, audit, and other quality assurance procedures to prevent integration of inappropriate or ineffective methods in policy guidance; 2) identifying and preventing conflicts of interest; and 3) ensuring that capabilities and functionality of AI governance tools align with policy goals. If governments, multilateral institutions, and others working with or creating AI governance tools can incorporate lessons learned from other mature fields such as data governance and quality assessment, the result will establish a healthier body of AI governance tools, and over time, healthier and more trustworthy AI ecosystems.

About the Authors

(Listed alphabetically)

Pam Dixon is the founder and executive director of the World Privacy Forum, a respected nonprofit, non-partisan, public interest research group. An author and researcher, she has written influential studies in the area of identity, AI, health, and complex data ecosystems and their governance for more than 20 years. Dixon has worked extensively on data governance and privacy across multiple jurisdictions, including the US, India, Africa, Asia, the EU, and additional jurisdictions. Her field research on India's Aadhaar identity ecosystem, peer-reviewed and published in Nature Springer, was cited in India's landmark Aadhaar Privacy Supreme Court opinion. Dixon currently serves as the co-chair of the UN Statistics Data Governance and Legal Frameworks working group, and is co-chair of WHO's Research, Academic, and Technical network. At OECD, Dixon is a member of the OECD.AI Network of Experts and serves in multiple expert groups, including the AI Futures group. In prior work at OECD, Dixon was part of the original AI expert group that crafted the OECD AI Principles, which were ratified in 2019. Dixon has presented her work on complex data ecosystems governance to the The National Academies of Sciences, Engineering, and Medicine and to the Royal Academies of Science. She is the author of nine books and numerous studies and articles, and she serves on the editorial board of the Journal of Technology Science, a Harvard-based publication. Dixon was named one of the most influential global experts in digital identity in 2021. Dixon received the Electronic Frontier Foundation Pioneer Award in 2021 for her ongoing oeuvre of groundbreaking research regarding privacy and data ecosystems.

Kate Kaye, deputy director of the World Privacy Forum, is a researcher, author, and award-winning journalist. She is a member of the OECD.AI Network of Experts, where she contributes to the Expert Group on AI Risk and Accountability. In addition to her extensive research and reporting on data and AI, Kate is the recipient of the Montreal AI Ethics Institute Research Internship, and was a member of UNHCR's Hive Data Advisory Board.

Editor: Robert Gellman

About The World Privacy Forum

The World Privacy Forum is a respected NGO and non-partisan public interest research group focused on conducting research and analysis in the area of privacy and complex data ecosystems and their governance, including in the areas of identity, AI, health, and others. WPF works extensively on privacy and governance across multiple jurisdictions, including the US, India, Africa, Asia, the EU, and additional jurisdictions. For more than 20 years WPF has written in-depth, influential studies, including groundbreaking research regarding systemic medical identity theft, India's Aadhaar identity ecosystem —peer-reviewed work which was cited in the landmark Aadhaar Privacy Opinion of the Indian Supreme Court — and The Scoring of America, an early and influential report on machine learning and consumer scores. WPF co-chairs the UN Statistics Data Governance and Legal Frameworks working group, and is co-chair of the WHO Research, Academia, and Technical Constituency. At OECD, WPF researchers participate in the OECD.AI AI Expert Groups, among other activities. WPF participated as part of the first core group of AI experts that collaborated to write the OECD Recommendation on Artificial Intelligence, now widely viewed as the leading normative principles regarding AI. WPF research on complex data ecosystems governance has been presented at the National Academies of Science and the Royal Academies of Science. World Privacy Forum: https://www.worldprivacyforum.org.[1]

1 World Privacy Forum's home page includes information about our activities, as well as numerous data governance and privacy research, data visualizations, and resources. https://www.worldprivacyforum.org.

Risky Analysis: Assessing and Improving AI Governance Tools

An international review of AI Governance Tools and suggestions for pathways forward

World Privacy Forum

www.worldprivacyforum.org

Cover and design by John Emerson

EBook/Digital: ISBN: 978-0-9914500-2-2

Publication Date: November 2023

Nothing in this material constitutes legal advice.

This report is available free of charge at: **https://www.worldprivacyforum.org/2023/12/risky-analysis-assessing-and-improving-ai-governance-tools**

Updates to the report will be made at : **https://www.worldprivacyforum.org/2023/12/risky-analysis-assessing-and-improving-ai-governance-tools**

Acknowledgements

The World Privacy Forum extends our gratitude to the following people who reviewed and/or were interviewed for this report:

Report Reviewers

(Listed alphabetically)

Ryan Calo, Lane Powell and D. Wayne Gittinger professor, University of Washington School of Law; adjunct professor, UW's Paul G. Allen School of Computer Science and Engineering; and faculty co-founder, UW Tech Policy Lab

Abhishek Gupta, founder and principal researcher at the Montreal AI Ethics Institute

Cristina Pombo Rivera, principal advisor and fAIr LAC coordinator, Social Sector, Inter-American Development Bank

Jason Tamara Widjaja, director of artificial intelligence and responsible AI lead, Singapore Tech Center, MSD

Report Sources

(Listed alphabetically)

Dr. Rumman Chowdhury, CEO, Humane Intelligence

Maria Paz Hermosilla Cornejo, director of GobLab UAI, in the School of Government at Adolfo Ibáñez University, Chile

Abigail Jacobs, assistant professor of information, School of Information, and assistant professor of complex systems, College of Literature, Science, and the Arts, University of Michigan

Lizzie Kumar, PhD candidate in computer science at the Brown University Center for Tech Responsibility

Tim Miller, professor in artificial intelligence at the School of Electrical Engineering and Computer Science at The University of Queensland

Luca Nannini, PhD student in the Information Technology Research PhD program at CiTIUS of the University of Santiago de Compostela, Spain

Ndapewa Onyothi Wilhelmina Nekoto, research and community lead at the Masakhane Research Foundation

Abigail Oppong, independent researcher and Masakhane Research Foundation contributor

Eike Petersen, postdoctoral researcher, DTU Compute, Technical University of Denmark

Cristina Pombo Rivera, principal advisor and fAIr LAC coordinator, Social Sector, Inter-American Development Bank

Cynthia Rudin, Earl D. McLean, Jr. professor of Computer Science and Electrical and Computer Engineering, Duke University

Ian Rutherford, statistician, United Nations Statistics Division

Jason Tamara Widjaja, director of artificial intelligence and responsible AI lead, Singapore Tech Center, MSD (known as Merck and Co. in the US and Canada).

Jane K. Winn, professor of law, University of Washington School of Law

Executive Summary:

Why the World Privacy Forum Conducted This Research

The World Privacy Forum conducted the research, writing, and background work necessary to complete this report to address the risks posed by profound changes and advances in the AI ecosystem. These changes and advances impact people, groups of people, and communities, and require evidence-based policy responses as soon as possible. Currently, there is a meaningful lack of evidence regarding how to implement and ensure trustworthy AI; this is true for older AI systems, and it is also true for newer, more advanced AI systems.

This report is intended to begin building much-needed evidence and procedures regarding how to implement trustworthy AI by analyzing AI governance tools and their functions. This is a critically important task because AI governance tools form a pivotal component of AI systems and their lifecycle. This report defines AI governance tools as:

> "Socio-technical tools for mapping, measuring, or managing AI systems and their risks in a manner that operationalizes or implements trustworthy AI."[2]

This report also documents what these tools do, where they are located and used, their range of maturity, some of the specific risks they pose, the practices currently in place in relation to these tools, and initial steps to take to begin creating improvements.

AI governance tools are important as an area of focus because they sit at the implementation layer of the AI ecosystem and operate across AI system types. AI governance tools, when they function well, can assist the people, businesses, governments, and organizations implementing AI or researching AI to delve into various aspects of how AI models are functioning, and if they are performing in expected or intended ways. For example, some AI governance tools are meant to measure fairness or to "de-bias" AI systems. Some AI governance tools are meant to explain AI system outputs. And some AI governance tools are designed to measure and improve system robustness, among other tasks. However, when AI governance tools do not function well, they can exacerbate existing problems with AI systems.

The timing of this report is noteworthy. In 2007, WPF began work on an extensively researched report on machine learning and its impacts, *The Scoring of America*. Published in 2014, the report articulated the problems with the deep machine learning of the time and discussed why policymakers needed to address problems with bias, transparency, interpretability, fairness, and other issues. It would have been impossible to know that in just a few years, groundbreaking research introducing a new approach to AI network architecture[3] would begin to evolve AI and its capabilities in novel ways.[4] In a sense, WPF's 2014 publication marked the last years of an earlier deep learning AI era. In contrast, this report, *Risky Analysis*—while building on WPF's earlier AI work—sits at the start of a burgeoning new era in AI.

As such, *Risky Analysis* is a different kind of report. It documents the existing evidence regarding AI governance tools with an intent to begin building the larger evidentiary repository needed to create an evaluation environment that supports a transparent and healthy body of AI tools, which will in turn facilitate a healthier AI ecosystem. WPF intends to continue building on this work. For these reasons, WPF is treating this report as a living document, which WPF will update on a regular basis.

---

2 This definition excludes statutes, regulations, and common law.

3 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lion James, Aidan N. Gomez, Lukas Kaiser, Illia Polosukhin, *Attention is all you need*, arXiv:1706.03762v7 [cs.CL], https://doi.org/10.48550/arXiv.1706.03762. This paper was written by eight individuals, who at the time were Google researchers in various capacities. The paper was first presented at the 31st Conference on Neural Information Processing Systems (NIPS 22017), Long Beach, CA, USA.

4 Madhumita Murgia, *Generative AI exists because of the transformer, This is how it: Writes, Works, Learns, Thinks and Hallucinates*, Financial Times, (September 11, 2023) https://ig.ft.com/generative-ai/.

Research Conducted for This Report: Building an Evidence Base Regarding AI Governance Tools

This report surveys the international landscape of AI governance tools and provides an early evidentiary foundation that documents multiple aspects of these tools. The report focuses on AI governance tools published by multilateral organizations and by governments. The report utilizes the evidence from the survey of tools in conjunction with in-depth case studies and scholarly literature review to construct a lexicon of AI governance tool types. Based on the evidence gathered, these tool types include: *practical guidance*, *self assessment questionnaires*, *process frameworks*, *technical frameworks, technical code*, and *software*.

Figure 1: ***AI Governance Tool Types Lexicon***

Practical Guidance	Includes general educational information, practical guidance, or other consideration factors
Self-assessment Questions	Includes assessment questions or detailed questionnaire
Procedural Framework	Includes process steps or suggested workflow for AI system assessments and/or improvements
Technical Framework	Includes technical methods or detailed technical process guidance or steps
Technical Code or Software	Includes technical methods, including use of specific code or software
Scoring or Classification Output	Includes criteria for determining a classification, or a mechanism for producing a quantifiable score or rating reflecting a particular aspect of an AI system

Source: World Privacy Forum, Research: Kate Kaye, Pam Dixon. Image: John Emerson.

For more information regarding methodology guiding the evaluation of AI governance tools and Finding 2—Some AI governance tools feature off-label, unsuitable, or out-of-context uses of measurement methods—see Appendix C.

The survey of AI governance tools in this report includes tools from each region. Some examples include:

- An updated process for acquisition of public sector AI from Chile's public procurement directorate, ChileCompra
- Self-assessment-based scoring systems from the Governments of Canada and Dubai and Kwame Nkrumah University of Science and Technology in Ghana
- Software and a technical testing framework from Singapore's Infocomm Media Development Authority
- An AI risk management framework from the US National Institute of Standards and Technology (NIST)
- A culturally-sensitive process for reducing risk and protecting data privacy throughout the lifecycle of an algorithm from New Zealand's Ministry for Social Development
- A vast repository of AI governance tool types from The Organization of Economic Cooperation and Development (OECD), a multilateral institution

The report also includes two detailed case studies in AI fairness and explainability and provides suggestions for how to begin building an evidence basis for an AI governance ecosystem. Going forward, AI governance tools, when fit for purpose, can help provide better health at the implementation layer of AI. AI governance tools are nascent, flexible, and when designed and applied for their intended purpose, can improve the health of the AI ecosystem. However, much more work needs to be done to build the evidentiary basis to create an evaluation environment for this ecosystem.

This report analyzes and discusses the existing governance structures in place today that are intended to protect privacy and govern large data ecosystems by facilitating effective and trustworthy management of data flows. This report discusses how data privacy and governance regulations that were installed as recently as 10 years ago no longer retain the same fit and effectiveness in some AI environments, particularly in environments where advanced versions of AI are in use. The research and analysis conducted for this report indicates that we do not yet know what will be effective replacement or evolutionary structures to protect privacy and govern data in an advanced AI era. It is essential to gather the evidence now for what will work, and to develop an AI governance ecosystem based on this and other evidence.

The report research includes analysis of what could be helpful to create improvements in the AI ecosystem. The research found that there are many unknowns regarding AI governance tools, and what standards, methods, or measurements could best be applied to create transparency and a basis for AI that is trustworthy. The research found hopeful avenues and places to start; these include use of the Plan-Do-Study/Check-Act cycle to improve management of AI governance tools and working to improve documentation of AI governance tools. The report also suggests an adaptation of an early AI governance tools framework from the OECD to provide improved gatekeeper functions for entities that publish collections of AI governance tools.

Going Forward

The research for *Risky Analysis* indicates that an evaluative environment in which AI governance tools can be tested, matured, and validated will require an evidentiary foundation. The work to create this foundation is just beginning and will require multistakeholder cooperation. The World Privacy Forum is committed to continuing to do the work necessary to gather and analyze the evidence that will facilitate the building of an AI governance ecosystem that is based on evidence, protects privacy and other values, is trustworthy, and provides a genuine foundation for regulatory structures and implementation practices that are fit for purpose today and for the coming AI era.

Index

Brief Summary of Report 1

Executive Summary: Why the World Privacy Forum Conducted This Research 5

List of Figures 9

Background and Introduction 10

Methodology 18

Findings 18

Part I: Discussion: Critical Analysis of AI Governance Tools 23

Measuring AI Fairness Measures 24

Use Cases in AI Fairness 25

Pathways for Building an Evaluation Environment and Creating Improvements in the AI Governance Tools Ecosystem 36

PART II: A Survey of AI Governance Tools and Other Notable AI Governance Efforts from around the World 47

Intergovernmental Organization Toolkits and Use Cases from International and Regional Multilateral Institutions 51

Regional Development Banks 57

AI Governance Tools and Use Cases from National Governments and NGOs 60

Appendix A: AI Governance Tool Types Lexicon 87

Appendix B: AI Governance Tools and Features Comparison Chart 88

Appendix C: Some AI Governance Tools Feature Off-label, Unsuitable, or Out-of-context Uses of Measurement Methods 90

Appendix D: OECD Catalog of Tools and Metrics Framework 95

List of Figures

Figure 1: *AI Governance Tool Types Lexicon*
(Source: World Privacy Forum, Research: Kate Kaye, Pam Dixon. Image: John Emerson. 6

Figure 2: *Table of Global Privacy Laws*
(Source: World Privacy Forum. Research: Pam Dixon, Kate Kaye. Data Visualization: John Emerson. 14

Figure 3: *Table of Global Privacy Laws, Regional Breakout*
(Source: World Privacy Forum. Research: Pam Dixon, Kate Kaye. Data Visualization: John Emerson. 14

Figure 4: *OECD Catalogue of Tools and Metrics Framework*
(Source: OECD) 42

Figure 5: *AI Governance Tool Types and Features Comparison Chart*
(Source: World Privacy Forum, Research: Kate Kaye, Pam Dixon. Image/Data Visualization: John Emerson.) 50

Figure 6: *Cross-Validation Model*
(Source: Inter-American Development Bank, Responsible Use of AI for Public Policy: Data Science Toolkit) 59

Figure 7: *Dimensions of Explainable AI Tools Usability*
(Source: Partnership on AI) 71

Figure 8: *Financial AI and Data Analytics Model Management*
(Source: Monetary Authority of Singapore) 73

Figure 9: *DEEP MAX Scorecard*
(Source: Tamil Nadu Information Technology Department) 80

Figure 10: *AI System Ethics Self-Assessment Tool Report*
(Source: Government of Dubai) 82

Figure 11: *Model Development Lifecycle Operational Algorithm Governance Structure*
(Source: New Zealand Ministry of Social Development) 86

Figure 12: *AI Governance Tools Including Off-Label Measures*
(Source: World Privacy Forum, Research: Kate Kaye, Pam Dixon. 92

Figure 13: *OECD Catalogue of Tools and Metrics Framework, full size*
(Source: OECD). 95

Background and Introduction

AI governance tools are important because they can map, measure, and manage complex AI governance challenges, particularly at the level of practical implementation. The tools are intended to remove bias from AI systems,[5] or increase the explainability of AI systems, among other tasks. Seeking an orderly, automated way of solving complex problems in AI systems can create efficiencies. But those same efficiencies, if not well-understood and appropriately constrained, can themselves exacerbate existing problems in systems and in some cases create new ones. This is the case with AI governance tools, an important and nascent part of AI ecosystems which this report defines as:

> **AI Governance Tools:**
>
> Socio-technical tools for mapping, measuring, or managing AI systems and their risks in a manner that operationalizes or implements trustworthy AI.[6] [7]

An AI governance tool can be used to evaluate, score, audit, classify, or improve an AI system, its decision outputs, or the impacts of those outputs. These tools come in many forms. This report classifies AI governance tools in the following categories: *practical guidance*, *self assessment questionnaires*, *process frameworks*, *technical frameworks*, *technical code*, and *software*.

While AI governance tools offer the promise of improving the understanding of various aspects of AI systems or their implementations, not all AI governance tools accomplish the goals of mapping, measuring, or managing AI systems and their risks, which we argue are essential features of an effective AI governance tool. Further, given the lack of systematic guidance, procedures, or oversight for their context, use, and interpretation, AI governance tools can be utilized improperly or out of context, creating the potential for errors ranging from small to significant.

For example, AI governance tools can be used in novel or "off-label"[8] ways, which can lead to meaningful errors in contextualization and interpretation. Some of the more complex AI governance tools can create additional risk by

5 An *AI system* is defined in the NIST AI Risk Management Framework as: "An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. Al systems are designed to operate with varying levels of autonomy (Adapted from: OECD Recommendation on AI:2019; ISO/IEC 22989:2022)." *See*: NIST AI RMF, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.The OECD has updated its definition of an AI System as of 2023. The new definition is: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." Definition available at: *OECD AI Principles Overview*, OECD, https://oecd.ai/en/ai-principles.

6 The definition for *AI governance tools* was developed by the authors of this report at the World Privacy Forum. It is based on the research for this report, the scholarly literature, and consultation with a wide range of technical, standards, legal, and policy experts. This definition maps to the OECD AI Principles, the National Institutes of Standards and Technology Trustworthy and Responsible AI principles, and the general outlines of the EU AI Act. The definition was finalized November 10, 2023 in Paris, France.

7 The definition for *AI governance tools* excludes statutes, regulations, and common law.

8 The term "off label use" originally stemmed from the practice in clinical settings of using prescription drugs in a way that differs from what is approved by the FDA and printed on the original prescription label. In the AI context, "off-label" refers to the practice of taking a technology that was created for one context, and using it in another outside of the original use case. NIST mentions "off label use" in its AI Risk Management Framework: "...existing frameworks and guidance are unable to....consider risks associated with third-party AI technologies, transfer learning, and off-label use where AI systems may be trained for decision-making outside an organization's security controls or trained in one domain and then "fine-tuned" for another." *NIST AI Risk Management Framework,* National Institute of Standards and Technology, Feb. 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. p. 39. In a study of off-label use of imaging databases, a National Academy of Sciences study found that the practice could lead to bias in AI algorithms. See: Efrat Shimron, Jonathan I. Tamir, Ke Wang, and Michael Lustig, *Implicit data crimes: Machine learning bias arising from misuse of public data*, March 21, 2022. PNAS, https://doi.org/10.1073/pnas.2117203119. And finally, increased risk is also associated with the term as used in its clinical context. See: Rebecca Dresser and Joel Trader, *Off-label prescribing: A call for heightened professional and governmental oversight*, Journal of Law and Medical Ethics, 2009 Fall: 37(3) 476-396. doi: 10.1111/j.1748-720X.2009.00408.x. "The potential for harm is greatest when an off-label use lacks a solid evidentiary basis. A 2006 study examining prescribing practices for 169 commonly prescribed drugs found high rates of off-label use with little or no scientific support."

producing a rating or score that in and of itself can be subject to error or misinterpretation, especially if there is a lack of documentation and guidance for use of the tool. All told, flawed usage and interpretation can result in a gap between what people want these tools to accomplish, and what these tools actually do accomplish.

Lessons learned from earlier AI policy implementations

AI governance tools can exacerbate risk when there are gaps in controls and standards for the tools, their context of use, and other items, such as their outputs or results. The many types of AI governance tools can range from simple questionnaires all the way to software. Therefore, the manner in which some AI governance tools are used to map, measure, and manage the risks of AI systems can differ substantially. The quality of an AI governance tool and any quantifications it produces matters greatly. For instance, an AI output in the form of a score can seem deceptively simple to interpret. However, AI score outputs are historically notorious for requiring great care and proper contextualization to accurately interpret and apply.[9] This is an important area to understand because challenges with today's AI governance tool outputs, which are in some cases scores, reflect a long history of broader AI scoring approaches (of which there are many) that for decades have been studied, understood, and in some cases, regulated.[10]

Some AI governance tools, such as those intended to remove bias from an AI system, may provide a score or rating to indicate the prominence or presence of certain biases. Not all AI governance tools produce quantified measures or scores. However, in such cases that they do produce scores, it will be necessary to properly interpret the score, and to validate the score for the specific context in which it is used with objective criteria. Accomplishing these kinds of tasks consistently across the full body of AI governance tools and tool types requires a range of policy guidance from informal technical and policy guidance to formal legal guidance or regulation, depending on the tool being used, its output, and the context and purpose of its use.

Using examples from the classical AI machine learning context, some countries regulate AI systems that specifically impact decisions related to eligibility, including AI systems that automate decision-making associated with credit reporting. AI systems trained to analyze credit eligibility often produce a *credit score* as an output. As mentioned, some of these systems currently have regulations in place. This is true across multiple jurisdictions.[11] Most credit scoring regulations are intended to provide transparency regarding automated decisions, error correction, and redress, among other features of credit scoring systems. In regulated scoring models, scores are evaluated for fit, accuracy, and other factors based on the evidence. Evaluative techniques and processes guide proper implementation of the scoring systems.[12] There is enough history and established policy around credit scoring systems and their risks, that credit scores, their use, and their interpretation is well-understood. For example, errors or problems introduced by flaws in either the data, the analysis, or even the implementation or interpretation of the scoring can create meaningful impacts for people, groups of people, and communities.[13]

---

9 Camille Olivia Little, Debolina Halder Lina, and Genevera Allen, *Fair feature importance scores for interpreting tree-based methods and surrogates,* Rice University (Department of Computer Science, Department of Statistics, Department of Electrical and Computer Engineering) October 9 2023. https://arxiv.org/pdf/2310.04352.pdf.

10 Note: Credit scoring regulations enacted in the 1970s are among the oldest and most salient exemplars of existing AI model regulations. See: Fair Credit Reporting Act, 15 U.S.C. § 1681 (US).

11 Credit score regulations are widespread, with regional differences. *See also* Fair Credit Reporting Act, 15 U.S.C. § 1681 (US). Administrative Guideline for Credit Reporting Business (China). Personal Data Protection, 18.331 (Uruguay).

12 Pam Dixon & Robert Gellman, *The Scoring of America*, World Privacy Forum, 42-80 (Apr. 2, 2014), https://www.worldprivacyforum.org/2014/04/wpf-report-the-scoring-of-america-how-secret-consumer-scores-threaten-your-privacy-and-your-future/ (Examples of scoring types include financial and risk scores, fraud scores, identity and authentication scores, smart grid and energy scores, social scoring, law enforcement and judicial scores; cited pages provide descriptions of exemplars of each score type). Compare for example, the legal protections in place for credit scores and the lack of legal protections in place for other scores, such as scores reflecting level of poverty and wealth. See Pam Dixon & Robert Gellman, *The Scoring of America*, at 42-80.

13 *Consumer Response Annual Report, January 1- December 31 2022*, US Consumer Financial Protection Bureau, March 2023. https://files.consumerfinance.gov/f/documents/cfpb_2022-consumer-response-annual-report_2023-03.pdf.

Credit scoring systems and regulations provide many lessons. However, there is an additional policy lesson here beyond just the importance of regulating credit scoring: many thousands of other AI scoring systems exist, and most of these are not formally regulated.[14] In fact, regulated scores are rare. *The Scoring of America* analyzed many kinds of AI scores beyond credit scoring—from patient frailty scores to consumer prominence scores indicating purchasing power to identity scores used to quantify the validity of an identity.[15]

The 2014 *Scoring* report found that unregulated AI scores of that time operated a lot like analytical plumbing, humming along in the background of many business processes. These scores are plentiful and largely unseen, but nevertheless could have an impact on bias, fairness, privacy, transparency, and other issues. The lack of broader policy action regarding the various AI scoring systems that were not in scope of credit scoring regulation resulted in widespread and opaque use of a variety of scores spanning a range of risk levels.

Today, AI governance tools are in an intriguingly similar position in that they, too, are increasingly common and are poised to become a critical part of the evolution of the "AI analytical plumbing," despite the fact that they are often not subject to evidence-based assessments or regulation. Even though AI governance tools are nascent and largely unregulated, they are already in widespread use across the world, and across sectors. For example, some AI governance tools are already in use in eligibility contexts, such as to measure AI systems used in relation to employment. While legal scholars who research these fields know about some of these tools, their uses, and their risks, the knowledge is not yet widespread.[16]

Additionally, because AI governance tools are often made available with minimal documentation and have little to no regulation, these tools exist in various stages of quality assurance. Many new AI governance tools are in development today. But their risks, and the policies that will address these risks consistently are largely not yet well-developed, or in some cases, not present at all. Given the beneficial potential of AI governance tools,[17] it is worth meaningful efforts to understand more about them, how they operate in today's AI environments, and how the AI governance tools environment can be improved while incorporating existing legal and policy guardrails from other regulatory regimes.

Addressing policy disruptions stemming from old and new forms of AI: incorporating lessons learned from the data governance and privacy domain

Human rights, privacy, and data governance laws and policies are currently in various stages of change as a result of AI. Some older AI systems that emerged in past decades have had regulatory oversight in place for many years, for example, credit scoring models were regulated beginning as early as the 1970s. However, the emergence of advanced AI models [18] is creating novel disruptions, and questions abound about what new regulations for newer models should look like. Even though existing approaches are not necessarily responsive to the changes in AI or are being bypassed, there is still much to be learned from the past history of data governance and privacy.

14 Pam Dixon & Robert Gellman, *The Scoring of America*, World Privacy Forum, 42-80 (Apr. 2, 2014), https://www.worldprivacyforum.org/2014/04/wpf-report-the-scoring-of-america-how-secret-consumer-scores-threaten-your-privacy-and-your-future/.

15 Pam Dixon & Robert Gellman, *supra* note 13.

16 For a detailed discussion of two specific AI governance tools use cases, see Part I of this report, *Discussion: Critical Analysis of AI Governance Tools.*

17 For a nuanced discussion of the benefits and risks of newer forms of AI, see Elham Tabassi, *Minimizing harms and maximizing the potential of generative AI*, Taking Measure Blog, NIST, Nov. 20, 2023. https://www.nist.gov/blogs/taking-measure/minimizing-harms-and-maximizing-potential-generative-ai.

18 A type of machine learning model of note in AI is the "transformer." The transformer AI model is considered by AI scientists to be a significant evolutionary advancement. The creator of AlexNet wrote of the development of transformers that "It's a milestone by any measure, if not an inflection point..." Fei-Fei Li, *My North Star for the future of AI*, The Atlantic, Nov. 7 2023, excerpted from Fei-Fei Li, The Worlds I see: Curiosity, Exploration, and Discovery at the Dawn of AI, Flatiron Books, Nov. 2023. Regarding transformer models and generative AI, *see also*: Madhumita Murgia, *Generative AI exists because of the transformer, This is how it: Writes, Works, Learns, Thinks and Hallucinates*, Financial Times, (September 11, 2023) https://ig.ft.com/generative-ai/.

We begin with a discussion of terminology. Data governance and privacy are related, but they are not interchangeable. Data governance is a comprehensive approach to the entirety of data of an organization or entity that ensures the information is managed through the full data lifecycle. This can include data collection practices, data security, quality, documentation, classification, lineage, cataloging, auditing, sharing, and other aspects. Data privacy is a subset of data governance and is best defined in context as forms of protecting either personal data, or the personal data of a group of people. Privacy is often seen in terms of individual data rights, such as the right to deletion, and so forth. While the conception of privacy as an individual right is currently ascendant in terms of legislation today,[19] conceptions of privacy as a group or community-based privacy right are emerging as well, and can be found, for example, in Māori approaches to privacy.[20]

In response to changes in technology, new opportunities, and other developments, data governance policy, laws, and institutions were introduced, developed, and adjusted mightily over the decades. These evolutions were driven by an urgent need to respond to then-radical changes in technology and governance policy. The impetus was to provide responses to a variety of emerging threats and opportunities related first to the emergence of the computer, and later, to the emergence of the Internet and subsequent factors, such as the emergence of social media platforms.

In the late 1960s, attention to data governance, data protection, and privacy began slowly, with small developments here and there around the world. Responding to the growth of personal computing, countries enacted different privacy laws beginning in the 1970s and 1980s.[21] It did not take long before the differences and limits in these national laws created problems with international data flows. Europe began to address these problems, and the EU, after some significant effort, adopted a Data Protection Directive [22] in the 1990s. The shortcomings of the Directive and the challenges with its implementation resulted in its replacement by the EU General Data Protection Regulation[23] which has been enforced since 2018. Many other countries around the world now follow the EU privacy model. There is no question GDPR forms a near-worldwide regulatory structure.[24]

---

19 For example, a European-influenced articulation of individual privacy may be seen in OECD's Recommendation on Privacy (the Fair Information Practice Principles). A full articulation of the European approach may be seen in Directive 95/46/EC, 1995 O.J. (L 281) 31 and in the current EU General Data Protection Regulation.

20 Te Mana Raraunga, the Māori Data Sovereignty Network, https://www.temanararaunga.maori.nz. See also: First Nations Information Governance Centre, *The First Nations Principles of OCAP,* https://fnigc.ca/ocap-training/ (establishes how First Nations' data and information will be collected, protected, used, or shared). For a general discussion of privacy, See Kenneth A. Bamberger and Deirdre K. Mulligan, *Privacy on the books and on the ground*, Stanford Law Review, Vol. 63, January 2011. UC Berkeley Public Law Research Paper No. 1568385. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1568385.

21 The state of Hesse, Germany passed a federal law that regulated automated data processing in the public sector on 7 October 1970. (Bundesdatenschutzgesetz, or BDSG.) Full text: https://www.gesetze-im-internet.de/bdsg_2018/index.html . In 1970, the US passed its first major privacy law, the Fair Credit Reporting Act, which also is among the first laws to regulate machine learning. Full text: https://www.ecfr.gov/current/title-12/chapter-X/part-1022 . Other laws followed in the EU and the US. In 1981, the EU opened its *Convention 108* for signature by EU members, and by other countries. Full text and list of signatories: https://coe.int/en/web/data-protection/convention108-and-protocol . In the 1990s, the EU passed its landmark Data Protection Directive EU 95/46, https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=celex:31995L0046 . More than 160 jurisdictions across the world now have some form of data governance / data protection legislation mostly following the pattern of the EU General Data Protection Regulation. Full text: (EU) 2016/679 (GDPR), Full text: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504&qid=1532348683434. The uptake of the GDPR comprises a mature and nearly global regulatory footprint although significant differences in policy and implementation remain.

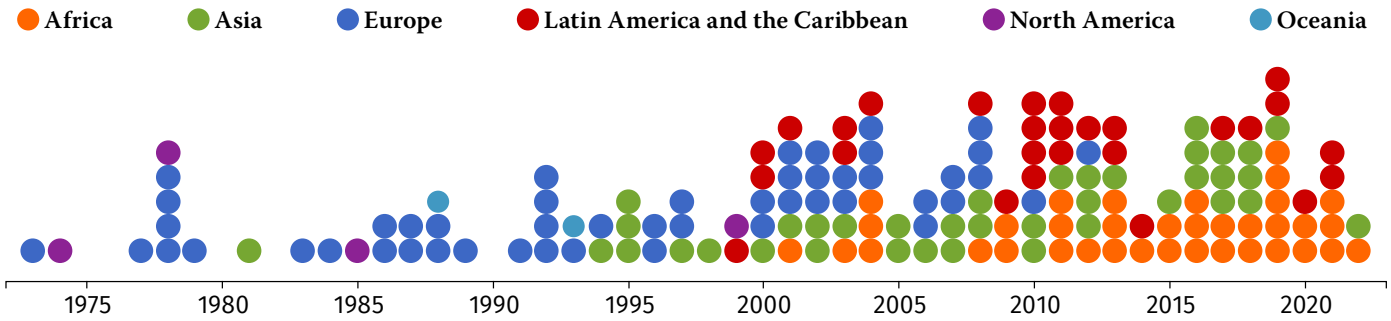
22 Directive 95/46/EC, 1995 O.J. (L 281) 31

23 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, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1

24 Another important data protection governance instrument is Convention 108, and its update, Convention 108+. The Council of Europe crafted Conv. 108 and 108+ so that it could be ratified by countries outside of Europe. See: *Convention 108* for signature by EU members, and by other countries. Full text and list of signatories: https://coe.int/en/web/data-protection/convention108-and-protocol. See also: *Convention 108 and Protocols:* https://www.coe.int/en/web/data-protection/convention108-and-protocol.

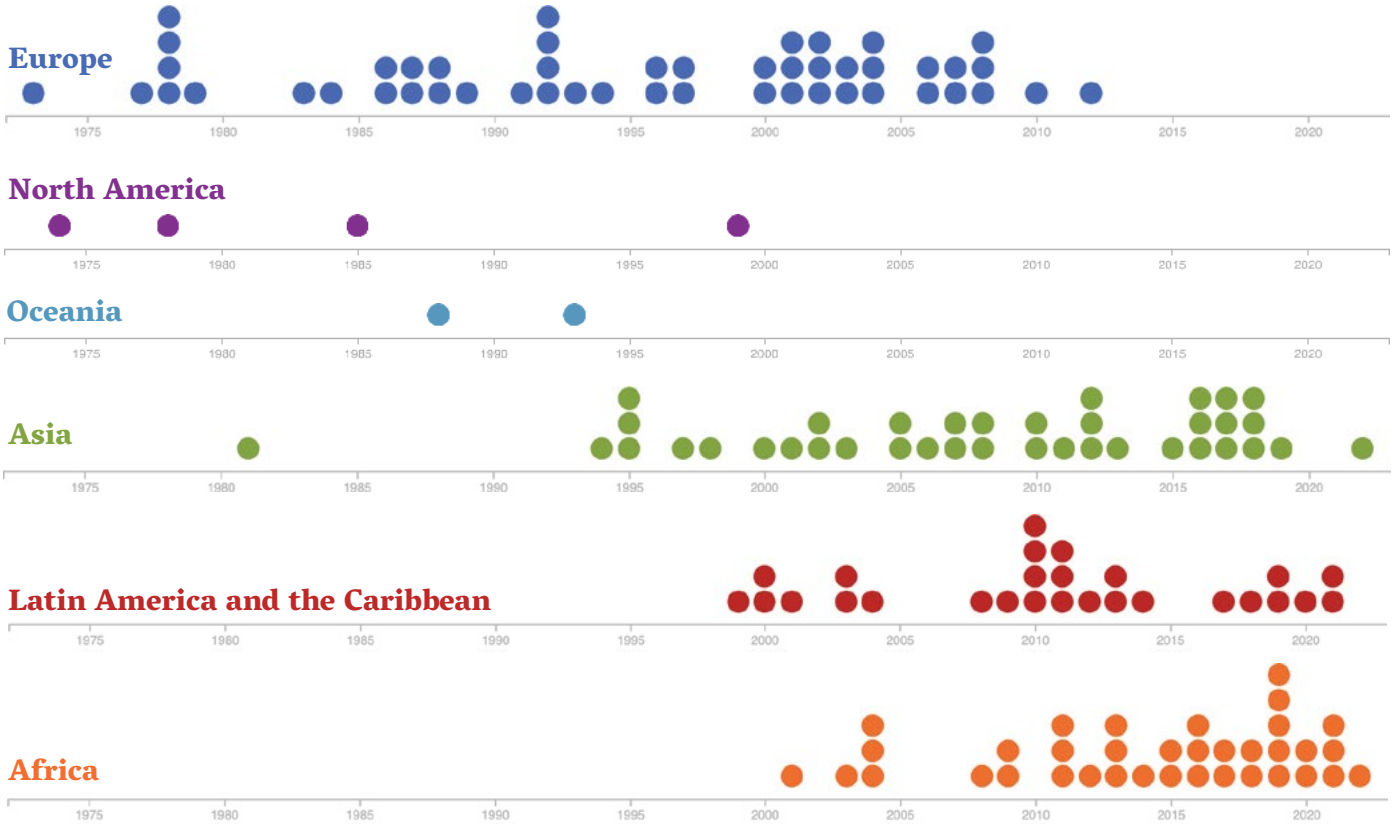
See for example, Figures 1 and 2, which visualize the distribution of data governance and data protection laws across the world. Notice the distinct patterns of distribution of data protection laws through global regions over time.

Figure 2: Table of Global Privacy Laws



Source: World Privacy Forum. Research: Pam Dixon, Kate Kaye. Data Visualization: John Emerson.

Figure 3: Table of Global Privacy Laws, Regional Breakout



Source: World Privacy Forum. Research: Pam Dixon, Kate Kaye. Data Visualization: John Emerson.

For decades data policy and technical developments came about more or less at approximately the same time, albeit with some delay, and at a much slower pace. For example, at about the 20-year mark of early credit model development, credit score models were regulated.[25] These first major credit system regulations developed in the

25 This statement refers to the US context for credit regulation. While US-based Fair & Isaac had developed their credit scoring model in the 1950s, it wasn't until the 1970s that AI credit models were regulated in the US. See *Scoring of America*, *supra note 13.*

1970s, at the beginning of a series of worldwide data governance and privacy developments that then unfolded in incremental steps over decades.[26]

Fair Information Practices (FIPs),[27] the core statement of data governance and privacy values started in 1973 in the United States, was restated by the Organization of Economic Cooperation and Development (OECD) in 1980,[28] and became the basis for many privacy laws and policies around the world. Eventually, FIPs faded into the background, not because the policies were wrong, but because the general policies that served so well for so long were not specific enough to address ongoing developments in technology, industry, and government. To offer one example, FIPs did not call for privacy agencies, but countries quickly recognized the value of privacy agencies or data protection authorities, and the idea spread around the world. Data protection authorities function as enforcers of data protection and governance laws, and help guide the implementation data governance ecosystems at the ground level effectively.[29]

As privacy laws and institutions matured, it became clear over time that solutions which had seemed responsive in theory did not always work well in practice, or, sometimes, ideas that worked in one jurisdiction or social context did not fit in others. For example, GDPR and GDPR-like legislation, which grew from the 1995 EU Data Protection Directive, both of which focused on individual privacy rights, does not always fit well in some contexts, including Indigenous contexts, where privacy and data are often handled as community rights.[30] [31] Additional ideas from jurisdictions and stakeholders came along. There was a lot of experimentation, which created an evidentiary basis over time. The data governance and privacy learning curve stretched over decades, and the various stakeholders in the data ecosystems are still learning. Similarly, AI regulations have been moving along for the most part in incremental steps.[32]

---

26 See *supra* note 19 for details of developments.

27 Robert Gellman, *Fair Information Practices: A basic history*, BobGellman.com, Version 2.22 (Apr. 6, 2022), https://bobgellman.com/rg-docs/rg-FIPShistory.pdf.

28 *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, OECD (Feb. 12, 2002), https://doi.org/10.1787/9789264196391-en.

29 *See generally* Global Privacy Assembly, https://globalprivacyassembly.org, (the Assembly is comprised of the international data protection and privacy commissioners or authorities. They met the first time in 1979). *See also* Irish Data Protection Commission, https://www.dataprotection.ie; Data Protection Office Mauritius, https://dataprotection.govmu.org/SitePages/Index.aspx; Personal Information Protection Commission Japan, https://www.ppc.go.jp/en/; Office of the Privacy Commissioner for New Zealand (Te Mana Matapono Matatapu), https://www.privacy.org.nz/privacy-act-2020/privacy-principles (examples of the work of data protection authorities).

30 There have been significant advances in regards to the data rights of Indigenous peoples. This extends to the rights of Indigenous people to develop their own methods of data governance, which can, depending on context, grant community-level privacy rights which operate substantially differently than individual privacy rights enshrined in the GDPR. These contextual differences have meaningful implications for AI governance tools and their use. *See: United Nations Declaration on the Rights of Indigenous Peoples,* United Nations, General Assembly Res 61/295 art. 18 (Sept. 13, 2007) http://www.un-documents.net/a61r295.htm.(provides Indigenous peoples' right to participate in decision-making in matters which would affect their rights, through representatives chosen by them in accordance with their own procedures, as well as to maintain and develop their own Indigenous decision-making institutions). *See also:* First Nations Information Governance Centre, *The First Nations Principles of OCAP,* https://fnigc.ca/ocap-training/ (establishes how First Nations' data and information will be collected, protected, used, or shared). S*ee als*o Te Mana Raraunga, the Māori Data Sovereignty Network, https://www.temanararaunga.maori.nz;

31 Michael Pisa, Pam Dixon, Benno Ndulu, Ugonma Nwankwo, *Governing Data for Development: Trends, Challenges, and Opportunities*, Center for Global Development, November 12, 2020. https://www.cgdev.org/publication/governing-data-development-trends-challenges-and-opportunities.

32 For example, the Fair Credit Reporting Act, originally passed in the 1970s, is updated in various ways, including regulatory updates, to incorporate changes and advances in AI and policy understanding periodically. *Small business advisory review panel for consumer reporting rulemaking: Outline of proposals and alternatives under consideration*, Consumer Financial Protection Bureau September 15 2023. https://files.consumerfinance.gov/f/documents/cfpb_consumer-reporting-rule-sbrefa_outline-of-proposals.pdf.

In contrast to the long evolution of data governance and privacy laws and norms, advanced forms of AI, though in development since 2017, jumped to public awareness seemingly overnight in 2022.[33] The presence of newer and more advanced AI models brought quick regulatory reactions and proposals that in some cases derogated from decades of established knowledge and lessons. Quick response is on the whole hopeful. However, responses will need to ensure there is input from all stakeholders, and ensure that existing legal and other guardrails, including existing human rights and privacy guardrails, are integrated. One lesson from data governance and privacy is history is that it takes time to understand what works and what does not.

It is worth recalling that various forms of machine learning have been used and regulated for many decades. As discussed, credit score regulations—which address data inputs, algorithms, set points, and other aspects of machine learning—have existed in some jurisdictions since the 1970s. These early forms of machine learning regulations often include well-understood governance mechanisms that are common today, such as error correction, a formal dispute process, extensive government oversight, and other forms of consumer redress. The procedural, and administrative controls used in these types of regulations were new at one time, but are now international standards. These standards, norms, and older governance models enshrined into law need to be taken into full account by those seeking to address the risks of emerging advanced AI systems.

AI governance tools hold out great promise for mapping, measuring, and managing new AI risks. Work to address how AI governance tools can be managed competently with appropriate and helpful guardrails is important, and will entail building the necessary evidence and measurement environments to facilitate this work. Without an evidentiary basis for policy, we are all just making guesses, which is not sufficient to address the actual risks the developing AI ecosystem may have.

The importance of acknowledging what we do not yet know

The more advanced forms of AI that are in place introduce novel problems and new architectures, and these may require a range of new governance approaches. Necessary responses include changes in the way privacy, human rights, and data governance are operationalized and implemented in AI and other ecosystems.

This is new territory, and sufficient evidence regarding valid and fit-for-purpose governance of these systems does not exist yet in a world filled with AI activities. We simply do not how privacy and data governance models must adjust to remain effective. There is a continuing need to construct evidence-based models of privacy and data governance and to use these as a basis to respond to the new challenges and opportunities introduced by advanced AI models.

At the same time, governments, academics, companies, and others have jumped into what appears to be a high-speed race to regulate AI. The reaction appears somewhat instantaneous, especially in comparison to the long ramp that data privacy and data governance policy experienced. It appears that the socio- technical and policy responses to AI will not have the luxury of developing slowly over decades, or perhaps even years. Regulations of varying quality, validation level, and enforceability will be coming soon. This contrasts to the environment of a decade ago when AI governance was not seen as needing immediate action.

As discussed, many of the concerns at the heart of data governance and protection policymaking – fairness, due process, discrimination, openness, rights and responsibilities of data controllers and data subjects, and limits on data use and disclosure – apply to AI activities that include processing of information about individuals, and increasingly, about groups of people and communities. Evidence built up over time shows what works and does not for these systems. However, some of the older governance models are too narrow to address the full range of

---

33 In 2017, Google researchers published a landmark paper regarding advanced AI models called transformer models. This paper effectively marks the beginning of the "Transformer AI era." See: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lion James, Aidan N. Gomez, Lukas Kaiser, Illia Polosukhin, *Attention is all you need*, arXiv:1706.03762v7 [cs.CL], https://doi.org/10.48550/arXiv.1706.03762. In November 2022, OpenAI's announcement and launch of ChatGPT 3.5 brought a landmark proof of concept of what a Large Language Model built on a transformer model could accomplish to the attention of the broader public. The launch of ChatGPT 3.5 was also the spark that lit the regulatory attention to AI policy beginning in 2022 and extending through 2023. *Introducing ChatGPT*, OpenAI, Nov. 30, 2022. https://openai.com/blog/chatgpt.

issues today. Yet some of the newer AI governance models proposed today lack the deep policy knowledge and experience from recent decades.

The world is beginning to acknowledge the substantial impacts of modern AI on privacy, fairness, bias, transparency, and other values. However, there is not yet enough evidence to enable an analysis of what those full range of impacts might be, or how AI activities will develop.

It is critically important to engage in humility regarding what we do and do not know about advanced AI models.[34] We must ensure that evidence gathering and better understanding precede a rush to solutions -- regulation must relate to the reality on the ground. We can accomplish a better result through testing and validation of the technical and policy systems at hand. Otherwise, regulation may well be unfit for purpose and fail to accomplish the goals which are vitally important to attain.

Creating a healthier AI ecosystem and fit-for-purpose guardrails: Toward building an evidence-based AI governance tools environment to operationalize and implement trustworthy AI goals

AI governance tools hold great promise to create improvements at the implementation level. They can function as implementation interfaces for AI systems and ecosystems, and they are already in widespread distribution internationally. This report provides an international survey of the tools as they exist today, a detailed analysis of these tools and their benefits and risks, as well as how they operate in various contexts.The report suggests several concrete pathways for improving outcomes and ensuring that AI governance tools, which are intended to improve AI systems, do just that.

While this report is specifically focused on AI governance tools as an important body of tools to create improvements in AI systems, many additional pathways to improvement regarding more broadly defined AI systems exist, and these possible pathways can be experimented with. Privacy Impact Assessments (PIAs), now commonplace across the world, could be helpful. PIAs were developed in the mid-1990's and reached maturity around 2005-2009. PIA development is still undergoing ongoing cycles of improvement.[35] However, PIAs alone will not be enough to address the full range of challenges that AI presents. Additional tools will be needed.

For example, the development of ethical or trustworthy AI Impact Assessments is already underway.[36] [37]A high-quality and verifiable AI impact assessment that evaluates impacts and validity is essential for those relying on machine learning models and AI system outputs.[38] The need for assessment is especially urgent for those models and systems that support decisions about patient health; matters pertaining to employment and other eligibility-related or eligibility-adjacent decisions; law enforcement and criminal justice decisions; and other activities directly affecting the lives of individuals, groups of individuals, and communities.

Recognize, however, that assessments for privacy, human rights, certain aspects of governance, and other assessments and validation in relation to today's AI activities are far from mature. And even the most perfect

34 See *supra* note 31.

35 Roger Clarke, *Privacy impact assessment: Its origins and developmen*t, Computer Law & Security Review, Volume 25, Issue 2, 2009, Pages 123-135, ISSN 0267-3649, https://doi.org/10.1016/j.clsr.2009.02.002 or https://www.sciencedirect.com/science/article/pii/S0267364909000302. See also Roger Clarke, *An Evaluation of Privacy Impact Assessment Guidance Documents*, 3 November 2010. http://www.rogerclarke.com/DV/PIAG-Eval.html

36 *UNESCO Ethical Impact Assessment*, UNESCO, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000386276.

37 *AI Impact Assessment definition*, NIST AI Risk Management Framework. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. NIST describes AI Impact Assessment tasks as those which "... include assessing and evaluating requirements for AI system accountability, combating harmful bias, examining impacts of AI systems, product safety, liability, and security, among others. AI actors such as impact assessors and evaluators provide technical, human factor, socio-cultural, and legal expertise." Page 226.

38 *Algorithmic Impact Assessment Tool*, Responsible use of artificial intelligence (AI). Government of Canada. Most recent version: 25 April 2023. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

assessment tools will not be enough to address the full range of challenges through the AI lifecycle. More will be needed to address all of the aspects of the lifecycle, including implementation through AI governance tools, and this work has barely begun in regards to addressing advanced forms of AI.

Methodology

This report surveys AI governance tools with to assess how widespread they are, where they are located, and how they might—or might not—improve AI systems. This research examines two use cases in depth plus additional smaller exemplars.

To conduct the preliminary work for this report, the first part of the research sought to determine: 1) how widespread AI governance tools are; 2) where they are published geographically or virtually; 3) what entities created them; 4) the goals of the tools; and 5) the extent to which the tools met their stated objectives. To assess these questions, we sampled a limited set of AI governance tools across jurisdictions and organizations.

This research also analyzed scholarly literature that assesses the quality, functionality and applicability of AI governance tools. The analysis encompasses literature published between 2017 and 2023 that analyzes AI fairness and explainability tools. The research for this report concluded Oct. 31, 2023. Only a few minor updates were added after that time.

An extensive comparative study and analysis of similar fields to determine existing norms and standards for documentation, quality assessment, testing, ongoing monitoring, and other aspects of assessment and post-market improvement cycles was conducted, as well as multiple interviews with experts in the AI field.

This report adopts the term AI Governance Tools and introduces a lexicon of AI Governance Tool Types. The research conducted for Part II of this report made it clear that there was a need to clarify and distinguish among the items commonly labeled generically as "AI Tool" or "AI Toolkit" in the international AI policy and governance sphere.

For more information regarding methodology guiding the evaluation of AI governance tools and Finding 2: *Some AI governance tools feature off-label, unsuitable, or out-of-context uses of measurement methods,* see Appendix C.

Findings

1. AI governance tools are widely published and offered by governments, multilateral institutions, and other organizations.

AI governance tools exist across Africa, Asia, Europe, North America, South America, and Oceania (Australia and New Zealand), at varying levels of maturity and dispersion. Governments, multilateral organizations, academia, civil society, business, and others utilize these tools in different types of AI implementations.[39] This research focuses on AI governance tools used, promoted, or cataloged primarily by governments and multilateral institutions, especially those tools that seek to implement principles of trustworthy AI.[40] It remains difficult to quantify precisely how many tools exist.

2. Some AI governance tools feature off-label, unsuitable, or out-of-context uses of measurement methods.

---

39 The findings are based on recent analysis of select tools. It is not the universe of all tools. All of the AI governance tools analyzed for this report address algorithmic fairness, discrimination and bias, and all but one addresses explainable, transparent and interpretable AI systems. Many of the remaining related items reviewed in Part II also address these two issues, which are prominent in AI governance.

40 This research did not examine all tools available from academia or industry. By "principles of trustworthy AI," this research refers to, for example, the OECD Recommendation on AI and UNESCO Recommendation on the Ethics of Artificial Intelligence, UNESCO, adopted by 193 member states in 2021.

More than 38% of AI governance tools reviewed in this report either mention, recommend, or incorporate at least one of three measures shown in scholarly literature to be problematic. These include off-label, unsuitable, or out-of-context applications when used to measure AI systems.[41]

3. AI Governance Tool providers and hosts have important gatekeeper and quality assurance roles.

Some collections of AI governance tools are published online. This research focused on tools and tool catalogs published by governments and multilateral organizations. These tools and tool catalogs vary significantly in size and types of offerings. Some are comprised of a listing of an AI governance tool with little to no additional information. Some tool collections go further and provide certain levels of assessment or at least a detailed description of the tools.

The OECD, for example, publishes a catalog of AI governance tools that is among the largest offered to date.[42] The OECD framework for its tool catalog may become a helpful model going forward. For example, at least 12 items featured in the OECD's Catalogue of AI Tools and Metrics either mention, recommend, or incorporate off-label measures discussed in Part I of this report, which features use cases of problematic AI fairness and explainability measures. Tool catalog hosts and publishers have important roles as gatekeepers with responsibilities to ensure tool quality and transparency.

Secondary Findings

1. Standards and guidance for quality assessment and assurance of AI governance tools do not appear to be consistent across the AI ecosystem.

This research did not seek to determine as a primary goal whether quality assessments are in place for each AI governance tool or tool catalog. However, it became apparent during the research process that while some AI governance tool providers have conducted some quality assessments of those tools, some have not; if they do conduct quality assessments, AI governance tool providers do not always conduct them according to an internationally recognized standard.

Complete product labeling, documentation, provision for user feedback, requirements for testing, or provision of redress in the case of problems are important features of traditional products, but these features are not always present in AI governance tools.

Pathways for Improvements: Summary

The following is a high-level summary of the solutions and steps that will begin to address the problems and opportunities the research for this report identified. At the end of Part I of this report, a section titled *Pathways for Building an Evaluation Environment and Creating Improvements* discusses in detail potential pathways and solutions toward improving the AI governance tools environment.

Establishing an Evaluation Environment for AI Governance Tools

---

41 Of the select 18 AI governance tools reviewed in detail in this report, 7—or more than 38%—mention or recommend using one of three problematic measures: fairness tools incorporating the US Four-Fifths or 80% Rule, or SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) for AI explainability. Each of these measurement methods have been shown to be unsuitable including when used in an "off-label" manner if applied to measure many types AI systems. See Part I for use cases describing these measures. *See also* Appendix C for a detailed accounting of this finding.

42 *Government AI Readiness Index*, OECD.AI Observatory (May 23, 2023), https://oecd.ai/en/catalogue/tools/government-ai-readiness-index.

There is not enough data yet about how AI governance tools interface with specific standards. As a result, foundational work needs to be done to build an evaluative AI governance tools environment that facilitates validation, transparency, and other measurements. Establishing an evaluation environment for AI governance tools will be crucial to create a healthy AI governance tools ecosystem, and more broadly, a healthier AI ecosystem.

In considering what might help build a transparent, evaluative environment for AI governance tools, the application of international and other standards holds potential. For example, the extensive quality assurance ecosystem articulated in formal standards and norms is well-understood across many mature sectors.

Although many established standards already exist and are important to acknowledge, currently, there is limited knowledge about the functionality of these standards as applied to AI governance tools. Testing of available tools would improve understanding of how existing standards might apply, and it would also support the ecosystem based on evidence. The Plan-Do-Check (or Study)-Act cycle will be a key tool to assist in this maturation.

Establishing Baseline Requirements for Documentation and Labeling of AI Governance Tools:

The research found high variability in the documentation and labeling of AI governance tools. This suggests that developing norms regarding documentation and labeling of AI governance tools could produce meaningful levels of improvements. For example, it would be helpful if tools routinely include information about the developer, date of release, results of any validation or quality assurance testing, and instructions on the contexts in which the methods should or should not be used. A privacy and data policy is also important and should be included in the documentation of AI governance tools.

Additional items can be provided in the documentation, for example:

- Appropriate performance metrics for validity and reliability
- Documentation should provide the suggested context for the use of an AI governance tool. AI systems are about context, which is important when it comes to applicable uses, environment, and user interactions. A concern is that tools originally designed for application in one use case or context may potentially be used in an inappropriate context or use case or "off-label" manner due to lack of guidance for the end user.
- Documentation should give end users an idea of how simple or complex it would be to utilize a given AI governance tool.
- Cost analysis for utilizing the method: How much would it cost to use the tool and validate the results?
- A data policy: A detailed data policy should be posted in conjunction with each AI governance tool. For example, if applicable, this information could include the kind of data used to create the tool, if data is collected or used in the operation of the tool, and if that information is used for further AI model training, analysis, or other purposes.
- Complaint and feedback mechanism: AI governance tools should provide a mechanism to collect feedback from users.
- Cycle of continuous improvement: Developers of AI governance tools should maintain and update the tools at a reasonable pace.
- Conflict of interest: The identities of those who financed, resourced, provided, and published AI governance tools should be made public in a prominent manner in conjunction with publication or distribution of the tool.

The crucial role of NIST and the OECD in convening stakeholders and developing an evaluative environment and multistakeholder consensus procedures for high-quality AI governance tools and catalogs

The **National Institute of Standards and Technology (NIST)** could play an additional role in AI by building an environment in which an evidentiary basis for the socio-technical contexts and best practices for AI governance tools could be created. WPF urges NIST to undertake this work, including developing recommendations for a process for developing, evaluating, and using AI governance tools.

The **OECD** could play an additional role in AI by creating a definitive best-practice framework for AI governance tool or tool catalog publishers by further developing and refining its existing work in this area. This report includes initial suggestions for this work in the *Pathways for Building an Evaluation Environment and Creating Improvements* discussion at the end of Part I., This builds on OECD's existing work on a framework for AI governance tools. WPF urges the OECD to gather international stakeholders to further this work.

Measurement Modeling: A structured Approach Aligning AI Governance Tools and Policy Goals

Measurement modeling could play a positive role in improving outcomes and the quality of AI governance tools. A shorthand for understanding measurement modeling is that it is a structured method that can be used to illuminate gaps between the actual results of measurement systems and policy goals.[43]

When embedded in policy rules and guidance, specific methods or metrics for building more fair, accountable and transparent AI systems and gauging AI risks can have a lasting impact on the ways society comprehends AI systems and their effects on people's lives. What and how we measure something[44] not only reflects our understanding of it, but imposes frameworks or structures for our future understanding.

Measurement modeling is one approach that can assist in this process.[45] For example, measurement modeling has been proposed as a method for recognizing gaps in relation to fairness gaps in computational systems. [46] [47]

Measurement modeling essentially asks evaluators to distinguish between what or how a metric or tool measures, and the goals of the measurement. In other words, is there proper alignment between a metric or tool and the goals of policy? Does the metric or tool actually measure for the same things the policy aims to achieve? The method might be applied when vetting or validating AI governance tools or metrics used to gauge AI fairness, for example.

Some researchers have devised an audit framework for assessing the validity and stability of specific measures such as personality testing metrics used in automated hiring systems. For instance, a socio-technical algorithmic

---

43 Luca Mari et al., Measurement Across the Sciences: Developing a shared concept system for measurement 19-48 and 213-263, (2d ed., 2023) (regarding "Fundamental Concepts in Measurement" and "Modeling Measurement and its Quality").

44 David J. Hand, Measurement: A Very Short Introduction (2016).

45 *Measurement management systems — Requirements for measurement processes and measuring equipment,* Int'l Org. for Standardization, https://www.iso.org/standard/26033.html (Figure 1 in the ISO text shows a model of a measurement management system).

46 Abigail Z. Jacobs & Hanna Wallach, *Measurement and Fairness: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Ass'n for Computing Machinery, 375–385 (Mar. 2021), https://doi.org/10.1145/3442188.3445901.

47 WPF conducted interviews with Abigail Jacobs, an assistant professor of information at the School of Information and an assistant professor of complex systems within the College of Literature, Science, and the Arts at University of Michigan in May and June 2023.

auditing framework found that two real-world personality prediction systems showed "substantial instability with respect to key facets of measurement, and hence cannot be considered valid testing instruments."[48]

This and other frameworks for assessing measurement methods could be helpful to policymakers as they inspect AI governance tools.

Going forward, ensuring alignment of AI governance tools with policy goals for trustworthy AI will be of the utmost importance. For this reason, it will be helpful to assess underlying assumptions about what the measurement mechanisms or methods used in AI governance tools actually do. This very issue is at the heart of the next section of the report, which covers detailed use cases of AI governance tools in their implementation contexts.

48 Alene K. Rhea et al., *An external stability audit framework to test the validity of personality prediction in AI hiring*, 36 Data Mining Knowledge Discovery, issue 6, at 2153-2193 (Sept. 17, 2022).

Part I:

Discussion: Critical Analysis of AI Governance Tools

Governments from Australia to Singapore, Ghana to India, and Europe to the US, have begun to put AI principles into practice by presenting methods for measuring and improving the impacts of AI systems through a variety of AI governance tools. The overwhelming majority of AI governance tools reviewed in this report emphasize two particular governance goals among many[49]: 1.) fairness, or avoidance of bias and discrimination in AI-based decisions and outputs, and 2.) explainability, or interpretability of those systems.[50]

The impulse to operationalize AI principles by measuring and improving AI impacts is a positive one, which will ideally guide users, developers, and other actors on a path toward more beneficial and trustworthy AI systems. Measuring the world around us is one way humans make sense of it. It's only natural that people want to quantify fairness, explainability, and other aspects of AI.

The measures established today could have lasting effects on how the impacts of AI are reflected and interpreted for years to come. Today's measurements will form the foundation of risk scores, consumer scores,[51] ratings, and other statistics we rely on to help make sense of these systems and enforce the rules and regulations addressing them. No one wants to standardize ill-suited methods or embed them in policy in ways that could introduce new problems or harms. That's why it is so important to make sure measurement approaches align with policy goals.

Part II of this report reviews and analyzes a wide-ranging group of more than 30 AI governance tools and adjacent guidance distributed in 13 national jurisdictions across a number of regions. This section's focus is intentionally narrowed to encompass literature that analyzes AI fairness and explainability tools. There is a wealth of relevant literature from scholars in technical and socio-technical fields published between 2017 and 2023. This rich and growing body of work investigates a variety of approaches to measuring and improving AI fairness and explainability. Put simply, it seeks to "measure the measures."

The literature cited and reviewed in this section paints a vivid portrait of what could go wrong if AI governance tools are applied without rigorous evaluation or in inappropriate contexts. This body of literature questions some commonly-used approaches for assessing or improving AI fairness or explainability.[52] It shows that AI measurement methods can lead to AI system accuracy failures, unintentional harms to individuals or groups, or manipulation of metrics to produce tainted measurement outcomes.

Scholars interviewed for this report generally agree that the mission to guide AI actors toward developing and operating more trustworthy AI systems through AI governance tools is beneficial. However, their work may not always be known to policymakers or others influencing those tools. As noted in this report's findings, some AI governance tools mention, recommend, or incorporate off-label uses of potentially faulty or ill-suited tools that are scrutinized in the growing body of scholarly literature.[53] Our intention is to point out salient areas in which the scholarly research we've reviewed might inform future AI governance and AI governance tools.

49 Mohammad Hossin, & M.N Sulaiman, *A Review on Evaluation Metrics for Data Classification Evaluations,* International Journal of Data Mining & Knowledge Management Process, Vol.5, No.2, (March 2015)

50 For additional discussion of terms and definitions, see Part I.

51 Pam Dixon & Robert Gellman, *The Scoring of America*, World Privacy Forum (Apr. 2, 2014), https://www.worldprivacyforum.org/2014/04/wpf-report-the-scoring-of-america-how-secret-consumer-scores-threaten-your-privacy-and-your-future/ (for a definition and discussion of consumer scores and AI scoring in general).

52 Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems: FAT* '19 Proceedings of the Conference on Fairness, Accountability, and Transparency,* Ass'n for Computing Machinery, 59-68 (Jan. 29, 2019), https://doi.org/10.1145/3287560.3287598.

53 *E.g.*, Elizabeth Kumar et al., *Problems with Shapley-value-based explanations as feature importance measures,*119 Proceedings of the 37th International Conference on Machine Learning (ICML'20), art. 509, at 5491–5500 (June 30, 2020), https://arxiv.org/pdf/2002.11097.pdf.

Measuring AI Fairness Measures

A considerable body of research has emerged in recent years highlighting the potential problems when AI actors use automated AI governance tools that promise to create systems that are more fair,[54] but do so without properly assessing those methods or their applicability to a chosen purpose. For example, particularly in the past few years, scholars from around the world have raised alarms about application of metrics that do not align with specific AI fairness-related tasks, such as measuring bias in a dataset used to train an AI model or assessing the risk of unfair decisions made by an AI system.

Recent research intended to elucidate basic requirements of appropriate fairness metrics suggests that "the choice of the most appropriate metrics to consider will always be application-dependent." This scholarly literature finds that assessment of a risk model's fairness in itself is crucial because such models are used to inform human decision-makers.[55]

Governments are just beginning to recognize the need to assess methods intended to improve AI fairness. For example, as discussed in Part II of this report, Chile's 2022 bidding and quality assurance requirements for government acquisition of AI systems stress the importance not only of evaluating the system's impacts on equity, but of evaluating the equity metrics themselves.[56]

Detailed guidance for implementing those requirements states that the type of metric employed is important; it calls on the public sector entity making the purchase to determine appropriate metrics, rather than the technology vendor. In the end, the goal is for both parties to collaborate on determining the most appropriate metrics.[57]

Part II of this report illustrates that governments and other organizations want to put AI principles into practice, and many also want to find ways to produce quantifiable AI risk and analysis measures in the form of fairness scores or ratings. Yet, some of the literature referenced here reminds us that there are pitfalls inherent in quantifying fairness through one-size-fits-all assessments, encoded technical tools, or other quick technical fixes.

This section contains two use cases. The first spotlights an inappropriate use of metrics to automatically alleviate bias from disparate impacts of AI systems. The second highlights the use of SHAP and LIME, two related approaches intended to explain how AI systems produce particular outputs or decisions, both of which have attracted scrutiny among computer science researchers.

---

54 Wesley Hanwen Deng et al., *Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Ass'n for Computing Machinery, 473–484 (June 20, 2022), https://doi.org/10.1145/3531146.3533113.

55 Eike Petersen et al., *On (assessing) the fairness of risk score models: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, Ass'n for Computing Machinery, 817–829 (June 12, 2023), https://doi.org/10.1145/3593013.3594045.

56 *Direccion de Compras y Contratación Publica Aprueba Formato Tipo de Bases Administrativas Para la Adquisición de Proyectos de Ciencia de Datos e Inteligencia Artificial, Resolución N°60*, ChileCompra (Dec. 28, 2022) https://www.chilecompra.cl/wp-content/uploads/2023/01/Bases-Tipo-Ciencia-de-Datos.pdf (the National Artificial Intelligence Policy approved by Supreme Decree No. 20 of Dec. 3 2021, of the Ministry of Science, Technology, Knowledge and Innovation of Chile).

57 *Direccion de Compras y Contratación Publica Aprueba Formato Tipo de Bases Administrativas Para la Adquisición de Proyectos de Ciencia de Datos e Inteligencia Artificial, Resolución N°60*, ChileCompra at 54.

Use Cases in AI Fairness

The Risks of Using the US Four-Fifths Employment Rule for AI Fairness Without Appropriate Context

In the laudable mission to ensure that AI systems do not produce negative impacts on specific groups of people, an array of tools and metrics intended to remove disparate impacts from AI datasets and systems has emerged. Some of these tools use as their foundation encoded translations[58] of a complex US rule: the "Four-Fifths rule."[59]

The Four-Fifths Rule is well-known in the US labor recruitment field as a measure of adverse impact and fairness in hiring selection practices. Detailed in the Equal Employment Opportunity Commission *Uniform Guidelines on Employee Selection Procedures of 1978*,[60] the rule is based on the concept that a selection rate for any race, sex or ethnic group that is less than four-fifths—or 80%—of the rate reflecting the group with the highest selection rate is evidence of adverse impact on the groups with lower selection rates. The rule has been widely applied by employers,[61] lawyers,[62] and social scientists[63] to determine if hiring practices are lawful and if they result in disparate or adverse impacts against certain groups of people.

Employers with more than 100 employees are required to maintain information regarding disparate impact in hiring selection rates, according to the Uniform Guidelines.[64] While the guidelines state that the Four-Fifths rule is "generally" regarded by federal enforcement agencies as evidence of adverse impact, it explains that in some cases, smaller differences in selection rate may constitute adverse impact, and in others, greater differences in selection rate may not constitute adverse impact. In other words, context matters.[65]

Despite its widespread use, legal, employment, and technical experts have cautioned against use of the Four-Fifths Rule as a singular means of assessing disparate impact.[66] Many experts warn against simplistic applications of the rule, both within its historical use in US labor contexts as well as for its use in AI contexts.[67]

In June 2023, the chair of the U.S. Equal Employment Opportunity Commission cautioned against relying solely on meeting the 80% threshold. Calling the Four-Fifths rule "a check" and just one single standard used at the start

---

58 *See* Margaret Rouse, *What does encoding mean?,* Techopedia (Sept. 20, 2023), https://www.techopedia.com/definition/948/encoding. (encoded translations are intended to reflect theoretical concepts in the form of computer code).

59 Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures, 44 Fed. Reg. 43, (March 2, 1979) (question 11 regarding rate of selection).

60 Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607(1978)

61 *4/5ths Rule - Meaning & Definition*, MBA Skool, (Aug. 16, 2023, 11:01 AM), https://www.mbaskool.com/business-concepts/human-resources-hr-terms/13006-45ths-rule.html.

62 1607.4 Information on impact, Legal Information Institute at Cornell Law School, 29 CFR § 1607.4 (Aug. 16, 2023, 11:07 AM), https://www.law.cornell.edu/cfr/text/29/1607.4.

63 Alexander P. Burgoyne et al., *Reducing adverse impact in high-stakes testing*, 87 Intelligence, art. 101561 (July-Aug. 2021), https://doi.org/10.1016/j.intell.2021.101561.

64 Uniform Guidelines on Employee Selection Procedures, *supra*.

65 *Id.*

66 *E.g.*, M.S.A. Lee & L. Floridi, *Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs,* 31 Minds & Machines 165,191 (June 9, 2020), https://doi.org/10.1007/s11023-020-09529-4 ("Feldman et al. (2015) have formalized the approach to identifying disparate impact, but their methodology for pre-processing the data to remove the bias has shown instability in performance of the technique").

67 Philip Roth et al., *Modeling the Behavior of the 4/5ths Rule for Determining Adverse Impact: Reasons for Caution,* 91 J. Applied Psych. 507, 522 (May 2006).

of federal investigations, rather than the only measure used for gauging disparate impact, she said that "smaller differences in selection rates may constitute disparate impact."[68]

Further, according to a U.S. Justice Department legal manual addressing disparate impact, "not every type of disparity lends itself to the use of the Four-Fifths rule, even with respect to employment decisions."[69] Legal scholars also have questioned the limits of the Four-Fifths rule, noting its failure to statistically reflect hiring disparity impact adequately.[70]

Despite those caveats, the Four-Fifths Rule and its 80% benchmark have been repurposed in computer code form and used in a variety of AI fairness metrics and tools.[71] The rule is applied in both employment[72] and non-employment contexts [73] as a means of measuring or "removing" bias or disparate impacts.[74] It is also used outside of the US employment context and is encoded into AI governance tools offered in other jurisdictions. [75]

In a 2019 study of 18 vendors offering algorithmic pre-employment assessments, researchers found that three vendors "explicitly mentioned the 4/5 rule" and several "claimed to test models for bias, 'fixing' it when it appeared."[76] A more recent review indicates this is still happening. As of August 2023, some companies providing AI software publicly mentioned the four-fifths rule as a basis for addressing disparate impact in their systems. [77]

It is not known at this time how many of the entities and individuals using metrics or tools that incorporate the rule's 80% benchmark are aware of the full background, context, and underlying rationale of the four-fifths rule as encoded in those tools. It is also unknown how many of those using the tools outside of a US employment context would continue using them if they were aware of the potential problems.

---

68 Chair Burrows spoke during a keynote speech in June 2023 at the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) attended by World Privacy Forum representatives. She said that it is "worrisome" when employers or vendors suggest that meeting the 80% benchmark is enough to ensure that a hiring approach or system does not create disparate impact.

69 U.S. Dep't of Just., Just. Manual § 7 (1964).

70 Jennifer Peresie, *Toward a Coherent Test for Disparate Impact Discrimination.* 84 Ind. L. J. 773, 802 (2009), http://ilj.law.indiana.edu/articles/84/84_3_Peresie.pdf.

71 Elizabeth Watkins et al., *The Four-Fifths Rule is Not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness,* Parity Techs. Inc., (March 3, 2022), https://ssrn.com/abstract=4037022.

72 Hilke Schellmann, *Auditors are testing hiring algorithms for bias, but there's no easy fix,* MIT Technology Review (Feb. 11, 2021), https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/.

73 *Bias Mitigation with Disparate Impact Remover*, Jupyter nbviewer (Aug. 16, 2023, 11:18AM), https://nbviewer.org/github/srnghn/bias-mitigation-examples/blob/master/Bias Mitigation with Disparate Impact Remover.ipynb.

74 AIF360, GitHub, Trusted AI, Supported Bias Mitigation Algorithms, "Disparate Impact Remover." (November 11, 2023), https://github.com/Trusted-AI/AIF360/tree/master. (Documentation for the Disparate Impact Remover algorithm supported by AI Fairness 360 specifically cites 2015 research introducing a disparate impact measurement based on the Four-Fifths Rule's 80% benchmark.)

75 Multiple AI governance tools surveyed in Part II of this report mention or recommend fairness assessments that use encoded versions of the Four-Fifths or 80% Rule to measure disparate impact. See Part I and Appendix C for more detail.

76 Manish Raghavan et al., *Mitigating bias in algorithmic hiring: evaluating claims and practices, FAT* '20 Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Ass'n for Computing Machinery, 469–481 (Jan. 27, 2020), https://doi.org/10.1145/3351095.3372828.

77 *E.g.*, *Minimize Disparate Impact in Live Data (Pilot)*, Salesforce (2022), https://help.salesforce.com/s/articleView?id=release-notes.rn_bi_edd_bias_live_data.htm&release=238&type=5 (This feature was made available as a pilot with certain customers and was subject to additional terms and conditions); *See also* Karthik Guruswamy, *Mitigating Bias in AI/ML Models with Disparate Impact Analysis*, H2O.ai (Aug. 2, 2019), https://h2o.ai/blog/mitigating-bias-in-ai-ml-models-with-disparate-impact-analysis/; *See also Responsible AI Overview, From Explainability to Responsibility*, H2O.ai (Apr. 2020), https://h2o.ai/content/dam/h2o/en/marketing/documents/2020/04/Responsible-AI-Overview.pdf; *See also* Rabah Abdul Khalek, *How to test the fairness of ML models? The 80% rule to measure the disparate impact,* Giskard (Feb. 1, 2023), https://www.giskard.ai/knowledge/how-to-test-ml-models-5-the-80-rule-to-measure-disparity.

Scholarly researchers also argue that application of the rule in algorithmic recruitment systems is "coarse as it is agnostic to quality of candidates" and does not "account for uncertainties and biases in the data systematically."[78] Scholars also find that codifying the Four-Fifths Rule into AI fairness software should not be used in contexts outside hiring in the US or US labor law and compliance.[79] Scholars also assert that tools intended to produce non-discriminatory AI systems that incorporate the Four-Fifths Rule may miss other important factors weighed in traditional assessments, such as which subsections of applicant groups should be measured using the rule.[80]

Meanwhile, some civil rights and employment lawyers argue that use of the Four-Fifths Rule rule as a test for disparate impact is unreliable in some cases[81] and, particularly in relation to AI used in labor recruitment, "is not only unsupported by the case law, but it is also bad policy." [82]

Use Cases in Automating Fairness: A Compendium of Potential Risks

The movement toward establishing practices that create fairer AI outcomes is positive. However, scholarly literature reviewed for this report indicates an array of unintended consequences of applying metrics or other technical approaches to measure or improve AI fairness.

For example, attempting to de-bias AI systems by abstracting, simplifying and de-contextualizing complex concepts such as disparate impact is just one problematic approach emerging within the AI governance tool environment among many.

It is worth noting there are significant distinctions among definitions of fairness, which may complicate the efficacy of technical approaches designed according to one perception of fairness when used in other contexts.[83] Some key concerns and potential problems:

"Fairness gerrymandering":

"Fairness gerrymandering" in AI fairness tools is a term of art utilized in the scholarly literature to represent when algorithms that take fairness into account have the paradoxical effect of making their outcomes particularly unfair to one subgroup.[84] Technically speaking, this occurs when "a classifier appears to be fair on each individual group,

---

78 Jad Salem et al., *Don't let Ricci v. DeStefano Hold You Back: A Bias-Aware Legal Solution to the Hiring Paradox, FAccT '22 Proceedings in ACM Conference on Fairness, Accountability, and Transparency*, Ass'n for Computing Machinery (June 20, 2022), https://doi.org/10.1145/3531146.3533129.

79 *See* Elizabeth Watkins et al., *supra.* (some researchers suggest that use of the Four-Fifths Rule outside US employment contexts amounts to problematic "epistemic trespassing").

80 Kate Kaye, *Why AI fairness tools might actually cause more problems*, Protocol (June 18, 2022), https://www.protocol.com/enterprise/ai-fairness-tool-disparate-impact.

81 Marion Gross Sobol & Charles J. Ellard, *Measures of Employment Discrimination: A Statistical Alternative to the Four-Fifths Rule,* 10 Indus. Rels. L. J. 3, 381, 399 (1988).

82 Christine Webber & Samantha German, *AI Bias Panel Shows EEOC Should Ditch Four-Fifths Rule*, Law360 (Feb. 15, 2023), https://www.law360.com/articles/1576150.

83 Michelle Seng Ah Lee & Jat Singh. 2021. *The Landscape and Gaps in Open Source Fairness Toolkits, CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Ass'n for Computing Machinery, art. 699 at 1–13 (May 7, 2021), https://doi.org/10.1145/3411764.3445261.

84 Evan Lerner, *Combatting "Fairness Gerrymandering" with Socially Conscious Algorithms*, Penn Eng'g Today (Jan. 31, 2018), https://blog.seas.upenn.edu/combatting-fairness-gerrymandering-with-socially-conscious-algorithms-17e3e63cdbd1/. (This article describes the work of Michael Kearns, founding director of the Warren Center and National Center Professor of Management & Technology in Penn Engineering's Department of Computer and Information Science (CIS), and fellow Warren Center member Aaron Roth, Class of 1940 Bicentennial Term Associate Professor in CIS).

but badly violates the fairness constraint on one or more structured subgroups defined over the protected attributes (such as certain combinations of protected attribute values)."[85]

Fairness gerrymandering might occur if a method for achieving algorithmic fairness is applied in the context of only a small number of pre-defined groups. For example: when, in relation to two binary features corresponding to race and gender, a classifier is considered equitable if it corresponds to one combination of those binary features (such as if it corresponds to a "Black man," or a "white woman"), but not another combination, such as "Black woman."[86] The risk is that an analytical process may result in unfairness in relation to other groups not explicitly considered. In other words, a process that creates more equitable outcomes for some groups might produce undesirable side effects for other groups.[87]

Abstraction traps, oversimplification, and lack of critical context:

Because "abstractions are essential to computer science, and in particular machine learning,"[88] they are inherent in technical interventions that can create "abstraction traps" when used in societal contexts.[89] The aforementioned abstraction of the four-fifths rule is just one example of an abstraction trap. Another such trap might result if an algorithm designed to solve a problem in one social setting, such as predicting risk of recidivism, is also used in relation to loan default. Such abstractions may "render technical interventions ineffective, inaccurate, and sometimes dangerously misguided when they enter the societal context that surrounds decision-making systems."[90]

Also, in an effort to codify governance goals such as fairness and explainability, AI governance tools may lack critical context. Removing or reducing the proper context for an AI governance tool "may flatten nuance and suggest that the tools to solve complex problems lie within the confines of the kit," or can "[abstract] away" crucial elements of the social context in which AI systems are deployed.[91]

For example, the NIST AI Risk Management Framework, an AI governance tool reviewed in Part II of this report, recognizes that metrics used to measure AI risk "can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts."[92]

Scholarly literature also addresses problems that result from application of formal mathematical models of "fair" decision-making used in policy, analyzing the potential for decision-making using algorithms to "violate at least one normatively desirable fairness principle."[93]

Fairness and risk scoring model tradeoffs:

Maximizing fairness across different legally protected groups of people and also achieving maximal accuracy is a topic of intense scrutiny in the literature—because this is nearly impossible to accomplish. Research evaluating

85 Michael Kearns et al., *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness, Proc. of the 35th Int'l Conf. on Machine Learning* (2018).

86 *Id*. at p. 1, example 1.1.

87 Abigail Z. Jacobs & Hanna Wallach, *supra,* at 4.2.

88 Andrew D. Selbst et al., *supra*, at 59–68. .

89 Andreas Tsamados et al., *The Ethics of Algorithms: Key Problems and Solutions,* 37 AI & Society 215 (Sept. 8, 2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3662302.(This paper includes a discussion of five specific abstraction traps, or failures to account for the social context is which algorithms operate).

90 Andrew D. Selbst et al., *supra,* at 59.

91 Richmond Y. Wong et al.,. *Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics.* 67 Proc. ACM Hum.-Comput. Interact. CSCW1, art. 145 (Apr. 16, 2023), https://doi.org/10.1145/3579621 (*See* 145:3).

92 *Artificial Intelligence Risk Management Framework*, *supra*, *at* 29.

93 Ben Green, *Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness*, 35 Philos. Technol., art. 90, 90 (2022), https://doi.org/10.1007/s13347-022-00584-6.

methods and metrics used to score or predict AI risk levels indicates the distribution of true risks may differ among groups, and in particular, may not be proportional to one another.

This could lead to unfair allocation or access to resources, among other potentially negative outcomes.[94] This issue can apply when risk assessments are used by decision-makers across sectors. Examples include mortgage lending, employment, college admissions, child welfare, and medical diagnoses.While some risk scoring models are regulated, many others are not.

Limited applicability throughout AI life cycle:

Some AI governance tools or fairness AI auditing software may only apply during limited phases of the AI life cycle. For example, some AI fairness tools may only apply to the model training stage of AI development. While this is important, determining fairness at one life cycle stage does not mean that fairness is imbued thereafter through the AI life cycle.

For example, if models are adjusted post-deployment the fairness of their outputs can be affected negatively. In addition, some fairness tools do not support early stages of the ML development lifecycle, such as problem formulation stages.[95] Also, there is a risk of AI fairness tools being applied to inappropriate use cases, misinterpreted, or misused.[96]

Limited applicability to third-party AI systems:

Third-party AI systems, including third-party software, hardware, and data components, among others, "may complicate risk measurement."[97] The inner workings of third-party AI systems are not always transparent to users. In addition, certain AI governance tools and metrics may not be applicable when attempting to assess third-party AI software or systems built using unstructured data, as opposed to structured data.[98] [99] The OECD notes the importance of understanding where, when, and what parts of an AI system are built in-house or by a third party, noting three configurations for how this might be operationalized.[100]

Regional contextual constraints: Off-the-shelf AI auditing software products or open-source governance tools may have been designed for use in specific countries that have high AI capacity.[101] Because of this, these products do not always address concerns in all Asian, African, Caribbean, or Latin American jurisdictions, among others.

For instance, an AI governance tool may not recognize nuances among the large variety of Asian and African sub-populations, demographics, and languages. Masakhane Research Foundation, a grassroots organization based

---

94 Eike Petersen et al.

95 Wesley Hanwen Deng et al., *supra*, at 3.1.4 Implications, 477.

96 Michelle Seng Ah Lee & Jat Singh, *The Landscape and Gaps in Open Source Fairness Toolkits, CHI '21 Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Ass'n for Computing Machinery art. 699, 1–13 (May 7, 2021), https://doi.org/10.1145/3411764.3445261.

97 *Artificial Intelligence Risk Management Framework, supra,* at 5.

98 Based on a WPF interview in March 2023 with Jason Tamara Widjaja, director of Artificial Intelligence and Responsible AI Lead, Singapore Tech Center, MSD, known as Merck and Co. in the US and Canada.

99 Herman Sugiharto et al., *Unveiling document structures with YOLOv5 Layout Detection* , ArXiv (Oct. 2, 2023), https://arxiv.org/pdf/2309.17033.pdf.

100 *Advancing Accountability in AI: Governing and managing risks throughout the lifecycle of trustworthy AI,* 349 OECD 22 (Feb. 23, 2023) https://www.oecd-ilibrary.org/docserver/2448f04b-en.pdf?expires=1698704093&id=id&accname=guest&checksum=4466CC55A462E97C223D622680107C7F (describes three scenarios— Universal, Customizable, or Tailed—in respect to third-party AI integrations).

101 *Government AI Readiness Index*, OECD.AI, https://oecd.ai/en/catalogue/tools/government-ai-readiness-index.

in Kilifi, Kenya, that is mentioned in Section II of this report,[102] generates, curates, and annotates datasets that are inclusive of the languages people speak throughout the African continent.[103]

Singapore's Generative AI Evaluation Catalogue[104] also states that LLM evaluation techniques tend to be Western-centric, and should consider user demographics and cultural sensitivities. In addition, India's Tamil Nadu State Policy for Safe and Ethical AI[105] points to the importance of cultural relevance for AI governance tools. This work is deeply embedded within the regional AI and cultural contexts.

Lack of definitional consistency:

There are significant distinctions among definitions of fairness, which may complicate the efficacy of technical approaches designed according to one perception of fairness when used in other contexts.[106]

The difficulty of assessing fairness and privacy in AI systems:

Assessing AI model fairness may be in conflict with data minimization and data protection goals, as well as existing regulations in some circumstances. Measurement for disparate impacts against particular groups requires knowledge of sensitive attributes such as race or age, the very types of data attributes that some data governance regulations may restrict. AI fairness researchers have documented this phenomenon[107] as well as introduced methods for measuring fairness while protecting sensitive data.[108]

It is important to note that this paradox is not present in all data ecosystems. For example, National Statistical Organizations (NSOs) operate under a derogation in most countries of the world, even where data governance legislation is present. There is a unique set of rules providing ethical guardrails for NSOs in precisely these kinds of analytical circumstances.[109] In addition, there are many other exemptions for conducting analysis using sensitive data; for example, public health data during a national public health emergency is often treated more leniently because data protection rules may be suspended in certain emergency situations.[110]

---

102 *Priority Africa Flagship Programmes and Actions*, UNESCO (May 11, 2023), https://www.unesco.org/en/africa-flagship-programmes.

103 Masakhane (July 28, 2023, 10:24AM), https://www.masakhane.io.

104 *Cataloguing LLM Evaluations,* Infocomm Media Dev. Auth. and AI Verify Found. (Oct. 2023), https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf.

105 *India's Tamil Nadu State Policy for Safe and Ethical AI,* Tamil Nadu Information Technology Department (2020), https://it.tn.gov.in/sites/default/files/2021-06/TN_Safe_Ethical_AI_policy_2020.pdf.

106 Michelle Seng Ah Lee & Jat Singh, *The Landscape and Gaps in Open Source Fairness Toolkits, CHI '21 Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Ass'n for Computing Machinery art. 699, 1–13 (May 7, 2021), https://doi.org/10.1145/3411764.3445261.

107 H. Chang & R. Shokri, *On the Privacy Risks of Algorithmic Fairness*, IEEE Eur. Symp. on Sec. and Priv., 292-303 (Sept. 2021), https://ieeexplore.ieee.org/document/9581219.

108 Michael Veale & Reuben Binns, *Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data,* Big Data & Soc'y (Nov. 20, 2017), https://doi.org/10.1177/2053951717743530.

109 *Fundamental Principles of Official Statistics*, United Nations Statistical Comm'n (Jan. 29, 2014), https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

110 Robert Gellman & Pam Dixon, *Covid-19 and HIPAA: HHS's troubled approach to waiving privacy and security rules for the pandemic,* World Privacy Forum. (Sept. 16, 2020), https://www.worldprivacyforum.org/2020/09/covid-19-and-hipaa/.

Inspecting AI Explainability

Governments, corporations, and others using AI systems, along with those affected by these systems, want to understand how these systems make predictions and decisions. Through what is referred to as explainability, explicability, or interpretability, governments and others hope to illuminate the unseen aspects of AI systems.[111]

How AI Transparency, Explainability, and Interpretability Differ

Like many terms related to AI, discrepancies abound regarding the meanings of explainability-related terminology.[112] Some policymakers have distinguished among meanings of the terms transparency, explainability, and interpretability. Research shows that knowing precisely how some AI systems produce outputs can be extremely difficult, even though the components of these systems can be made transparent for evaluation: such as the parameters or weights affecting how a model behaves[113] or the datasets used to train and test a model.

Also, according to some definitions used in the AI policy sphere, there are nuanced differences between AI interpretability and explainability. These definitions suggest that interpretability is intended to satisfy the inquiries of end users or people affected by an AI system, possibly to facilitate some form of redress. Explainability, on the other hand, is the aim of technical practitioners attempting to describe the mechanisms that lead to AI system or algorithmic outputs, possibly to determine what is needed to adjust and improve them.[114]

AI explainability represents the capacity of an AI system to reveal how it arrived at a particular output, such as a decision, prediction or score. (For more details, see the sidebar on transparency, explainability, and interpretability.)

AI developers and practitioners are working to find ways to illuminate the inner workings of AI systems, some of which are becoming increasingly complex, such as neural networks.[115] [116] However, there is no consensus

---

111 *See* Saurabh Bagchi, *Why we need to see inside AI's black box*, The Conversation (May 26, 2023), https://www.scientificamerican.com/article/why-we-need-to-see-inside-ais-black-box/ (the term "black box" is often used to describe the opacity of some AI systems).

112 Cynthia Rudin et al., *Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges*, ArXiv, 2 (July 10, 2021), https://arxiv.org/abs/2103.11251 (It is important to note that some computer scientists don't recognize the same distinctions described here, and argue that literature confounding explainability with interpretability or comprehensibility obscures important arguments, and suggest that attempts to devise new taxonomies related to explainability "miss vast topics within interpretable ML.")

113 *Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods*, Int'l Org. for Standardization, https://www.iso.org/standard/79804.html (in particular sections 3.12, *pieces linear neural network*; 3.13, *binarized neural network*; 3:14, *recurrent neural network*; and 3.15, *transformer neural network*); *see also Artificial Neural Network*, Wikipedia, (Aug. 17, 2023), https://en.wikipedia.org/wiki/Artificial_neural_network.

114 David A. Broniatowski, *Psychological Foundations of Explainability and Interpretability in Artificial Intelligence*, Nat'l Inst. of Standards and Tech., 1-2 (April 2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8367.pdf.

115 *The 2nd Explainable AI for Computer Vision (XAI4CV) Workshop at CVPR 2023*, XAI4CV (June 19, 2023), https://xai4cv.github.io/workshop_cvpr23 (there are many types of neural networks, including, for example, convolutional neural networks and transformer neural networks, among others); *see also* Richard E. Turner, *An Introduction to Transformers*, ArXiv (Oct. 19, 2023), https://arxiv.org/abs/2304.10557.

116 *See* Kate Kaye, *Why AI software companies are betting on small data to spot manufacturing defects*, Protocol (Jan. 12, 2022), https://www.protocol.com/enterprise/landing-mariner-ai-manufacturing-defect (also, in contrast to efforts to build large, complex AI models trained on massive volumes of data intended for a wide variety of purposes and applications, there is also a movement to build highly customized AI models using very small datasets for very specific purposes.

regarding whether it is possible to achieve genuine AI explainability or interpretability.[117] Nevertheless, some scholarly literature indeed shows that AI models that are designed to be interpretable are possible, and that in some cases, such as situations involving high-stakes decisions, "interpretable models should be used if possible, rather than 'explained' black box models."[118]

Some literature goes further still, cautioning against the emphasis on AI explainability goals, suggesting that demands for AI explainability "nurture a new kind of 'transparency fallacy.'"[119]

In addition, some AI auditing experts doubt rhetoric suggesting that some AI systems are too densely complicated to be explained, and suggest that with additional transparency, "the mystery disappears."[120]

The research for this report identified specific questions and concerns related to AI explainability and interpretability detailed in the literature. Here, we highlight a specific use case involving SHAP and LIME, two related approaches intended to explain how AI systems produce particular outputs or decisions, both of which have attracted scrutiny among computer science and AI researchers.

In addition, later in this section, we discuss scholarly literature addressing the limits and unintended consequences of explainable AI methods such as risk of manipulation and inappropriate applications.

SHAP and LIME: Popular but Faulty AI Explainability Metrics

In the absence of widely-adopted AI explainability standards, two approaches—SHAP and LIME—have grown in popularity, despite attracting an abundance of criticism from scholars who have found them to be unreliable methods of explaining many types of complex AI systems.[121]

Use of both SHAP[122] and LIME[123] has increased in part because they are model agnostic, meaning they can be applied to any type of model that data scientists build. An abundance of accessible and easy-to-use documentation related to the two methods has also fostered interest in them.[124]

---

117 Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM Comput. Surv. 5, art. 93, 93:2 (Sept. 2019), https://doi.org/10.1145/3236009.

118 Cynthia Rudin et al., *supra,* at Principle 5.

119 Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 *Duke Law & Technology Review* 18-84 (2017).

120 Lorena O'Neil, *These Women Tried to Warn Us About AI,* Rolling Stone (Aug. 12, 2023) ("Many leaders at these firms even claim that elements of their AI systems are unknowable -- like the inner workings of the human mind, only more novel, more dense. Rumman Chowdhury firmly believes this is nonsense, noting, "When codes can be picked apart and analyzed by outsiders, the mystery disappears." In an email exchange with WPF in September 2023, Chowdhury elaborated on this point, noting that her research suggests that when people seek AI explainability, most aim to have the mechanisms of inputs that led to certain outputs explained, rather than to delve deeply into the technical aspects of a system. Explainability is not a destination or a solution in and of itself, she said; it has to come with accountability).

121 Dylan Slack et al., *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods, AIES '20 Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Ass'n for Computing Machinery (Feb. 7, 2020), https://doi.org/10.1145/3375627.3375830.

122 *shap*, GitHub, https://github.com/shap.

123 *lime*, GitHub, marcotcr, https://github.com/marcotcr/lime.

124 This is based on a description of how SHAP and LIME work and their problems, as intended for a layperson, provided by Tim Miller, professor in artificial intelligence at the School of Electrical Engineering and Computer Science at The University of Queensland, during interviews conducted by WPF in June and November 2023. Miller was professor in the School of Computing and Information Systems at The University of Melbourne, and co-director of its Centre of AI and Digital Ethics, when WPF spoke with him in June 2023. In general, Miller said LIME is unstable and inappropriate as an explainability metric for machine learning, while SHAP-based methods are also limited in effectiveness.
Professor Tim Miller, Univ. Of Queensland Australia, https://eecs.uq.edu.au/profile/9477/tim-miller.

The proliferation and adoption of SHAP and LIME as AI explainability methods recognized and used around the world is evident in documentation related to AI governance tools reviewed in Part II of this report. The review found that six AI governance tools from national governments reference or mention SHAP or LIME or both. In addition, a catalog of tools from a multilateral organization includes 12 items recommending SHAP and/or LIME.[125]

However, the applicability and efficacy of both SHAP and LIME are limited, particularly when used in an attempt to explain complex AI systems comprised of non-linear machine or deep learning models. In a typical use case, an AI practitioner might employ SHAP or LIME to explain a single instance of a model output, such as one decision or prediction, rather than the whole model. Because both methods work by approximating more complex, non-linear models (the types that are often called "black-box" models) with more straightforward linear models, they may produce misleading results.[126]

Short for Shapley Additive exPlanations, SHAP is based on a concept known as the the Shapley Value, introduced by Lloyd Shapley in 1951[127] in the context of cooperative game theory. The Shapley Value is a method used to determine the importance or contribution of each player to an overall competition between groups.[128]

Today, SHAP is used for another purpose entirely: in an attempt to expose and quantify feature importance, or the importance of factors that contribute to predictions of machine learning models.[129] Oftentimes, SHAP is used in the hopes of revealing how factors affect the outputs of opaque, "black box" AI systems such as deep learning models and neural networks, which are difficult to interpret.

SHAP has grown in popularity since around 2017.[130] By 2020, use of SHAP for AI explainability had become widely adopted. When researchers asked people from 30 organizations in 2020 which explainability techniques they used and how, they reported that "feature importance was the most common explainability technique, and Shapley values were the most common type of feature importance explanation."[131]

Why SHAP and LIME Can Produce Misleading Explanations

SHAP reflects feature importance numerically. For instance, when using SHAP to determine how certain input features affect a more straightforward linear regression model trained on a California housing dataset, the SHAP value of the median house age in a block group might be expressed as -0.22, and the SHAP value of median

125 As noted in the findings of this report, several AI governance tools from national governments and multilaterals mention or recommend LIME and/or SHAP, including Chile's procurement form and process for government acquisition of algorithmic systems, IDB FairLAC's *Responsible use of AI for public policy data science handbook*, India's Responsible AI #AIFORALL Approach Document for India Part 1 – *Principles for Responsible AI*, Monetary Authority of Singapore's *FEAT Fairness Principles Assessment Methodology*, 12 items featured in the OECD's *Catalogue of Tools and Metrics*, and Singapore's *AI Verify*..

126 November 2023 WPF interview with Tim Miller.

127 Lloyd S. Shapley, *Notes on the N-Person Game — II: The Value of an N-Person Game*, RAND Corp. (1951), https://www.rand.org/pubs/research_memoranda/RM0670.html.

128 S. Hart, *Shapley Value*, in The New Palgrave Dictionary of Economics 1-6 (1987), https://doi.org/10.1057/978-1-349-95121-5_1369-1.

129 This description is based on an overview of how SHAPley Values work intended for a layperson as provided by Elizabeth Kumar, a Computer Science PhD candidate at Brown University, during interviews conducted by WPF in April and November 2023. Lizzie Kumar personal website, https://iekumar.com/.

130 Scott M. Lundberg & Su-In Lee, *A unified approach to interpreting model predictions, in Proceedings of the 31st International Conference on Neural Information Processing Systems*, Arxiv, 4768-4777 (Nov. 25, 2017), https://arxiv.org/abs/1705.07874 (a research paper presented at the NeurIPS conference in 2017 that is considered instrumental in popularizing the use of SHAP in AI explanations).

131 Umang Bhatt et al., *Explainable machine learning in deployment, FAT* '20 Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency)*, Ass'n for Computing Machinery, 648–657 (Jan. 27, 2020), https://doi.org/10.1145/3351095.3375624.

income as +0.92. The process would be used to add other features, such as the average number of rooms or average home occupancy, until the current model output is reached.[132]

Although Shapley values have been applied in the context of feature importance for decades,[133] researchers have found several mathematical, practical, contextual, and epistemological problems associated with use of the method for explaining AI systems. For example, when attempting to attribute influence to a large set of features affecting AI model decisions or predictions, the approach relies on the modeler to decide which features count as "players" and which are redundant; these subjective decisions can affect the resulting explanations.[134]

Scholarly research also indicates that some users of SHAP may not understand how to interpret its results. A survey of data scientists using SHAP-based tools showed that many were unable to accurately describe what SHAP values or scores represented.[135] The study also found that the popularity of SHAP-based tools influenced some data scientists to trust the tools even if they did not understand what they did or how to interpret their results.

In addition, research shows that use of SHAP in AI explainability tools may lead users to falsely believe they discovered a precise explanation for why or how a system produced a specific output, such as a decision or prediction. This in turn may lead to misconceptions about what SHAP values represent and the actionable information that can be gleaned from them.[136]

Even scholars who acknowledge benefits of using SHAP to provide insight into certain aspects of models and data suggest they "can lead to wrong conclusions if applied incorrectly,"[137]and argue that they can be expensive to compute.[138]

LIME, a similar AI explainability method that has grown in adoption, was first introduced in 2016.[139]Short for Local Interpretable Model-agnostic Explanations, LIME produces explanations by randomly sampling "locally" around the singular instance chosen to be explained. But its randomness is a pitfall: If LIME is used again in an attempt to explain the very same instance, its explanation will be different.[140] The use of LIME for AI explainability has been criticized, and research shows the method can lead to inaccurate results,[141] or be manipulated or "gamed."[142]

---

132 Vinícius Trevisan, *Towards Data Science*, Medium. Jan 17, 2022. https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137.

133 W. Kruskal, *Relative importance by averaging over orderings*, The American Statistician, 41(1):6–10, 1987.

134 I. Elizabeth Kumar et al., *Problems with Shapley-value-based explanations as feature importance measures.*

135 Harmanpreet Kaur et al., *Interpreting interpretability: Understanding data scientists use of interpretability tools for machine learning, CHI '20 Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Ass'n for Computing Machinery, 114 (Apr. 23, 2020), https://doi. org/10.1145/3313831.3376219.

136 Elizabeth Kumar et al., *Shapley Residuals: Quantifying the limits of the Shapley value for explanations*, Neural Info. Processing Sys. (2021).

137 Christoph Molnar et al., *General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models*, Arxiv (2022), https://arxiv.org/pdf/2007.04131.pdf.

138 Christoph Molnar, *SHAP Is Not All You Need*, Mindful Modeler (Feb. 7, 2023), https://mindfulmodeler.substack.com/p/shap-is-not-all-you-need.

139 Marco Tulio Ribeiro et al., *"Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Ass'n for Computing Machinery, 1135–1144 (Aug. 13, 2016), https://doi.org/10.1145/2939672.2939778.

140 According to a November 2023 interview with Tim Miller.

141 Romaric Gaudel et al., *s-LIME: Reconciling Locality and Fidelity in Linear Explanations*, Arxiv, (Aug. 2, 2022), https://arxiv.org/abs/2208.01510.

142 Dylan Slack et al.

Overall, the research indicating that there are vulnerabilities in these popular explainability measures is not reassuring; however, it is not completely unexpected. Trustworthy AI implementation is still nascent, with much work and refinement yet to come.

Risks Inherent in the Rush to Explain AI

As policymakers have pushed for ways to interpret or explain how AI systems make decisions, a growing body of scholarly literature revealing the limits and unintended consequences of explainable AI methods has emerged.

Wrong Tool for the Job?

Some researchers have found that, even if people are provided with explanations for AI decisions or predictions, they may not actually take the explanations into consideration when they make their decisions.[143] For example, people might disregard AI-based recommendations and their explanations associated with a medical diagnosis because they usurp their human decision-making agency or control. Also, common explainability approaches might not provide relevant explanations.[144]

Explainability metrics are not always intended solely to show end users how a system made a decision or how it weighted certain factors. Explainability metrics might also be used by AI practitioners and evaluators to validate and debug models. Developers might use these metrics to help expose ways to adjust models and their data inputs in an attempt to reduce unintended outputs, such as inaccurate predictions or decisions that disfavor specific groups. In other words, explainability metrics and methods are not one-size-fits-all; rather, they are "data, task, and algorithm-specific."[145]

Ultimately, improper application of some AI governance tools can create a false sense of trust and confidence in their ability to explain AI systems.[146] [147]

The limitations of AI explainability methods also can extend to AI policy documents. Recent scholarly literature found a lack of common AI explainability-related terminology and definitions. It also found misalignments between on-the-ground, often-nascent technical research addressing explainability and policy that demands AI explainability be addressed in order to facilitate civil rights goals.[148]

So, the terminology is confusing, and the research itself has not reached a point where it is settled. AI governance tools in the area of explainability are still in an early phase of development. More work may be required on the technical and policy sides before even a rough consensus or middle ground emerges.

---

143 Tim Miller, *Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI*, *FAccT '23 Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, Ass'n for Computing Machinery, 333–342 (June 12, 2023), https://doi.org/10.1145/3593013.3594001.

144 Tim Miller, *Contrastive explanation: A structural-model approach,* 36 The Knowledge Engineering Review, E14 (Oct. 20, 2021), https://arxiv.org/abs/1811.03163.

145 Luca Nannini et al. . *Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK, FAccT '23 Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, Ass'n for Computing Machinery, 1198–1212 (June 12, 2023), https://doi.org/10.1145/3593013.3594074.

146 Elizabeth Kumar et al., *Shapley Residuals: Quantifying the limits of the Shapley value for explanations.*

147 Harmanpreet Kauret al.

148 Luca Nannini et al.

Risk of Manipulating Explanations

Recent research indicates that interpretation of deep learning predictions can be extremely fragile and manipulated relatively easily.[149] Other research highlights the potential for fundamental conflicts between providers and recipients of AI explanations, warning that "the provider might manipulate the explanation for her own ends." For example, feature importance indicated in financial or medical AI system outputs could be sensitive to random or targeted perturbation or disturbance.[150]

Pathways for Building an Evaluation Environment and Creating Improvements in the AI Governance Tools Ecosystem

It is the goal of this research to help gather evidence that will assist in the building of a more reliable body of AI governance tools. In working through the use cases for this report, the research and scholarly literature points to a variety of evaluation, validation, and quality weaknesses that are present in the uses of a number of AI governance tools. If we do not understand the limitations of AI governance tools, we cannot use them to establish a trustworthy ecosystem of AI systems. Indeed, use of the tools without understanding their limitations is more likely to achieve the opposite result.

One of the most significant limitations of AI governance tools is the lack of knowledge about which contexts are and are not appropriate for the use of a particular tool. Further, even when some may be aware of the limitations of a tool, others using it may not. To cite a specific example of this from the research, challenges in using SHAP for AI explainability mentioned in this report's case studies here in Part I are openly discussed amongst technical experts and specialized researchers; however, the problems of applying the four-fifths rule—another measurement approach for AI fairness described in detail in a Part I use case— may be less widely known or understood. This is especially true when the four-fifths rule is encoded into an AI governance tool in a way that is opaque, and then used outside of its originally intended context-sensitive use case.

The research for this report posits several reasons why this and other breakdowns in contextual understanding, among other problems, are occurring. For example:

- AI governance tools are nascent; as such, a transparent, evaluative community basing their judgments on the evidence has yet not been fully constructed.
- The scrutiny and detailed research found in the scholarly literature has not reached all AI governance tool end users, tool publishers, or regulators.
- Some problems may be deeply encoded into the AI governance tools, and these problems can be very difficult to see by even careful researchers, much less by end users of the tools.

In considering what might help build a transparent, evaluative community for AI governance tools, the quality assurance system developed for international and other technical standards holds potential. The extensive body of quality assurance structures developed for standards is well-understood and used across a variety of business and professional sectors. For example, several practices are dependent upon quality assurance and quality control

---

149 Amirata Ghorbani et al., *Interpretation of Neural Networks Is Fragile,* In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019; The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019; The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, 3681–3688 (Jan. 27-Feb. 1, 2019), https://arxiv.org/abs/1710.10547.

150 Sebastian Bordt et al., *Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts, In FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency,* Ass'n of Computing Machinery, 891–905 (June 21, 2022), https://arxiv.org/pdf/2201.10295.pdf.

standardization, including engineering, medicine, nuclear facilities and activities,[151] the development of drugs and therapeutics,[152] and so forth. A premise of this report is that AI governance tools can learn from these existing structures of governance.

Ideally, in a matured evaluative environment, AI governance tools would be of high quality and standardized so that when end-users encounter an AI governance tool, they can use that tool in accordance with well-documented standards. At the same time, AI governance tool publishers and managers could contribute to that process by checking for the presence of quality assurance and standardization to ensure the tools they are making available to the public are trustworthy.

Drawing from existing and well-established standards and norms, this report distills a selection of administrative procedures, tools, and methods that articulate how AI governance tools could be created, documented, managed, and maintained. This research focuses on 1) what AI governance tool developers can use; and 2) what "gatekeeper" organizations that host catalogs of AI governance tools can use for quality assurance. There is still much work to be done to test, adapt and create procedures and norms specific to AI governance tools. The distillation here can provide a starting point for further work in helping to build an evaluation environment for AI governance tools.

Creating an Evaluation Environment for AI Governance Tools

This section provides ideas and suggestions about what standards, norms, guidance, and information may prove helpful as AI governance tools continue to evolve. Although many established standards already exist and are important to acknowledge, there remains limited knowledge about their functionality and trustworthiness as applied to AI governance tools. Testing of available tools improves understanding of the current capabilities and quality and encourages building the evaluative environment based on evidence. The Plan-Do-Check (or Study)-Act cycle will be a key tool to assist in this maturation.

For the nascent body of AI government tools, the focus in the beginning can reasonably be on creating a uniform system of measurement for AI governance tools, and then over time, through experimentation and evidence gathering, building the knowledge of how to improve the standards, or write new ones, if they are required. Quality assessment and management of AI governance tools eventually needs to be a routine part of the AI tools and metrics lifecycle. In time, and with work, it will hopefully be consistent across the AI ecosystem.

AI governance tool developers, end-users and researchers should consider using the existing body of evaluation and measurement standards to assess the quality of AI governance tools in a consistent way. Ideally, actors in the community using AI governance tools can, through cycles of continuous improvement, eventually coalesce around one or more standards that already exist, or work to create new standards over time based on the evidence.

For example, one pathway could be via utilizing ISO 9001 as a starting point. ISO 9001 is a significant family of standards that lays out specific-yet-flexible criteria for a quality management system.[153] This group of standards is among the most widely used international quality management systems standards today. The 9001 standards are scalable, flexible, and adaptable for use with AI governance tools and the broader network of AI governance tool catalogs.

The ISO 9001:2015 is a specific member of the ISO 9001 series of standards focusing on quality management systems. The ISO 9001:2015 *Quality management systems - Requirements* standard defines *quality* as the "degree to

---

151 *Quality Assurance and Quality Control in Nuclear Facilities and Activities: Good Practices and Lessons Learned*, Int'l Atomic Energy Agency (2020), https://www.iaea.org/publications/13656/quality-assurance-and-quality-control-in-nuclear-facilities-and-activities.

152 21st Century Cures Act, H.R. 34, 114th Cong. (2016).

153 *ISO 9001, Quality Management Standard,* Int'l Org. for Standardization), https://www.iso.org/iso-9001-quality-management.html (the 9001 standard family is built on 7 quality management principles, such as process approach, improvement, and evidence-based decision making, among others).

which a set of inherent characteristics of an object fulfills requirements." This definition could apply to many types of "objects" including products and services supplied to or created for the machine learning and AI governance tools ecosystem. There is some existing scholarly work that assesses quality aspects of machine learning systems,[154] but this kind of quality assessment has not yet been systemically applied to AI governance tools.[155]

The implementation of a continuous improvement cycle, which takes the form of the Plan-Do-Check-Act (PDCA) cycle in the ISO 9001:2015, could be a practical and specific starting point in this standard which could be usefully applied to AI governance tools include in particular.[156]

Plan-Do-Check-Act: The PDCA Cycle

The ISO 9001:2015 standard opens with the *Plan-Do-Check-Act cycle*, a process-based approach focusing on continuous cycles of improvement and risk-based thinking. From the standard:

> The PDCA cycle enables an organization to ensure that its processes are adequately resourced and managed, and that opportunities for improvement are determined and acted on.
>
> Risk-based thinking enables an organization to determine the factors that could cause its processes and its quality management system to deviate from the planned results, to put in place preventive controls to minimize negative effects and to make maximum use of opportunities as they arise.[157]

ISO 9001:2015 is a flexible standard. In applying the PDCA cycle to the development or publication of AI governance tools, several implementation models are instructive.[158] Fortunately, there is a meaningful body of work on implementing the PDCA cycle in various contexts.[159] The Deming Cycle, also known as PDSA, or Plan-Do-Study-Act cycle, is also of relevance.[160] Although the PDCA cycle and the Deming Cycle are often used interchangeably, the Deming Institute views the "study" aspect of the PDSA cycle as distinct from the "check" aspect of the PDCA cycle. The Deming Institute properly evaluates the "check" aspect as focused more on the implementation of a change, with success or failure attached.

A detailed history of the evolution of the PDCA and PDSA cycles describes both of their roots in the scientific method, and how the cycles differ.[161] For AI governance tools, there will likely be a period of adjustment and

---

154 J. Siebert, et al., *Towards Guidelines for Assessing Qualities of Machine Learning Systems*, 1266 Commc'ns in Comput. and Info. Science (Shepperd, M., Brito e Abreu, F., Rodrigues da Silva, A., Pérez-Castillo, R. eds., 2020), https://doi.org/10.1007/978-3-030-58793-2_2.

155 Teresa Datta et al., *Tensions Between the Proxies of Human Values in AI, Contributed Talk, NeurIPS 2022 Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*, 2023 IEEE Conf, on Secure and Trustworthy Machine Learning (forthcoming), https://arxiv.org/pdf/1911.02508.pdf (the authors argue that proxies for fairness, particularly those with mathematical constructions, are poor substitutes for more robust analysis. The authors make a potent argument that some interventions, such as some fairness interventions, actually cause harm).

156 It is worth noting that the PDCA cycle is reflected in some ways in iterative "agile" development processes commonly used throughout the software and broader technology industries: How is Scrum Related to Plan-Do-Check-Act (PDCA) Process?, ArchiMetric, (January 9, 2019), https://www.archimetric.com/how-is-scrum-related-to-plan-do-check-act-pdca-process/.

157 ISO 9001:2015 (E), Introduction, 0.1, General. https://www.iso.org/standard/62085.html. *Note:* WPF worked from the full 2021 standard for the analysis in this report; it is the most current version of the standard. There are sector-specific applications of the standard that are available separately.

158 Many types of AI governance tools exist. AI governance tools can be a simple questionnaire, or they can be an entire AI system. Quality assessment will necessarily scale with the complexity of the AI governance tool. However, the core components of documentation, testing, evaluation, etc. can still be attained, even for the simplest tool.

159 *Project planning and implementing tools: Plan-Do-Check-Act Cycle,* Am. Soc'y for Quality (Oct. 2023), https://asq.org/quality-resources/pdca-cycle.

160 *The PDSA Cycle*, The Deming Institute, https://deming.org/explore/pdsa/.

161 Ronald D. Moon & Clifford L. Norman, *Circling back: clearing up the myths about the Deming cycle and seeing how it keeps evolving*, Quality Progress (2010), http://www.apiweb.org/circling-back.pdf.

experimentation as tool developers, publishers, and others test and fine-tune the PDCA and/or the PDSA cycle specifically for AI governance tools.

The following are some potential avenues to consider in implementing the four key aspects of the PDCA cycle:

1. **Plan:** Determine the purpose of the AI governance tool and then assess it to see if it is fit for purpose or accomplishes the intended goals.Then test its functionality and trustworthiness: Does it respect privacy? Is interpretable? Is it secure? Has the context been well-understood and identified?

For example:

- Identify and understand the use case or context in which an AI governance tool will be used. Choosing the right tool for the proper context is an important aspect of fitness for purpose, which is why it is high on this list.
- Develop the criteria for assessment, aligned with the context. For example, what are the key aspects of the AI governance tool that need to be assessed?
- Assess the risks involved in implementing the tool. Risk analysis comprises an extremely large literature for AI as well as other fields. AI Impact Assessments,[162] Data Privacy Impact Assessments,[163] and Safety Impact Assessments[164] are helpful tools to experiment with as the evaluation environment is built.
- Determine relevant evaluation criteria to address the stated purpose of the AI governance tool being tested. The criteria will reflect any key aspects that will be addressed in the testing. For example, is the tool accurate, does it scale, if so how much or how well, and how usable is the tool? Are the results from using the tool interpretable, and if so, how interpretable?
- Some AI governance tools will utilize benchmark datasets. Choosing benchmark datasets for testing is a domain of research in and of itself. The suggestion here is that a diverse set of benchmark datasets could be chosen, and they should be clearly representative of the types of data the tool or metric addresses.
- Identify appropriate performance metrics and cut-off points. For example, if a developer is testing tool *quality*, the *performance metric* will indicate the point after which the tool or metric ceases to be of acceptable quality for its described purpose.

2. **Do: Execute the plan.**

For example:

- Conduct experiments: What experiments assess the AI governance tool? In this step, the performance metrics from the Plan phase are fully tested. Ideally, the experiments will have a formal methodology that is fair and that can be made public and be evaluated. Testing should at a minimum follow the tool developer's recommendations.

---

162 Recommendation on the Ethics of Artificial Intelligence, UNESCO, November 2022. Available at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics .

163 See the UK Information Commissioner's materials on data protection impact assessments. Backgrounder with screening checklist, process checklist, and links to more information: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-governance/accountability-and-governance/data-protection-impact-assessments/ . See also the ICO's Template for a DPIA: https://ico.org.uk/media/for-organisations/documents/2553993/dpia-template.docx.

164 Safety impact assessment in this context refers to the standards literature available in ISO and other Standards Development Organizations. The ISO 31000 — Risk management series, IEC 31010:2019 Risk management, Risk assessment techniques, is a robust standard to explore regarding safety impact assessments that can be adapted to a wide range of situations. Another adaptable standard is the ANSI / ASSP Z590.3 standard, Prevention Through Design.

- Adversarial testing: Adversarial testing[165] can gauge the quality and utility of AI governance tools. This testing can be done prior to deployment by developers, and or after release by tool or tool catalog publishers, end users or researchers. There is empirical support for adversarial testing of AI governance tools. For example, a well-known early adversarial attack on LIME and SHAP usefully revealed that the methods did not fulfill their stated purpose.[166] Developing adversarial testing approaches specific to AI governance tools is another nascent area of work. Work to develop baselines and standardized approaches specifically for testing AI governance tools should be a priority.

- Analyze testing results: What do the results obtained from the experiments indicate? What patterns does the testing reveal? What strengths or weaknesses become apparent after testing? How well does the AI governance tool perform across diverse testing datasets?

3. Check/Study: Evaluate the methodology, evaluation methods, and implementation of testing. Conduct gap analysis. The evaluation result should also inform the quality and fitness of steps 1 and 2.

For example:

- Gather input about the evaluation process itself to assess whether it can be improved. For instance, how many different datasets were tested? Were a variety of different parameters tested? Were a variety of different settings used?

- Consider peer review and additional external validation: For example, involve domain experts or other researchers knowledgeable about the AI governance tool, including model evaluation where applicable, documentation and labeling evaluation including truth-in-advertising statements about the product, and evaluation of appropriate contexts and use cases.

4. Act: Improve transparency and functionality of the tool(s).

For example:

- Public dissemination of the evaluation methodology, results, and conclusions can improve transparency. This can take the form of a quality assessment report. Quality assessment reports are called "safety reports" in some fields. These reports contain "practical examples and detailed methods that can be used in support of safety standards."[167]

- End users should be made aware of the evaluations in a prominent manner, and the evaluation should be readily understandable by non-expert users. What problems can users expect? What outcomes can users reliably expect to achieve?

- Ensure AI governance tools are transparent and free from conflicts of interest. For example, core pieces of information to make available to end users should provide details about how development of AI governance tools are resourced and financed, by whom, and who published them. Ensuring that there is no conflict of interest is a quality measure that is readily achievable. Ensure that this information is included in each tool or technique's documentation. Conflict of interest statements will differ depending on the type

---

165 *See ISO/IEC 27050-1:2019, Information technology: electronic discovery,* Int'l Org. for Standardization. https://www.iso.org/standard/62085.html (adversarial testing is a well-developed concept in the domain of information security and data protection, and especially in the de-identification literature. It is becoming a well-understood concept in the AI literature as well. *Adversary* in these contexts typically means an individual or entity that can exploit potential vulnerabilities, intentionally or unintentionally. *Adversarial testing* refers to intentional testing of systems to find vulnerabilities or problems).

166 Dylan Slack et al.

167 For example, reports on safety, best practices, quality assurance, and training in nuclear activities are issued as "safety reports." *Quality Assurance and Quality Control in Nuclear Facilities and Activities*, Int'l Atomic Energy Agency (2020), https://www-pub.iaea.org/MTCD/Publications/PDF/TE-1910_web.pdf (the first 24 pages of this manual are highly relevant quality control discussions that contain helpful information about quality control in complex environments).

of tool. For example, a complex checklist may have one publisher and several authors. A set of metrics, models or algorithms used to address problems such as bias may also require disclosure of the entities who have authority to modify source code for that tool.

There is a great deal more work to be completed to reliably and consistently demonstrate that AI governance tools are fit for purpose. Ideally, within the next few years, a robust body of evidence will emerge to assist tool developers, users, and researchers in their analysis and understanding of the various aspects of AI governance tools.

- Solicit ongoing feedback about the tool/product. Ensure a feedback mechanism is routinely available.
- Iterate and refine.
- Repeat the evaluation with different tools, datasets, or evaluation criteria as necessary to continue to refine and improve the product.

Suggested Framework for Hosts, Publishers, or Managers of AI Governance Tools and Tool Catalogs

Multilateral institutions, governments, and other large entities that host AI governance tool repositories or collections are responsible for ensuring that the AI governance tools they host adhere to high product standards and are fit for purpose. In other mature product-focused realms, there are often specific laws addressing standardization and responsibilities for quality assurance. AI governance tools have not reached full maturity as a body. As a result, the quality controls for individual products, and, by extension, quality control for hosts and publishers of AI governance tools, are not yet fully developed. While this work is underway, much more needs to be done to support it.

Among the most important AI governance tool catalogs published today is at the OECD.AI Observatory,[168] which publishes a large catalog of AI governance tools called the OECD Catalogue of Tools and Metrics for Trustworthy AI.[169] Fortunately, the OECD has been experimenting and gathering evidence for several years in this area. In a foundational background paper, the OECD provides a framework articulating how it constructs its catalog.[170] On page 15 of the paper, the OECD includes a detailed chart depicting its organizational methodology for the catalog and discusses the structure and framework for its decision making in regards to its Catalogue of Tools.[171] This framework can be seen in Appendix D.

---

168 OECD Artificial Intelligence Observatory. OECD.AI, https://oecd.ai/en/ .

169 *OECD Catalogue of Tools and Metrics*, OECD.AI, https://oecd.ai/en/

170 *Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems*, 312 OECD 13 (June 2021), https://www.oecd-ilibrary.org/docserver/008232ec-en.pdf?expires=1699321405&id=id&accname=guest&checksum=AE9989A3DD8FA6F82EC930BD69F3758D Figure 1 in this document shows the high-level structure of the framework of tools for trustworthy AI. Figure 2 on page 15 of the document shows the full framework.

171 *Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems*, 312 OECD 13 (June 2021), https://www.oecd-ilibrary.org/docserver/008232ec-en.pdf?expires=1699321405&id=id&accname=guest&checksum=AE9989A3DD8FA6F82EC930BD69F3758D. See pages 10, 11 and 12 (Tables 2, 3 and 4) for the selection of technical, procedural and educational tools to implement trustworthy AI.

Figure 4: ***OECD Catalogue of Tools and Metrics Framework***

Type	Field	Definition	Options (if applicable)
Tool description	Name	The name of the tool	
	Link	A link to an up-to-date document	
	Description	A brief summary of the tool and its purpose	
Tool origin	Organisation	The organisation that developed the tool	
	Stakeholder group	The stakeholder group from which the initiative originates	Academia; Trade union/worker representative; Private sector; Civil society; Technical community; Public sector; International governmental organisation; Other
	Country	The country or region where the initiative originated	International; OECD countries; List of regions; List of countries; Other
	Date of publication	Date the tool was published in its first version	
	Contact email	Email of the contact person for the tool (not for public use)	
Tool categorisation	Type of Approach	High-level category of the tool	Process-related approach; Technical approach; Educational approach; Other
	Type of Tool	Category of the tool	Toolkits/toolboxes/software tools; Technical documentation; Technical certification; Technical standards; Product development/lifecycle tools; Technical validation tools; Guidelines; Governance frameworks; Risk management tools; Sector-specific codes of conduct; Collective agreements; Certification; Process-related documentation; Process standards; Change management processes; Capacity/awareness building tools; Inclusive design guidance; Educational materials/training programmes; Other
Scope	Technology platform	The technology platform(s) that the tool can be used for	Platform neutral; Platform specific; Multi-platform; Other
	Target stakeholder group	The stakeholder group where the tool is expected to be implemented	Academia; Trade union/worker representative; Private sector; Civil society; Technical community; Public sector; International governmental organisation; Other
	Primary and secondary policy area	The policy area(s) where the tool is expected to be implemented	Agriculture; Competition; Corporate governance; Development; Digital Economy; Economy; Education; Employment; Environment; Finance and insurance; Health; Industry and entrepreneurship; Innovation; Investment; Public governance; Science and technology; Social and welfare issues; Tax; Trade; Transport; All of the above; Not applicable; Other
	Geographical scope	The country or region that the initiative targets	International; OECD countries; List of regions; List of countries
	Target users of the tool	Users who are expected to use the tool to implement a project	AI system business leader; AI system technical developers; IT specialists; Researchers; AI system operators; Executive management; Government agencies; Data scientists; Project managers; HR managers; All employees; Other
	Impacted stakeholders	Groups of people that will be impacted by the implementation of the tool	Employees; Specific policy communities; Consumers; Regulators; Management; Other
	AI system lifecycle stage(s) covered	The stages of the AI system lifecycle that the tool helps to implement	Planning & design; Data collection & processing; Model building & interpretation; Verification & validation; Deployment; Operation & monitoring; All stages
Alignment with international AI Principles	Relevance to international AI Principles	Grade relevance to international AI Principles	Values-based Principles: Socio-economic and environmental impacts; Human-centred values & fairness; Transparency & explainability; Robustness, security, safety; Accountability; Human agency and oversight. Recommendations for policy makers: Investing in research; Data, compute, technologies; Enabling policy environment; Jobs, skills, transitions; International co-operation
Potential for adoption	Maturity of the tool	Project phase the tool is currently in	Project stage; In development; Running code; Implemented in one project; Implemented in multiple projects; Not relevant anymore; Other
	Degree tool is kept up to date	How the tool is kept up to date with evolving standards, requirements, etc.	No update mechanism planned; Periodic review; Always up to date; Other
	Degree of free use of the tool	Legal conditions for using the tool	Subscription fee; One-time license fee; Free-to-use (creative commons); Open source; Other
	Required resources to implement	The extent to which certain resources are needed to implement/use the tool	IT skills; Domain expertise; Data; IT infrastructure; Operational infrastructure; Financing
	Stakeholders involved	Stakeholders who will be involved in the implementation and operation of the tool	IT employees; Operations employees; All employees; Business unions; Trade unions/worker representatives; Clients; Suppliers; Government agencies; Other
Implementation incentives	Expected benefits	Expected benefits from using the tool	Reduction in risk of AI system failure; Reduction in cost of AI system implementation; Faster implementation of an AI system; Increased quality of AI system results; Improved ability of AI system's implementation to scale; Responsible implementation of AI system; Other
	Enforcement mechanisms	Enforcement mechanisms attached with the usage of this tool	Internal mediation (ombudsman); Ethics board; Certification; Enforcement body; Governmental regulation; Log registrars; Reporting frameworks; Collective agreements; N/A; Other

Source: OECD, June 2021, https://www.oecd-ilibrary.org/docserver/008232ec-en.pdf.

The OECD Catalogue of Tools and Metrics for Trustworthy AI is the first known multilateral framework of its kind to specifically address AI governance tools.

To understand how the OECD's current methodology could offer broader guidance for other tool catalogs and repositories, this research analyzed the OECD guidance, and compared it with product documentation standards in other areas of work, such as the ISO standards for product documentation, quality assessment and assurance, as well as consumer product safety standards.

In conducting this comparative work, this research found several areas where the OECD framework can be made more robust and provide greater quality controls for the tools listed. The structure below is adapted from the OECD's original catalog structure, with several additions from the ISO literature and near-worldwide normative laws including data governance law and consumer product safety law. The original OECD chart is reproduced in Appendix D.

Adapted framework for collections of AI governance tools (using as a foundation the OECD framework of tools for trustworthy AI)

The following framework utilizes the structure of the OECD framework for its Catalogue of Tools and Metrics, and adapts it to include elements that make it more robust in creating a healthy environment for AI governance tools.

1. General Information:

- Tool Name
- Context of Use
- Description
- Link to tool and documentation for the tool

- Organisation
- Country of Origin
- Date of Publication

2. Technical Specifications:

- Type of Tool (e.g., software, guideline, standard, certification, etc.)
- Technology Platform Compatibility (e.g., platform-neutral, specific, multi-platform)
- AI System Lifecycle Stages Covered (e.g., design, data collection, processing, deployment, monitoring, etc.)

3. Target Audience:

- Intended Users or User Groups (e.g., AI developers, researchers, and a variety of end users, such as affected consumers and communities, etc.)
- Applicable Policy Areas (e.g., health, finance, transportation, etc.)
- Geographical Scope

4. Usability and Accessibility:

- Tool Usability
- Accessibility (Assesses the tool's adherence to accessibility standards)
- Required Resources for Implementation (e.g., IT skills, domain expertise, infrastructure, cost, etc.)

5. Trustworthiness Metrics:

- Alignment with International AI Principles (e.g., transparency, interpretability, fairness, robustness, accountability, etc.)
- Ethical Considerations
- Privacy and Data Protection Measures (ensure a privacy policy is posted.)
- Security Measures

6. Performance and Effectiveness:

- Maturity of the Tool (e.g., prototype, in development, production-ready, etc.)
- Case Studies or Use Cases (if available)
- Validity and reliability of the tool in relation to fulfillment of tool and policy goals including in relation of trustworthiness metrics
- User Reviews and Ratings

7. Update and Maintenance:

- Update Process (e.g., how frequently the tool is updated)
- Support and Maintenance Availability

8. Cost and Licensing:

- Cost of tool use and testing (if any)
- Licensing Type

9. Data Policy:

This category would assess how the AI governance tool handles data from tool users. Key points might include:

- User Rights: What control do users of AI governance tools have over the data involved in the lifecycle of the tool?
- Does the AI governance tool train on any data? If so, is there a choice in the matter?
- What is the copyright policy of the AI governance tool? For example, is there a policy that ensures all data involved in the lifecycle of the tool will continue to be governed by a policy and made transparent?

10. Transparency and Conflict of Interest Notice:

This category lists applicable resources and funding sources, and/or financial interests in relation to the tool. Commercial Interests also should be noted, for instance, if the tool promotes specific commercial products or services. Affiliations, including relationships that potentially impact objectivity, including information about commercial or other entities that donated the tool for open-source use, should also be noted.

Individual Product-level Documentation for AI Governance Tools

The research conducted for this report found inconsistent documentation for AI governance tools. It was not a focus of the methodology for the report to analyze this particular aspect of AI governance tools, but it was difficult to avoid noticing it. Hosts and publishers of AI governance tools and tool catalogs could find experimentation regarding documentation to be fruitful. AI governance tool documentation deserves more discussion, given its importance, and given the inconsistent application of documentation when it is made available.

A great deal of existing work has already been done in other areas that could be helpful. For example, significant documentation standards and norms exist around consumer products, software products, and other technology products offered to the public. These norms are encapsulated in multiple ISO standards,[172] as well as OECD Responsible Business Conduct principles and implementation guidance.[173] [174]

Baseline documentation for individual AI governance tools is an area where rapid improvements may be achieved. Just as with any tool or product released to the public, developers should create and make robust documentation available. Entities publishing large collections of AI governance tools should require robust documentation prior to publication of any tool, no matter how simple or complex the tool may be.

---

172 See for example G. F. Hayhoe, "ISO standards for software user documentation," *2012 IEEE International Professional Communication Conference*, Orlando, FL, USA, 2012, pp. 1-3, doi: 10.1109/IPCC.2012.6408631. See also the work of NIST regarding recommended criteria for cybersecurity labeling of consumer software. While not directly related regarding topic, the procedures and ideas for labeling could be helpful, particularly if tested in the AI governance tools context. See: Recommended criteria for cybersecurity labeling of consumer software, National Institute of Standards and Technology, Feb. 4 2022. Available at: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.02042022-1.pdf .

173 *OECD Due diligence guidance for Responsible Business Conduct*, OECD. 31v May 2018. Available at: https://www.oecd.org/investment/due-diligence-guidance-for-responsible-business-conduct.htm .

174 Allan Jorgensen, Karine Perset, Rashad Abelson, *Recoding our understanding of RBC in science, tech, and innovation*, OECD. Oct 02 2023. Available at: https://www.oecd-forum.org/posts/recoding-our-understanding-of-rbc-in-science-tech-and-innovation-what-s-new-in-the-oecd-mne-guidelines .

As mentioned, many standards exist regarding product documentation, some sector-specific.[175] Ideally, over time, a set of product documentation characteristics specific to AI governance tools will develop through testing, experimentation, and evidence building.

The following suggestions for creating documentation for an AI governance tool are distilled from the large body of existing work in software documentation and consumer product documentation, among other areas. These suggestions are provided here as an initial step toward further experimentation and evidence building.[176]

Suggestions for Transparency and Documentation of Individual AI Governance Tools: [177]

Documentation will typically focus on what was done, what assumptions and constraints were present, what data was used, and other elements that a user, a developer, a researcher or tools publisher may want to know and use for validating a tool.

For example, the following is an adaptation of documentation found in the standards literature:

Title of AI Governance Tool

Introduction and Context - An overview of the AI tool or metric will often include a discussion of the purpose and background of a tool, a description of its key features, and a description of the target audience or users of the tool.

Getting Started and User Manual - This can include items like system requirements, installation guide when applicable, and use instructions. For software-based AI governance tools, a detailed guide would describe the features of the tool, instructions for use, and a guide to troubleshooting.

Socio-Technical Documentation - This documentation provides developers and technical users with information about an AI governance tool's architecture, interfaces, APIs if applicable, as well as integrations with other AI governance tools or systems. It could also include information about intended environments or contexts of use and intended human interaction.

Tutorials and Case Studies - These can discuss relevant use cases for the AI governance tool, and proposed uses of the tool.

Changelogs/Version History - This can provide historical data regarding iterative updates and changes made to a tool, including a version history.

Support and Contact - AI governance tool providers can let users know how to get support or report product issues. Ideally, the documentation will include a feedback process and formal complaint mechanism.

Funding and Conflict of Interest Statement - This category creates transparency regarding the resourcing, funding, affiliations, and objectivity of the tool. The importance of such transparency structures cannot be

175 See for example exemplars from the field of health, which can be quite specific: Ethical standards for clinical documentation integrity professionals, AHIMA. 2020. Available at: https://www.ahima.org/media/r2gmhlop/ethical-standards-for-clinical-documentation-integrity-cdi-professionals-2020.pdf?oid=301868 . See also: *Clinical documentation guide,* Marin Health and Human Services, 2021. Available at: https://www.marinhhs.org/sites/default/files/files/servicepages/2021_05/documentation_manual_2021_v_5-12-21_0.pdf .

176 Documentation can cover many aspects of products, from documenting the use of a product, or documenting other aspects of a product such as product features. See for example a comprehensive international standard on user instructions, which is now a multi-standards body standard. See: BS EN IEC/IEEE 822079-1:2020 *Preparation of information for use (instruction for use) of products,. Principles and general requirements. The definitive guide to writing instructions for use*, British Standards Institute. October 2020. Available at: https://knowledge.bsigroup.com/products/preparation-of-information-for-use-instructions-for-use-of-products-principles-and-general-requirements?version=tracked .

177 Note: The developer of the AI governance tool is the entity that would ideally provide or at least contribute to the documentation.

overstated, given the variety of ways both conscious as well as unconscious bias can undermine the objectivity of research and tool outcomes.

Concluding Thoughts Regarding How to Build an Evaluation Environment and Create Improvements in the AI Governance Tools Ecosystem

The suggestions in this discussion of how to begin creating and building improvements for the AI governance tools ecosystem discuss a body of standards, governance techniques, and approaches that are currently available. This is by no means to say that these are the only policy tools that should be considered—far from it. Many additional strategies exist. For example, regulator-approved codes of conduct may play an important role in addressing specific AI risks in focused use-cases or even specific types of AI systems.[178]

This being said, this discussion highlights the rich body of strategies that, while often overlooked in the governance of AI, may be among the most important. When it comes to AI governance, it is still early; right now, we are all crossing the river by feeling the stones.

178 Regulator- approved codes of conduct are not the same as industry self-regulation. For example, industry codes of conduct in countries with GDPR or GDPR-commensurate legislation will typically use Article 40 in the GDPR or its equivalent, which sets out specific rules for codes of conduct that fall somewhere between formal regulations and self-regulations. See: *Article 40*, GDPR: https://gdpr-info.eu/art-40-gdpr/. One exemplar of a code developed under GDPR Article 40 is the 2021 *EU Cloud Code of Conduct,* expressly approved by the European Data Protection Board and the Belgium Data Protection Authority. See: *EU Cloud Code of Conduct,* https://eucoc.cloud/en/home/. In the US, the *Voluntary Consensus Standards* enshrined in OMB Circular A-119 is the regulatory vehicle that facilitates a similar process. See: Office of Management and Budget, Circular A-119, revised, 1998. https://obamawhitehouse.archives.gov/omb/circulars_a119.

PART II:

A Survey of AI Governance Tools and Other Notable AI Governance Efforts from around the World

The movement to advance responsible and trustworthy AI beyond theoretical principles toward practical implementation is upon us. This research indicates there is a keen interest in devising workable policy approaches to measuring and improving AI systems according to established AI principles, and it is happening on a worldwide scale.

Primarily in the past four years, governments, multilateral organizations, non-governmental organizations (NGOs), development banks, standards bodies, academic institutions, and public-private partnerships have taken concrete steps toward establishing AI governance tools for measuring and improving AI fairness, explainability, robustness, privacy, and more. This section of the report discusses these tools.

Some of this important work has been out of the spotlight amid more prominent pressures for enforceable AI regulations and guardrails. The AI governance tools we survey here are implementation tools: they form the interface between the goals of AI governance and how it happens in reality. These tools generally do not fall under specific AI regulations at this time, depending on the jurisdiction. However, it would be a mistake to assume that only strict regulation would afford positive change in a technical ecosystem: many AI governance tools demonstrate a genuine commitment to address goals and concerns around AI.

In some cases, these reviews include constructive critiques and warnings regarding components of AI governance tools that demand improved inspection and quality assurance. Although development of AI governance tools is nascent, this is no time to ignore due diligence for the methods and measures that will form the basis of AI governance for years to come. AI systems affect real people, groups of people, and communities every day.[179] And, although sandbox policy approaches and toy technical projects have potential value, they still will need effective quality controls and assessment before dissemination.

As we detail in Part III of this report, a next step for many of the tools reviewed here will be reevaluation and possible adjustment to ensure a reliable and thriving AI governance tool ecosystem people can trust.

Highlights from Distinctive AI Governance Tools and Activities

Notable efforts to put goals for responsible AI into practice are happening in earnest around the world. Consider it a much-needed sign of unity in an often divided environment.

This landscape of AI governance tools is fertile ground. Some of the tools reviewed here represent efforts that have reached more advanced stages than others. All are important.

Although commonalities do exist in all of these AI governance tools, the efforts are in no way entirely homogenized. They occupy a vast spectrum of nuanced approaches and ideas, reflecting the panoply of communities and cultures that dot the globe.

A few of these tools in particular stand out for a variety of reasons.

---

179 AI fairness, disparate impact, and explainability are issues that affect people right now. However, sometimes these immediate impacts are dismissed or deemphasized by people who are more concerned with the "existential" risks that advanced AI systems may pose in the future. No matter which AI-related issues are considered most pressing, the importance of high-quality and appropriately-applied metrics and methods for measuring the impacts of AI systems and achieving legitimate system improvements cannot be denied.

- In South America, the government of Chile has established a new approach to acquisition of AI. In an effort to incentivize technology providers to build their systems better from the start, the country's updated technology bidding and procurement process ensures public sector agencies thoroughly assess the AI systems they seek to use.
- In India, a plan from the state of Tamil Nadu for adoption of AI-based systems proposes the use of a point-based rating system. The rating system considers commonly-mentioned factors including fairness and transparency, but it also measures something we did not see in many other AI governance tools: diversity, as well as relevance and performance of AI systems across geographies and societies. This is a key concern in India, where the world's largest population represents several distinct cultures, languages, and customs.
- In Kenya, a project created to help ensure that genuinely inclusive AI can be built has taken root. Sometimes, AI governance means tackling foundational problems before tools are needed to improve existing AI systems. One such problem plaguing Africa is a lack of inclusive language data. Without high-quality data reflecting diverse populations, AI systems will not be fair, inclusive, or beneficial for everyone. Masakhane, which roughly translates to "We build together" in isiZulu, has launched a variety of initiatives designed to create datasets representing low-resourced African languages.
- In New Zealand, a process for identifying and reducing risk throughout the life cycle of an algorithm not only centers on personal data collection, use, or disclosure throughout all stages of algorithm governance, but it also strives for relevance to communities not typically represented in government technology policy. The process recommends use of frameworks for assessing the cultural data implications of algorithms affecting Māori communities.
- In what may be considered the most technical approach to implementing AI principles among the AI governance tools reviewed in this report, Singapore has developed software and a technical testing framework for improving the fairness, explainability, and robustness of AI systems. In October 2023, it became one of the first national governments to recommend specific approaches to evaluating generative AI systems such as Large Language Models.

Notably, the research and analysis in this report does not include corporate tools, although some of the tools reviewed here do mention, recommend, or incorporate AI governance methods created by private corporations. And in general, we narrowed the scope of AI governance tools distributed by governments to national-level governments rather than state or local governments.

Still, this report is inclusive in a very deliberate way. For AI to be truly fair and equitable—this survey of AI governance tools indicates this is the most common goal of all AI principles—it must be inclusive. That means inclusive data or inclusive approaches to AI-related education, work, and development. It also means that when it comes to measuring AI's impacts, everyone gets a seat at the table.

Some entries in this survey are not fully mature tools yet; they are efforts or programs that are adjacent to, or building toward operationalizing responsible AI. These tools were worthy of highlighting in order to ensure geographical inclusivity and to provide an equitable overview of AI governance tools activities as they exist today.

Notes About How This Section is Organized

The AI governance tools in this section are organized using the M49 ISO standard for naming and listing regions, subregions, and nations. The M49 ISO standard is the United Nations/ISO joint standard for naming and classifying regions, subregions, and for articulating country names.[180] It is considered to be the authoritative standard for naming and structuring geographical-related data.

180 *M49 Standard*, United Nations. https://unstats.un.org/unsd/methodology/m49/.

For those unfamiliar with this standard, there are distinctions that may be unfamiliar. The major regions are as follows:

Africa
Americas
Antarctica
Asia
Europe
Oceania.

This report lists the tools in order first by the major regions. So, for example, AI governance tools from Africa are listed first, with tools from the Americas second, and so forth.

The M49 standard includes subregions, and this report includes subregions where necessary. For example: for tools from Singapore, the region is Asia and the subregion is South-eastern Asia.

The tools reviewed here appear in the following order:

- Tools from international organizations
- Tools from international standards organizations
- Tools from regional banks
- Tools from national governments, followed by tools from other organizations in countries, organized in alphabetical order by region name: subregion name: country name.

This report introduces and uses the term AI Governance Tools, which this report defines as:

AI Governance Tools:

Socio-technical tools for mapping, measuring, or managing AI systems and their risks in a manner that operationalizes or implements trustworthy AI.[181]

This definition encompasses the wide array of formats and methods reviewed here in Part II. There are different types of tools. The following lexicon of AI governance tool types further distinguishes differences among them.

[**Possible Image**] **AI Governance Tool Types**

- **Practical Guidance** - Includes general educational information, practical guidance, or other consideration factors
- **Self-assessment Questions** - Includes assessment questions or detailed questionnaire
- **Procedural Framework** - Includes process steps or suggested workflow for AI system assessments and/or improvements
- **Technical Framework** - Includes technical methods or detailed technical process guidance or steps
- **Technical Code or Software** - Includes technical methods, including use of specific code or software
- **Scoring or Classification Output** - Includes criteria for determining a classification, or a mechanism for producing a quantifiable score or rating reflecting a particular aspect of an AI system

181 The definition for *AI governance tools* was developed by the authors of this report at the World Privacy Forum. It is based on the research for this report, the scholarly literature, and consultation with a wide range of technical, standards, legal, and policy experts. This definition maps to the OECD AI Principles, the National Institutes of Standards and Technology Trustworthy and Responsible AI principles, and the general outlines of the EU AI Act. The definition was finalized November 10, 2023 in Paris, France.

The AI Governance Tools Comparison Chart spotlights key features of select mature AI Governance Tools from national governments and multilateral organizations.

Figure 5: ***AI Governance Tool Types and Features Comparison Chart***

Our AI Governance Tools Comparison chart spotlights key features of select AI Governance Tools from national governments and multilateral organizations. The features indicated here directly map to the tool types we assign to each tool, and reflect our AI Governance Tool Lexicon featured in Appendix A.

		Tool Features					
		Practical guidance	Assessment questions	Process steps	Technical process guidance, code or software	Score or classification output	Mentions specific metrics, code or software
Australia 2019	Automated Decision-making Better Practice Guide TYPE: Practical Guidance with Self-assessment Questions	✓	✓				
Canada 2019	Algorithmic Impact Assessment tool TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓	✓			✓	
Chile 2022	AI Procurement Directorate TYPE: Practical Guidance	✓					✓
Dubai 2019	AI System Ethics Self-Assessment Tool TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓	✓			✓	
Ghana 2023	FACETS Framework TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓	✓			✓	
India 2021	The Responsible AI Approach Document for India Part 1 TYPE: Practical Guidance with Self-assessment Questions	✓	✓				✓
Tamil Nadu, India 2020	Policy for Safe and Ethical AI TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓				✓	
Latin America & Caribbean/IDB 2019	fAIr LAC in a box TYPE: Catalog	✓	✓	✓	✓		✓
Global/OECD 2023	Catalogue of AI Tools and Metrics to Promote Trustworthy AI TYPE: Catalog	✓	✓	✓	✓	✓	✓
New Zealand 2020	Model Development Lifecycle TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓	✓	✓		✓	
Singapore 2022	AI Verify TYPE: Practical Guidance with Technical Framework & Software	✓	✓	✓	✓		✓
Singapore 2022	Veritas Initiative TYPE: Practical Guidance & Process Framework with Self-assessment Questions & Technical Code	✓	✓	✓	✓		✓
Singapore 2023	Generative AI Evaluation Catalogue TYPE: Practical Guidance	✓					✓
UK 2021	AI and Data Protection Risk Toolkit TYPE: Practical Guidance & Process Framework with Scoring Output	✓		✓		✓	
US 2021	Artificial Intelligence: An Accountability Framework TYPE: Practical Guidance & Process Framework with Self-Assessment Questions	✓	✓	✓			✓
US 2022	Artificial Intelligence Governance Toolkit TYPE: Practical Guidance & Process Framework with Self-Assessment Questions	✓	✓	✓			
US 2022	Blueprint for an AI Bill of Rights TYPE: Practical Guidance	✓					
US 2023	Artificial Intelligence Risk Management Framework TYPE: Practical Guidance & Process Framework with Self-Assessment Questions	✓		✓			✓

Source: World Privacy Forum, Research: Kate Kaye, Pam Dixon. Image/Data Visualization: John Emerson.

Intergovernmental Organization Toolkits and Use Cases from International and Regional Multilateral Institutions

A range of intergovernmental organizations[182] have begun working in earnest on implementing trustworthy AI. Among entities establishing ways to operationalize trustworthy AI principles, the intergovernmental organizations are especially important, as their work has a greater opportunity to become normative across multiple countries and, in some cases, multiple regions.

Organization of Economic Cooperation and Development (OECD)

The Organization of Economic Cooperation and Development (OECD) is a multilateral institution that, for much of its history, has focused on the most developed economies.[183] OECD is headquartered in Paris, France. Today, it has 38 government members with five accession members (Argentina, Brazil, Bulgaria, Croatia, and Peru) and five key partners (Brazil, China, India, Indonesia, and South Africa).[184]

The OECD is notable for its formal multistakeholder process, which includes representatives from governments, industry, trade unions, and civil society. OECD's Council Recommendations, such as the OECD Privacy Guidelines,[185] are normative, and are also Customary International Law,[186] which can be adjudicated under the auspices of the International Court of Justice.[187] The OECD Privacy Guidelines also went on to form the early basis of most international privacy law as it developed in the 1980s through the 1990s and beyond.

The OECD began work in 2018 crafting the first multilateral principles for AI. The OECD AI Secretariat appointed a large group of international AI experts called the AI Expert Group, or AIGO. Along with the OECD

182 *Pritzker Legal Research Center*, Northwestern Pritzker School of Law, https://library.law.northwestern.edu/InternationalResearch/IGO. (An intergovernmental organization (IGO) is an entity established by a treaty or agreement between member states that agree to work together on projects or issues of common interest. IGOs can cover all or most jurisdictions (such as the United Nations), regional (such as the Association of Southeast Asian Nations), or subject-specific (such as the European Free Trade Association)).

183 In 1961, the *OEEC* became the OECD with the U.S. as a full member. The goal was to provide a forum for discussion of economic problems of mutual concern post-WWII. *See. S. Rep. No. 935 (*1948); *see also OECD Observer*, June 1967, at 10-11; *see also* Curt Tarnoff, *The Marshall Plan: Design, accomplishments, and significance*, Congressional Research Service (Jan. 18, 2018), https://sgp.fas.org/crs/row/R45079.pdf.

184 *OECD Member Countries, Accession Countries, and Key Partners*, OECD (Sept. 2023), https://www.oecd.org/about/members-and-partners/.

185 *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data,* OECD, (September 23, 1980) https://www.oecd-ilibrary.org/science-and-technology/oecd-guidelines-on-the-protection-of-privacy-and-transborder-flows-of-personal-data_9789264196391-en.

186 Customary International Law, Cornell Law School Legal Information Institute, Cornell Law School, https://www.law.cornell.edu/wex/customary_international_law ("Customary international law refers to international obligations arising from established international practices, as opposed to obligations arising from formal written conventions and treaties"). Note: The International Court of Justice, which is the main judicial body of the United Nations, acts in matters of customary international law to settle disagreements between member states. See: Report of the International Law Commission on the work of its seventieth session, Report of the Sixth Committee, United Nations. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/376/74/PDF/N1837674.pdf?OpenElement. For additional background, see: Analytical guide to the work of the International Law Commission, United Nations. https://legal.un.org/ilc/guide/1_13.shtml. See: Resolution 73/203 of 20 December 2018. For additional background on customary international law See also: Alan Watson, An Approach to Customary Law (1984). See also: Ernest Gellner, Nations and Nationalism (1984).

187 *See International Court of Justice*, United Nations, https://www.icj-cij.org/home (the International Court of Justice is the main judicial body of the United Nations).

government delegations and formal stakeholder groups, AIGO formed the core of the OECD multistakeholder process related to AI. The OECD AI Principles were adopted as a formal OECD Council Recommendation in May 2019, and were ratified by member governments.[188] The OECD AI Principles were the first "soft law" AI principles to be developed.[189] Since this time, the OECD has published a handbook, established a dedicated AI Working Party, launched and populated an international AI Observatory (OECD.AI), and is moving forward on multiple AI projects inside expert groups. The OECD AI Principles have been adopted normatively. The US National Institute of Standards and Technology (NIST) has incorporated OECD definitions and framework into its work,[190] and the European Union also has adopted the OECD definition of an AI system in its EU AI Act proposed legislation. There are cooperative OECD efforts in the broader standards world and in legislative arenas as well.

OECD AI Principles: Recommendation of the Council on Artificial Intelligence

The OECD AI principles are written in a way that emulates the brevity of the OECD Privacy Guidelines. The principles include inclusive growth, sustainable development, and well-being; human-centered values and fairness; transparency and explainability; robustness, security, and safety; and accountability. The principles feature recommendations regarding national policies and international cooperation for trustworthy AI, including investing in AI research and development; fostering a digital ecosystem for AI; shaping an enabling policy environment for AI; building human capacity and preparing for labor market transformation; and international cooperation for trustworthy AI. The OECD is actively working to implement these principles through expert groups, including those focused on accountability, AI incidents, and other practical guidance.

OECD.AI Catalogue of AI Tools and Metrics to Promote Trustworthy AI

Tool Type: Catalog

One of the focus points for the OECD's current AI work is to guide the implementation of the OECD AI Principles. There are multiple aspects to this work, and one particularly significant component has been the 2023 launch of the OECD.AI Catalogue of AI Tools and Metrics to Promote Trustworthy AI (hereafter, the Catalogue of Tools). The Catalogue of Tools was created to facilitate broader accessibility to the available tools and metrics intended to produce more trustworthy AI systems[191] that respect the OECD AI Principles.[192]

The OECD's Catalogue of Tools repository features a wide variety of AI governance tool types. As of September 2023, the Catalogue of Tools includes technical tools for auditing or reducing bias; technical tools for robust and secure data collection and processing; metrics to measure model performance; and metrics to measure privacy. The collection also includes metrics to evaluate portions of AI datasets, research papers detailing methods for

---

188 *Recommendation of the Council on Artificial Intelligence*, OECD (May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

189 The Legal Information Institute describes customary international law as referring to "...international obligations arising from established international practices, as opposed to obligations arising from formal written conventions and treaties." See: *Customary International Law*, Legal Information Institute, Cornell Law School, https://www.law.cornell.edu/wex/customary_international_law. For more, see *supra* note 186.

190 *Artificial Intelligence Risk Management Framework, supra*, at *1, 2, 9 and 10.*

191 *About the Catalogue, Catalogue of AI Tools and Metrics to Promote Trustworthy AI,* OECD.AI, https://oecd.ai/en/catalogue/faq.

192 *Recommendation of the Council on Artificial Intelligence*, OECD (May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

addressing AI problems, links to software and auditing services sold by private corporations,[193] [194] and an open-source database for archiving real-world examples of AI failures and vulnerabilities.[195] The submissions of tools and metrics featured in the Catalogue of Tools are to be vetted by the OECD Secretariat as well as partner stakeholder organizations to ensure accuracy and objectivity.[196] [197]

As noted in the Findings of this report, multilateral institutions and other AI Governance Tool providers and hosts have an important role as quality assurance gatekeepers. The OECD's Catalogue of Tools includes tools intended to measure and improve fairness and explainability of AI systems, some of which have drawn sharp criticism in scholarly literature reviewed here in Part I. For instance, nine items hosted in the Catalogue feature use of SHAP, and three feature use of LIME. In addition, despite scrutiny among researchers of methods that abstract and encode the US Four-Fifths Employment Rule in an attempt to measure disparate impact of AI systems, three tools featured in the OECD's catalog include such methods.[198]

There are differences in opinion among researchers and practitioners regarding the effectiveness of these measures in practice when applied to explain or interpret AI or determine its potential disparate impacts on particular groups.[199]

United Nations Educational, Scientific, and Cultural Organization (UNESCO)

UNESCO Recommendation on the Ethics of Artificial Intelligence

In 2022, UNESCO published its global standard on AI ethics. It did so in its flagship Recommendation on the Ethics of Artificial Intelligence.[200] The text may be summarized into what UNESCO calls four core values: human rights and dignity; living in peaceful, just, and interconnected societies; ensuring diversity and inclusiveness; and environment and ecosystem flourishing. The core principles include the idea of proportionality as well as a "do no harm" principle. Other core principles include the right to privacy and data protection, sustainability, and fairness and non-discrimination.

---

193 *Fairly AI Oversight and Risk Management Platform, Catalogue of Tools and Metrics for Trustworthy AI*, OECD.AI, https://oecd.ai/en/catalogue/tools/fairly-ai-oversight-and-risk-management-platform.

194 *Holistic AI Audits, Catalogue of Tools and Metrics for Trustworthy AI*, OECD.AI, https://oecd.ai/en/catalogue/tools/holistic-ai-audits.

195 *AI Vulnerability Database, Catalogue of AI Tools and Metrics to Promote Trustworthy AI,* OECD.AI, https://oecd.ai/en/catalogue/tools/ai-vulnerability-database.

196 About the *Catalogue (*The OECD Catalogue of Tools vetting process includes the following: "The Catalogue of AI tools & metrics has mechanisms to ensure that content is accurate and up to date. It operates with an open submission process, where tools are submitted directly by the organisations or individuals who created them, and by third parties. Submissions are vetted by the OECD Secretariat to ensure accuracy and objectivity. There is a biannual review and updating process when organisations are encouraged to submit new initiatives and update existing ones. If an existing initiative isn't updated over a two-year period, it will be removed from the Catalogue. Partnerships with relevant stakeholders – including Business at the OECD, the OECD Civil Society Information Society Advisory Council and the OECD Trade Union Advisory Committee – facilitate this biannual review").

197 This research discusses the OECD.AI Catalogue of Tools and the framework OECD uses for the catalog, with specific discussions about what could be changed to make advances and improvements in quality assurance of the tools.

198 See Appendix C for a detailed list of such tools.

199 *See* Dylan Slack et al.; *see also* Elizabeth Kumar et al., *Problems with Shapley-value-based explanations as feature importance measures; see also* Josh Poduska, *SHAP and LIME Python Libraries: Part 1 - Great explainers, with pros and cons to both*, Domino Data Lab (Dec. 2018), https://www.dominodatalab.com/blog/shap-lime-python-libraries-part-1-great-explainers-pros-cons (there is a robust literature focused on AI governance tools that are intended to be used for facilitating fairness, explainability, and interpretability in AI. This literature is discussed further in Part I of this report); *see also* Watkins et al.

200 *Recommendation on the Ethics of Artificial Intelligence.*

UNESCO has also produced training around its AI Recommendation. For example, its introductory online training course on AI and the Rule of Law features sections addressing "best practices that translate ethical principles into practice both in terms of the use of AI in justice systems, and in cases involving AI impacting human rights."[201] The course, aimed at members of the judiciary, includes sections on algorithmic bias and its implications for judicial decision-making and AI ethics and governance concerning judicial operators.[202]

UNESCO AI Ethical Impact Assessment and Report

UNESCO's AI Ethical Impact Assessment (EIA), fully articulated and discussed in a 2023 report,[203] assesses algorithms according to their alignment with the principles and guidance contained in the UNESCO Recommendation on AI. The EIA also recommends creating transparency by surfacing information about AI systems across the life cycle. The EIA is designed for procurers of AI systems, as well as those who want to evaluate whether or not use of AI is appropriate for a given problem.

World Bank Group

Risk Assessment Framework for World Bank Projects

The World Bank Group (WBG) created a Risk Assessment Framework initiative to Identity and Classify Ethical Risks from AI use in World Bank projects.[204] The framework was based in part on the OECD Framework for the Classification of AI Systems.[205]

The WBG also has proposed an initiative to ensure World Bank operations staff have tools for establishing algorithmic accountability and identifying AI's impact on human rights.[206]

International Standards Organizations

International Standards Organizations play a pivotal role when it comes to putting policy into practice and establishing norms for technology design, development, and use internationally. So, it should come as no surprise that standards bodies have begun contributing to the AI governance tool ecosystem.

This section discusses the three main international standards development organizations, or SDOs. NIST is an SDO. Although technically based in the US, NIST is unique in that it works cooperatively with multilateral organizations in its standards settings. As such, NIST functions as an international SDO.

There are many regional SDOs, including the European Committee for Electrotechnical Standardization (CENELEC), The African Organization for Standards (ARSO), and the British Standards Institution (BSI), among many others. This report, however, does not include an analysis of these or other regional SDOs.

---

201 *Almost 4000 judicial operators worldwide join UNESCO's MOOC on AI and the Rule of Law*, UNESCO Newsroom (Apr. 20, 2023), https://www.unesco.org/en/articles/almost-4000-judicial-operators-worldwide-join-unescos-mooc-ai-and-rule-law.

202 *AI and the Rule of Law: Capacity Building for Judicial Systems*, UNESCO (Apr. 24, 2023), https://www.unesco.org/en/artificial-intelligence/rule-law/mooc-judges.

203 *Ethical Impact Assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence*, UNESCO (2023), https://www.unesco.org/en/articles/ethical-impact-assessment-tool-recommendation-ethics-artificial-intelligence.

204 *Risk Assessment Framework initiative to Identity and Classify Ethical Risks from AI use in World Bank Projects*, The World Bank Group, https://aiforgood.itu.int/about-ai-for-good/un-ai-actions/wbg/.

205 *OECD Framework for the Classification of AI systems*, OECD Digit. Econ. Papers No. 323 (Feb. 22, 2022), https://doi.org/10.1787/cb6d9eca-en.

206 *Risk Assessment Framework initiative to Identity and Classify Ethical Risks from AI use in World Bank Projects.*

International Organization for Standardization (ISO)

ISO is a significant international SDO, which to date has created a number of core technical standards for AI systems in addition to its overarching technical standards work.[207] ISO launched its AI technical subcommittee in 2018[208] to address AI computational methods, trustworthiness, and societal concerns through the development of standards and guidelines. Some key AI-related work from ISO includes a standard providing guidelines for development and use of products, systems, and services for managing AI risk;[209] technical reports on AI bias measurement; and assessment of neural network robustness.[210] ISO has taken a whole-of-ecosystem approach[211] to its work on AI standards: its current roster of 20 AI-related standards, technical specifications, and reports thus far addresses bias, robustness, and various aspects of risk management.[212]

National Institute of Standards and Technology (NIST)

Artificial Intelligence Risk Management Framework and AI RMF Playbook

The National Institute of Standards and Technology is based in the US; however, in the area of AI, it has functioned in substantive ways as an international standards organization due to its collaborative work with OECD on AI terminology and other related work. NIST published its Artificial Intelligence Risk Management Framework[213] and companion AI RMF Playbook in January 2023.

The framework is divided into two parts. First, it discusses how organizations can frame the risks related to AI, and then it describes in detail the four functions present in the AI life cycle—Govern, Map, Measure, and Manage—and how to address them. While the Govern function applies to all stages of an organization's AI risk management processes, the framework document explains that the Map, Measure, and Manage functions can be applied according to specific contexts at specific stages of the AI life cycle.

The NIST AI RMF recognizes the fallibility of metrics used to measure AI risk, noting that such metrics "can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts." To address potential problems with AI measures, the framework suggests that AI metrics and effectiveness of existing controls are regularly assessed and updated, and that the AI system measurement process should involve consultation with multidisciplinary stakeholders.[214]

---

207 *See ISO/IEC JTC 1/SC 42*, Int'l Electrotechnical Comm'n, https://www.iec.ch/ords/f?p=103:22:403202137104604::::FSP_ORG_ID,FSP_LANG_ID:21538,25 (ISO/IEC JTC 1/SC 42, or simply ISO subcommittee 42, is a joint ISO IEC (International Electrotechnical Commission) committee that focuses on developing international standards for multiple aspects of AI, from smart manufacturing to medical equipment to biometrics and sustainability. The working groups operating under SC 42 also cover extensive aspects of AI such as foundational standards, data, and trustworthiness).

208 Robert Bartram, *The new frontier for artificial intelligence*, Int'l Org. for Standardization, (Oct. 18, 2018), https://www.iso.org/news/ref2336.html.

209 *ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management*, Int'l Org. for Standardization, 35.020, 35, ICS, (July 27, 2023, 6:27PM), https://www.iso.org/standard/77304.html.

210 *Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making*, ISO/IEC JTC 1/SC 42, (July 27, 2023, 6:29PM), https://webstore.iec.ch/publication/71949.

211 Wael William Diab, *IEC and ISO work on artificial intelligence: Covering the entire AI ecosystem*, E-tech, (May 20, 2022), https://etech.iec.ch/issue/2022-03/iec-and-iso-work-on-artificial-intelligence.

212 *ISO IEC JTC 1 / SC 42 : Artificial Intelligence*, Int'l Electrotechnical Comm'n, https://www.iec.ch/ords/f?p=103:22:400018789550151::::FSP_ORG_ID,FSP_LANG_ID:21538,25

213 *Artificial Intelligence Risk Management Framework.*

214 *Artificial Intelligence Risk Management Framework, supra*, at 29.

The NIST AI Risk Management Framework acknowledges that metrics used to measure AI risk "can be oversimplified, gamed, lack critical nuance, become relied upon in unexpected ways, or fail to account for differences in affected groups and contexts."

There are numerous other examples of practical implementation in the NIST framework, which has substantial overlap and harmonization with the OECD implementation work as well as the work being done on AI implementation and standardization at ISO.

As of October 30, 2023, NIST has been given substantial additional responsibilities in relation to AI governance, safeguards, and standards.[215] For example, NIST is tasked with developing guidelines, standards, and best practices for AI safety and security to help ensure the development of trustworthy AI systems. As part of this, it will develop a companion resource to the AI RMF for generative AI.

Institute of Electrical and Electronic Engineers (IEEE)

GET Program for AI Ethics and Governance Standards

IEEE, a well-established international SDO, has developed the IEEE GET Program in conjunction with partners including product certification company TÜV SÜD.[216] The program provides free access to seven AI ethics and governance standards as of January 2023.[217]

The program is designed to support efforts regarding AI ethics and governance literacy as well as AI systems governance and standardization, among other topics. The GET program is tied to IEEE's broader global initiative regarding ethical AI, which aims to "ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity." [218]

Standards featured in the IEEE GET Program include guidance and procedural methods for assessing impacts of AI systems according to human well-being; procedures for achieving testable levels of transparency; standards for ethical system design and data privacy; and a set of ontologies to establish ethically driven methodologies for the design of robots and automation systems.

The GET Program's IEEE Standard for Transparency of Autonomous Systems, for instance, includes detailed methods for gauging the level of transparency of an AI model.[219]

In related work, IEEE's Global Initiative on Ethics of Autonomous and Intelligent Systems also includes a Standard for Algorithmic Bias Considerations, one of 11 IEEE ethics-related standards currently under development.[220] The

215 *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, The White House (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

216 *IEEE GET Program: GET Program for AI Ethics and Governance Standards*, IEEE Xplore, (July 27, 2023, 6:32PM), https://ieeexplore.ieee.org/browse/standards/get-program/page/series?id=93.

217 *IEEE Introduces New Program for Free Access to AI Ethics and Governance Standards*, IEEE Standards Ass'n, (July 27, 2023, 6:32PM), https://standards.ieee.org/news/get-program-ai-ethics/.

218 *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*, IEEE Standards Ass'n (July 27, 2023, 6:40PM), https://standards.ieee.org/industry-connections/ec/autonomous-systems/.

219 *IEEE Standard for Transparency of Autonomous Systems*, in IEEE Std 7001-2021, 1-54, (July 28, 2023, 7:32AM) https://ieeexplore.ieee.org/document/9726144.

220 A. Koene et al., *IEEE P7003TM Standard for Algorithmic Bias Considerations*, 2018 IEEE/ACM Int'l Workshop on Software Fairness, 38-41 (2018), https://ieeexplore.ieee.org/document/8452919.

standard includes a development framework intended to avoid unintended, unjustified, and inappropriately differential outcomes for users.

Regional Development Banks

Region: Africa

The African Development Bank (ADB)

Financial Inclusion Grant Program

The African Development Bank launched a financial inclusion grant program of $1 million in 2021 to develop AI-enabled systems to process customer complaints in several key local languages in Ghana, Rwanda,[221] and Zambia.[222] However, there does not appear to be any program, methods, or tools to facilitate implementation of AI governance by ADB at this time.

Region: Americas: Latin America and the Caribbean

Inter-American Development Bank (IDB)

Region: Americas

The Inter-American Development Bank (IDB), headquartered in Washington, D.C., is one of four major regional financial sector multilateral institutions.[223] In addition to lending money to countries in Latin America and the Caribbean (LAC), IDB provides member governments with technical assistance for a variety of projects intended to bring economic inclusion and growth to the region. IDB has actively assisted institutions to operationalize AI principles.[224] In 2019, IDB launched a significant regional alliance for ethical and responsible use of technology, with a focus on AI. The partnership is between public and private sectors, and includes civil society and academia. Called "fAIr LAC,"[225] this project seeks to assist regional governments as they navigate the ethical application of AI.

The program features a series of pilot projects; an observatory of responsible AI use cases; four regional hubs (Jalisco, Costa Rica, Colombia, and Uruguay); and frameworks and documents written in Spanish, Portuguese, and English for use by public servants, government ministers, and AI developers working for governments. The initiative has resulted in procedures for AI self-assessment by public sector and private entities, and practical

---

221 Ministry of ICT and Innovation, *The National AI Policy* (2022), https://www.minict.gov.rw/index.php?eID=dumpFile&t=f&f=67550&token=6195a53203e197efa47592f40ff4aaf24579640e (The Ministry of Information, Communication Technology and Innovation of Rwanda, developed by the Ministry of ICT and Innovation, in collaboration with the Rwanda Utilities Regulatory Authority (RURA), GIZ FAIR Forward, the Centre for the 4th Industrial Revolution Rwanda (C4IR), and The Future Society (TFS), published its National AI policy in April 2023).

222 *African Development Bank provides $1 million for AI-based national customer management systems in Ghana, Rwanda and Zambia,* African Development Bank Group, (March 10, 2021), https://www.afdb.org/en/news-and-events/press-releases/african-development-bank-provides-1-million-ai-based-national-customer-management-systems-ghana-rwanda-and-zambia-42602.

223 Jenny Ottenhoff, *Regional Development Banks*, Center for Global Development (Sept. 2011), https://www.cgdev.org/publication/regional-development-banks-abcs-ifis-brief.

224 In conjunction with research conducted for this report, World Privacy Forum interviewed Dr. Cristina Pombo Rivera, principal advisor and head of the digital and data cluster, Social Sector, Inter-American Development Bank, in April 2023.

225 *fAIr LAC*, Inter-American Development Bank, (2019) https://fairlac.iadb.org/en/fair-lac-box.

manuals such as a handbook for AI project directors for incorporating ethical considerations throughout the AI life cycle.[226]

As part of its fAIr LAC program, the IDB has also helped countries build AI-related projects, including public online dashboards detailing information about AI-based tools used by the government of Chile.[227]

IDB's fAIr LAC in a box

Tool Type: Catalog

IDB's fAIr LAC program produced a collection of five AI governance tools called fAIr LAC in a box.[228] Among other things, it includes a data science toolkit[229] featuring methods for reducing unintended bias in AI systems, as well as approaches for tracking features of machine learning systems and keeping track of actions taken to reduce problems. The handbook-style toolkit is complemented by a GitHub repository featuring suggested procedures and questions to be used at various stages of the AI life cycle. The repository also features Cuadernillos de trabajo, or workbooks featuring specific code for things such as evaluating prediction errors, fine-tuning model parameters, and evaluating model performance.

The data science handbook references some specific measures intended to reveal explanations for how AI systems make decisions, such as counterfactual explanations, Shapley values,[230] and integrated gradients for deep networks. In particular, the handbook mentions use of SHAP as a quantitative explainability method for deep neural networks, and includes it in a detailed workbook section.[231]

The collection also includes ethical self-assessment guides for public agencies and AI developers; an algorithmic impact audit guide for policymakers in Latin America and the Caribbean responsible for leading automatic decision system projects;[232] and a case study of an algorithmic audit of a clinical risk prediction system called Laura.[233]

---

226 Gabriela Deni et al., *Responsible use of AI for public policy: Project formulation manual*, IDB, (Aug. 2021), https://publications.iadb.org/publications/english/document/Responsible-use-of-AI-for-public-policy-Project-formulation-manual.pdf.

227 *Repositorio Algoritmos Públicos*, GobLab UAI, Escuela de Gobierno Universidad Adolfo Ibáñez Santiago (2022), https://algoritmospublicos.cl/; see also *Repositorio de Algoritmos Publicos de Chile, GobLab UAI (2022)*, https://goblab.uai.cl/wp-content/uploads/2022/02/Primer-Informe-Repositorio-Algoritmos-Publicos-en-Chile.pdf
(the platform is a public-private partnership implemented by Universidad Adolfo Ibáñez with support of IDB in conjunction with the government of Chile).

228 *FAIr LAC*.

229 *Responsible use of AI for public policy data science toolkit*, IDB (2020), https://publications.iadb.org/publications/english/document/Responsible-use-of-AI-for-public-policy-Data-science-toolkit.pdf .

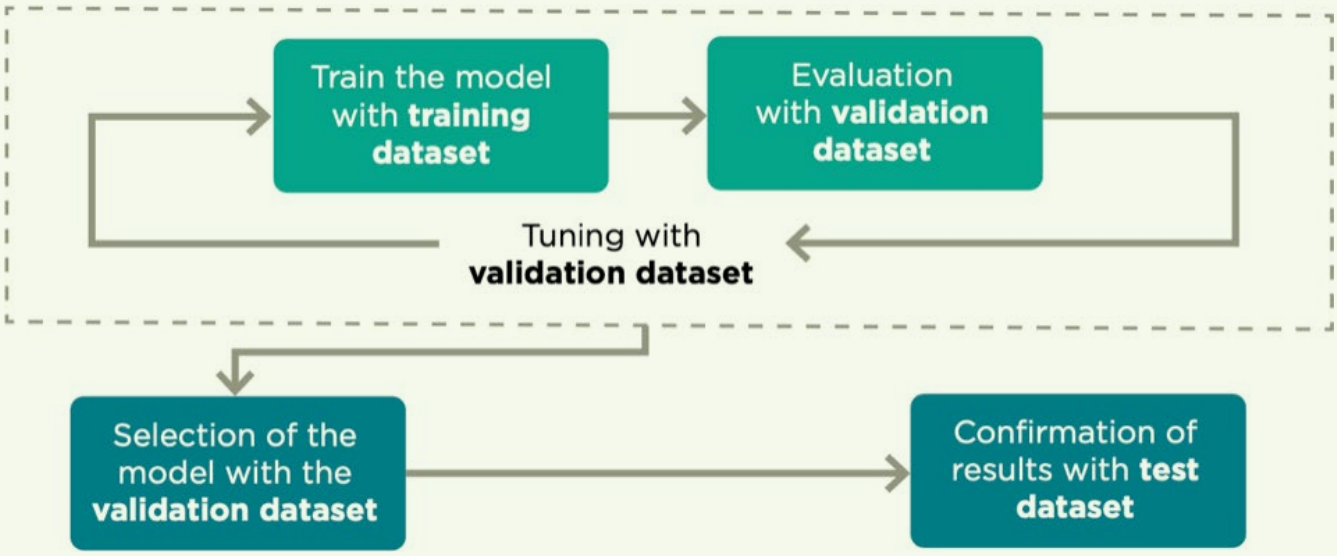
230 Use of SHAP for AI explainability has been scrutinized in scholarly literature. See Part I of this report for more information.

231 *Responsible use of AI for public policy data science toolkit*, *supra*, at 43 and 91-93.

232 Matías Aránguiz Villagrán, *Algorithmic Audit for Decision-Making or Decision Support Systems*, IDB (March 2022) https://publications.iadb.org/publications/english/document/Algorithmic-Audit-for-Decision-Making-or-Decision-Support-Systems.pdf.

233 Robot Laura Auditoría Algorítmica, IDB (Dec. 2021), https://publications.iadb.org/publications/english/document/Algorithmic-Audit-for-Decision-Making-or-Decision-Support-Systems.pdf.

Figure 6: ***Cross-Validation Model***



Source: Inter-American Development Bank, Responsible Use of AI for Public Policy: Data Science Toolkit

The case study mentions use of Aequitas, a tool that includes in its Fairness Criteria Assessments use of the Four-Fifths rule to measure disparate impact.[234] As noted in the Findings section of this report, use of SHAP and the Four-Fifths rule in AI measurement has drawn sharp criticism in scholarly literature reviewed here in Part I.

In relation to disparate treatment of subgroups, the case study makes a point of noting that although quantitative methods have been developed to capture and measure disparate impact on disadvantaged groups, some techniques do not reflect sociocultural context.

Rather than presenting the fAIr LAC in a box tools as infallible solutions, IDB's ongoing program is intended to inspire governments to consider critical questions and potential problems as part of the design and use of AI systems.

Region: Asia

Asia Development Bank (ADB)

Mapping poverty through data integration and AI

Programs involving AI are underway at the Asia Development Bank (ADB), including extensive work to map poverty more accurately in the Asian region through use of novel data types and AI.[235] The 2020 report, *Mapping poverty through data integration and Artificial Intelligence,* offers detailed guidance on how to prepare data for feasibility studies, how to ethically apply a convolutional neural network to poverty prediction, and advice on how to reduce bias and other undesirable components of sensitive poverty research.

234 aequitas/docs/output_data.html, DSSG (Data Science for Social Good), aequitas, GitHub, https://github.com/dssg

235 *Mapping poverty through data integration and Artificial Intelligence: A special supplement of the key indicators for Asia and the Pacific*, ADB, (Sept. 2020), https://www.adb.org/publications/mapping-poverty-data-integration-ai .

For example, the report provides three possible approaches to using big data and AI processes without contaminating the data samples with self-selection bias, including specific guidance regarding measurement to address these issues.[236]

The ADB's 2020 report, *AI in Social Protection — Exploring Opportunities and Mitigating Risks*, co-published with Germany's Agency for International Cooperation GIZ (Gesellschaft für Internationale Zusammenarbeit),[237] provides guidance for implementing ethical approaches for AI systems used in relation to social protection programs. The guidance addresses topics including inclusive data strategies and transparent and independent review of AI systems.

Region: Europe

The European Bank for Reconstruction and Development (EBRD)

Report: Approach to Accelerating the Digital Transition

In its 2021 report outlining its approach to accelerating the digital transition,[238] the European Bank for Reconstruction and Development set forth a series of pledges. The bank said it plans by 2025 to develop a legal and regulatory framework that promotes innovation, healthy competition in digital markets, and cybersecurity, while safeguarding financial stability and inclusion and ensuring diversity, the ethical use of artificial intelligence (AI), and appropriate data protection.[239] Thus far, the EBRD gives every indication that it is laying the foundations to implement ethical AI, but it has not yet published specific AI governance tools.

In the report, *The EBRD's approach to accelerating the digital transition, 2021-25*, the EBRD discusses digital transformation and AI, posing opportunities and risks related to issues ranging from privacy and bias problems to electronic waste. EBRD established its Digital Hub[240] in January 2022 to support and coordinate the implementation of its digital approach, which is intended to enable equal access to digital technology and skills, establish robust governance practices, and provide financial and technical support to companies and governments.

AI Governance Tools and Use Cases from National Governments and NGOs

This research unambiguously found that national governments and other organizations around the world throughout Africa, the Americas, Asia, Europe, Oceania, and the United Kingdom are crafting AI governance tools. Often, these tools are part of a larger national strategy on AI, with most of these efforts focused on moving from principles to practical guidance.

Observations of AI governance tools from this research indicate that national governments and NGOs have focused on operationalizing approaches to achieving AI fairness, explainability, robustness, and data minimization, among other important goals. AI governance tools reviewed here cover a wide range of formats, including self-assessment questionnaires, procedural workflow charts, detailed rules for AI system procurement, and

236 *Mapping poverty through data integration and Artificial Intelligence: A special supplement of the key indicators for Asia and the Pacific, supra,* at 6, 21, and 24.

237 *AI in Social Protection*, SocialProtection.org (2020), https://socialprotection.org/discover/publications/ai-social-protection.

238 *How the EBRD will achieve its digital transition,* EBRD (Nov. 2021), https://www.ebrd.com/ebrd-digital-transition.html.

239 *How the EBRD will achieve its digital transition, supra, at 3.*

240 *The EBRD Digital Hub Fact Sheet: Accelerating the Digital Transition*, EDRB (2022), https://www.ebrd.com/ebrd-digital-approach.html.

recommendations for oversight committees. Some feature scoring mechanisms to quantify risk or other aspects of AI systems. In more rare cases, governments and other organizations have provided detailed technical guidance and even technical software, all with the intent to assess AI systems and assuage particular problems.[241]

This section of the report primarily includes AI governance tools from national governments. In order to recognize efforts related to establishment of AI governance tools and practices taking place internationally, we also include here some work from NGOs and hybrid entities involving partnerships among NGOs, academia, and private industry. And, as previously mentioned, because the spectrum of AI development stages reflects a wide continuum across geographic regions, this report includes AI-related work that is building toward formal AI governance tools.

Africa: Northern Africa: Morocco

The International Artificial Intelligence Center of Morocco

AI Movement

The International Artificial Intelligence Center of Morocco, opened in November 2022,[242] has as its primary objective to deliver practical, resilient, and ethically sound approaches to AI-related challenges faced by society, the environment, the market, the economy, and technology.

The Center has launched a four-tiered certificate program called the "AI Governance and Applications" executive program, which includes a significant section on responsible and ethical AI and data management.[243]

A partner of the Sub-regional Forums on AI set up by UNESCO, the Center is involved with implementing the UNESCO Recommendation on the Ethics of AI in the African context through conferences and other work. These initiatives are conducted in relation to UNESCO's Global Priority Africa program.[244]

Africa: Eastern Africa: Kenya

Masakhane Research Foundation

Language Datasets for Africans, by Africans

AI systems require data as core inputs to learn how to recognize patterns in information and produce outputs such automated decisions and predictions. For example, datasets representing low-resourced languages are necessary to ensure that AI systems incorporating natural language processing (NLP) for products such as chatbots or mobile apps used to assist people in healthcare, finance, or social services are truly fair, inclusive, accurate, and trustworthy. However, for many low-resourced languages, high-quality datasets simply do not exist.

Work is underway to help ensure that the datasets needed to train AI models—reflecting the groups and communities who will use them or be affected by them—are built.

---

241 National governments and subnational governments are listed as *region: nation: subnational,* utilizing the M49/ISO- Standard.

242 AI movement, International Artificial Intelligence Center of Morocco, https://aim.um6p.ma/en/home/.

243 *AI Governance and Practice*, International Artificial Intelligence Center of Morocco, https://aim.um6p.ma/en/executive-master-ai-governance-practice/.

244 *Priority Africa Flagship Programmes and Actions*, UNESCO (May 11, 2023), https://www.unesco.org/en/africa-flagship-programmes.

Masakhane Research Foundation is a grassroots organization based in Kilifi, Kenya.[245] The group aims to ensure that low-resourced African language datasets exist, and that AI systems recognize African names, cultures, places, and history by conducting and strengthening NLP research in African languages: for Africans, by Africans.

Contributors to the Masakhane project are generating, curating, and annotating datasets that are inclusive of the languages people speak throughout the African continent.[246]

Masakhane roughly translates to "We build together" in isiZulu, one of South Africa's 12 official languages, which include Sepedi, Sesotho, Setswana, siSwati, Tshivenda, Xitsonga, Afrikaans, English, isiNdebele, isiXhosa, and South African Sign Language.[247]

> Masakhane aims for its multilingual scientific corpora of African research to "be used for the aims and goals shaped by Africans and in particular avoids exploitation from non-African big tech organizations."

The organization has launched a variety of initiatives designed to create new datasets including a project to deliver open, accessible, and high-quality text and speech datasets for low-resourced East African languages from Uganda, Tanzania, and Kenya.[248]

Another Masakhane project is intended to build a multilingual scientific corpora of African research translated into multiple African languages. That project will include creation of the Masakhane Ethical Manifesto for use as a reference for any NLP dataset creation efforts on the African continent. The manifesto will incorporate ubu-Ntu ethics, "notable for its strong focus on enriching relationships in ways that affirm human rights and human dignity through the equitable distribution of power, and individual and communal participation in mutually beneficial goals."[249]

Africa: Western Africa: Ghana

Kwame Nkrumah University of Science and Technology Responsible Artificial Intelligence Lab

FACETS Framework

Tool Type: Practical Guidance with Self-assessment Questions and Scoring Output

The Responsible Artificial Intelligence Lab (RAIL) at Kwame Nkrumah University of Science and Technology in Kumasi, Ghana[250] seeks to "improve the quality of life for all in Africa and beyond by partnering with Africa's science and policy communities to leverage AI through high-quality research, responsible innovation, and strengthening talent."

245 *Priority Africa Flagship Programmes and Actions*.

246 *Masakhane* (July 28, 2023, 10:24AM), https://www.masakhane.io.

247 *The NA Approves South African Sign Language as the 12th Official Language*, Parliament of the Republic of South Africa (May 3, 2023) http://www.parliament.gov.za/press-releases/na-approves-south-african-sign-language-12th-official-language.

248 *MakerereNLP: Text & Speech for East Africa*, Masakhane (July 28, 2023, 10:30AM), https://www.masakhane.io/ongoing-projects/makererenlp-text-speech-for-east-africa.

249 *Masakhane MT: Decolonise Science*, Masakhane (July 28, 2023, 10:30AM), https://www.masakhane.io/ongoing-projects/masakhane-mt-decolonise-science (according to Maskahane, "Corpora creation activity, in order to align with these local ethical principles, should affirm and enrich the human worth and dignity of Africans. To this extent, great concern should be taken to ensure that the creation of corpora should be used for the aims and goals shaped by Africans and in particular avoids exploitation from non-African big tech organizations").

250 Responsible Artificial Intelligence Lab (August 18, 2023, 6:53AM), https://rail.knust.edu.gh/.

Other notable African language dataset building projects include Ghana NLP, an open-source initiative focused on NLP in Ghanaian languages;[251] Lacuna Fund, a multistakeholder group building language and agriculture datasets;[252] and Mozilla Common Voice, an open-source voice database project from Mozilla Foundation featuring over 100 crowdsourced language datasets including several African language datasets.[253]

The lab is part of the Artificial Intelligence for Development in Africa program partnership between the Swedish International Development Agency and International Development Research Centre. It is also supported by GIZ, Germany's Agency for International Cooperation GmbH, through its FAIR Forward initiative.

RAIL introduced in January 2023 its framework for a quantitative approach to measuring responsible AI, called FACETS (Fairness, Accountability, Confidentiality, Ethics, Transparency, and Safety).[254] Based on established ISO 26000 social responsibility standards for businesses and organizations,[255] as well as other related frameworks for fairness, accountability, and transparency, FACETS is intended to provide a quantitative measure of ethical AI practices incorporated throughout the development of an AI system.

The framework includes an online self-assessment questionnaire[256] addressing the FACETS-related issues according to stages of an AI life cycle from origin to deployment, including data- and model-related questions. Self-assessment evaluators can respond "Yes," "No," or "I don't know."

Examples of questions featured in the FACETS assessment:

- "Did you consider inclusivity by putting people in the center from the beginning of the process (use-case definition, data collection, and system development)?"
- "Does the dataset follow acceptable standards best practices and specifications for data development like datasheets for datasets?"
- "Is the model published in a conference or journal?"
- "For Explainable AI (XAI) methods, is the explanation for the prediction shared in a decision report?"

Based on responses to the online self-assessment, the system calculates a numerical FACETS score.

---

251 *About Us,* GhanaNLP, (July 28, 2023, 10:33AM), https://ghananlp.org/about.

252 Lacuna Fund is supported by Rockefeller Foundation, Google.org, and Canada's International Development Research Centre, in addition to other development, philanthropic, and research institutions. The datasets produced by Lacuna Fund are licensed under the Creative Commons CC-BY 4.0 International license.

253 *Common Voice*, Mozilla (July 28, 2023, 10:35AM), https://commonvoice.mozilla.org.

254 *RAIL Introduces a Quantitative Approach to Measuring Responsible AI*, Responsible Artificial Intelligence Lab (Jan. 31, 2023), https://rail.knust.edu.gh/2023/01/31/rail-introduces-a-quantitative-approach-to-measuring-responsible-ai/.

255 *ISO 26000: Social responsibility*, Int'l Org. of Standardization (August 18, 2023, 6:51AM), https://www.iso.org/iso-26000-social-responsibility.html.

256 *Calculate the FACETS Score*, FACETS Responsible AI Framework (August 18, 2023, 6:58 AM), https://facets.netlify.app/facets#envision.

Americas: Latin America and the Caribbean: South America: Chile

Ethical, Responsible, and Transparent Algorithms Project and Procurement Requirements

Chile launched its Ethical, Responsible, and Transparent Algorithms Project in 2020.[257] The public-private partnership, established through funding from Inter-American Development Bank's IDB Lab, brings together the Universidad Adolfo Ibáñez and several public sector agencies to incorporate ethical standards in the purchase, use, and reporting of AI and automated decision algorithms by government agencies, as well as in the development of AI and automated systems by technology providers.

As part of the ongoing effort, the group has conducted pilots for the ethical implementation of automated systems in public institutions, and has developed guidelines, regulations, and other methods for ethical design and purchase of algorithmic systems.

Americas: Latin America and the Caribbean: South America: Chile

Government of Chile

ChileCompra Standard Bidding Terms for Data Science and AI Projects

Tool Type: Practical Guidance

Chile's Ethical, Responsible, and Transparent Algorithms Project has also led to new requirements for quality assurance for government agency procurement of AI and automated systems. Chile's purchasing and public procurement directorate, ChileCompra,[258] established its Standard Bidding Terms for Data Science and AI Projects in December 2022. Its Resolution No. 60 (Direccion de Compras y Contratación Publica Aprueba Formato Tipo de Bases Administrativas Para la Adquisición de Proyectos de Ciencia de Datos e Inteligencia Artificial) includes bidding terms, impact assessments, and guidance on measuring AI fairness and explainability. The requirements have been piloted with Chile's National Health Fund (FONASA) and its Public Criminal Defender's Office.[259]

> The project aims to incentivize industry players to build AI systems with responsible and trustworthy AI considerations if they want to win government contracts.

The project aims to incentivize industry players to build AI systems with responsible and trustworthy AI considerations if they want to win government contracts.[260] It uses these bidding and procurement requirements for public sector agencies to do that. The standard bidding conditions are promoted by ChileCompra to facilitate the participation of government suppliers in relation to large sum public sector tech acquisitions.

---

257 *Algoritmos Éticos*, GobLab UAI (Sept. 25, 2020), https://goblab.uai.cl/en/ethical-algorithms/.

258 *Direccion de Compras y Contratación Publica Aprueba Formato Tipo de Bases Administrativas Para la Adquisición de Proyectos de Ciencia de Datos e Inteligencia Artificial, Resolución N°60.*

259 *Ya se encuentra disponible Bases Tipo para licitar proyectos de algoritmos e inteligencia artificial con requisitos éticos*, ChileCompra (Jan. 30, 2023), https://www.chilecompra.cl/2023/01/ya-se-encuentra-disponible-bases-tipo-para-licitar-proyectos-de-algoritmos-e-inteligencia-artificial-con-requisitos-eticos/.

260 In conjunction with research conducted for this report, WPF interviewed Maria Paz Hermosilla, director of GobLab UAI in the Escuela Gobierno at Universidad Adolfo Ibáñez, in June 2023.

The AI procurement terms require that statistical equity metrics be taken into account to help conduct an ethical and responsible analysis of a system.[261] The terms also state that such metrics must be considered in a public sector agency's evaluation and choice of a model, based on the specific problem or context at hand.

ChileCompra's AI procurement requirement documentation suggests use of Datasheets for Datasets[262] to assist in detecting bias in data used to develop the system under assessment. It also suggests use of Model Cards for Model Reporting,[263] as well as risk analysis of personal data processing through an impact assessment.

In addition, the requirements make note of specific measures for AI explainability including counterfactual explanation, Local Interpretable Model-Agnostic Explanations (LIME), and the What-if Tool, an open-source tool original devised by Google. The What-if Tool attempts to assess the behavior of trained machine learning models, including models used for image recognition and conversational AI. Documentation associated with the What-if Tool features use of SHAP to reveal feature importance to analyze model fairness.[264]

As noted in the Findings section of this report, use of LIME and SHAP for AI explainability has drawn sharp criticism in scholarly literature reviewed here in Part I.

Americas: Latin America and the Caribbean: South America: Chile

Universidad Adolfo Ibáñez, Government of Chile and IDB

Repositorio Algoritmos Públicos

Universidad Adolfo Ibáñez's GobLab UAI is the public innovation laboratory of the research university's School of Government. To create more transparent use of algorithmic systems, the university helped build an online dashboard detailing information about algorithmic and AI-based tools used in the public sector. The platform was built with support from The Inter-American Development Bank in conjunction with the government of Chile.[265]

The digital platform featured information on 75 different automated systems used by public institutions in Chile by the end of 2022, from rules-based systems to deep learning and neural network-based systems.[266] Systems are categorized in the platform according to government divisions such as education, health, defense, and economic affairs.

---

261 *Direccion de Compras y Contratación Publica Aprueba Formato Tipo de Bases Administrativas Para la Adquisición de Proyectos de Ciencia de Datos e Inteligencia Artificial, Resolución N°60*, *supra,* at 54.

262 Timnit Gebru et al., *Datasheets for datasets*, Commun. 64 ACM 12, 86-92 (Dec. 2021), https://doi.org/10.1145/3458723.

263 Margaret Mitchell et al., *Model Cards for Model Reporting*, FAT* '19 *Proceedings of the Conference on Fairness, Accountability, and Transparency,* Ass'n for Computing Machinery 220–229 (Jan. 29, 2019), https://doi.org/10.1145/3287560.3287596.

264 What-If Tool and SHAP on COMPAS keras model, GitHub, PAIR-code, what-if-tool, https://github.com/PAIR-code/what-if-tool/blob/master/WIT_COMPAS_with_SHAP.ipynb

265 *Repositorio Algoritmos Públicos Inform Anual 2023*, GobLab UAI, Escuela de Gobierno Universidad Adolfo Ibáñez Santiago (2023), https://www.algoritmospublicos.cl/static/Informes/GobLab-UAI_Informe_Repositorio_Algoritmos_Publicos_2023.pdf.

266 *Repositorio Algoritmos Públicos Inform Anual 2023*, GobLab UAI, Escuela de Gobierno Universidad Adolfo Ibáñez Santiago (2023), https://www.algoritmospublicos.cl/static/Informes/GobLab-UAI_Informe_Repositorio_Algoritmos_Publicos_2023.pdf.

Americas: Northern America: Canada

Government of Canada

Algorithmic Impact Assessment Tool

Tool Type: Practical Guidance with Self-assessment Questions and Scoring Output

Canada's Algorithmic Impact Assessment Tool,[267] part of Canada's responsible use of artificial intelligence policy work, is mandatory for federal government institutions. It comes in the form of a questionnaire, and is answered by government agencies intending to use algorithmic systems. Inquiries address the purpose for AI systems, how they are designed, and how data is sourced and prepared.

The Algorithmic Impact Assessment (AIA) is designed to score the impact level of an algorithmic system according to factors including its design, algorithm, decision type, and data. Impact levels are classified from "little to no impact" to "very high impact" in relation to individual rights, health and well-being of individuals or communities, economic interests, and sustainability of an ecosystem.

Depending on impact level, Canada's AIA process requires peer review, data bias and quality testing, data governance implementation, analysis using Canada's "Gender-based Analysis Plus" analysis method,[268] human intervention during the decision-making process, and meaningful explanation of decision results.

Canada's AIA reports are available in Canada's Open Government Portal.[269]

AIA Use Case: Canada's Immigration, Refugees, and Citizenship Department

Advanced Analytics Triage of Visitor Record Applications

A December 2022 report reflecting the AIA of a system for Advanced Analytics Triage of Visitor Record Applications at Canada's Immigration, Refugees, and Citizenship department (IRCC) provides a case study of what Canada's AIA process looks like in practice. According to the report, the triage analytics system was scored as impact level 2, reflecting a moderate impact level.[270]

The report shows responses to questions about the triage system, which is intended to "help IRCC decision-makers process applications more efficiently." For example, when asked, "Will the system be replacing human decisions that require judgement or discretion?" IRCC answered, "Yes," which attributed four points in the scoring system. According to an official description of Canada's AIA process, "The value of each question is weighted based on the level of risk it introduces or mitigates in the automation project."[271]

---

267 *Algorithmic Impact Assessment tool*, Government of Canada.

268 *Gender-based Analysis Plus (GBA Plus)*, Government of Canada (July 28, 2023, 7:21AM). https://women-gender-equality.canada.ca/en/gender-based-analysis-plus.html.

269 *Open Government Portal, Collection Type: Algorithmic Impact Assessment*, Government of Canada (May 2023), https://search.open.canada.ca/opendata/?collection=aia&page=1&sort=date_modified+desc (as of May 2023, there were 11 AIA collection type results available).

270 *Algorithmic Impact Assessment - Advanced Analytics Triage of Overseas Temporary Resident Visa Applications*, Open Government Portal for Government of Canada, https://open.canada.ca/data/en/dataset/6cba99b1-ea2c-4f8a-b954-3843ecd3a7f0.

271 *Algorithmic Impact Assessment tool*, Government of Canada, *supra,* 2.1 scoring.

Americas: Northern America: United States

US federal agencies and the White House have produced multiple sets of voluntary guidance and procedural approaches to incorporating accountability, fairness, data privacy, and risk management into AI development and use in the US.

Americas: Northern America: United States

US Government Accountability Office

An Accountability Framework for Federal Agencies and Other Entities

Tool Type: Process Framework with Self-Assessment Questions

The US Government Accountability Office (GAO) published *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities* in 2021.[272] The guidance provides questions and audit procedures for promoting AI accountability, ensuring data quality, producing results that are consistent with objectives, and monitoring for reliability and relevance of AI systems over time.

The GAO's accountability framework mentions IBM's AI Fairness 360[273] as an example of "guidance on incorporating ethical principles such as fairness, accountability, transparency, and safety in AI use."[274]

In addition, the GAO framework mentions Microsoft's model drift monitoring guidance,[275] [276] and the use of data and model cards for data transparency purposes.[277]

Americas: Northern America: United States

US General Services Administration

The Artificial Intelligence Governance Toolkit

Tool Type: Practical Guidance and Process Framework with Self-Assessment Questions

272 *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, U.S. Government Accountability Office (June 2021), https://www.gao.gov/products/gao-21-519sp.

273 Documentation for the Disparate Impact Remover algorithm supported by AI Fairness 360 specifically cites 2015 research introducing a disparate impact measurement based on the Four-Fifths Rule's 80% benchmark. As noted in the Findings section of this report, this approach has drawn sharp criticism in scholarly literature reviewed here in Part I. See Appendix C for more detail. AIF360, GitHub, Trusted AI, Supported Bias Mitigation Algorithms, "Disparate Impact Remover," https://github.com/Trusted-AI/AIF360/tree/master

274 *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities, supra,* at 30.

275 Sushrut Shendre, *Model Drift in Machine Learning*, Towards Data Science via Medium (May 13, 2020), https://towardsdatascience.com/model-drift-in-machine-learning-models-8f7e7413b563.

276 *Detect data drift (preview) on datasets,* Microsoft Azure Machine Learning (Aug. 8, 2023), https://learn.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?view=azureml-api-1&tabs=python.

277 Margaret Mitchell et al.

The Artificial Intelligence Governance Toolkit[278] provides procedural guidance from the US General Services Administration (GSA). Published in 2022, it is complementary work to the GAO report on AI. The GSA's AI governance toolkit addresses both data privacy and data governance.

Data privacy and the substantial data collection and use that fuels AI are not always front and center in AI governance tools and guidance. In the GSA's toolkit, they are. Therein, the process guidance and questions to be considered throughout the AI development life cycle focus on privacy-related issues informed by the US Privacy Act's Fair Information Practice Principles.[279]

For instance, the GSA framework provides a series of questions addressing the origin and provenance of data used for AI, as well as whether data is sufficiently representative in order to prevent bias. It also recognizes the need for entities to minimize the amount of sensitive or personally-identifiable data they collect or process. To inspire sensitive data minimization, it asks, "Can you achieve similar/effective results with less (PII) data?"

Americas: Northern America: United States

National Institute of Standards and Technology (NIST)

Artificial Intelligence Risk Management Framework

Tool Type: Practical Guidance and Process Framework with Self-Assessment Questions

The National Institute of Standards and Technology published its Artificial Intelligence Risk Management Framework (RMF)[280] and companion AI RMF Playbook in January 2023. The framework and playbook are organized according to four broad categories or functions present in the AI life cycle: Govern, Map, Measure, Manage. See more detailed discussion of this important framework in the Standards section in Part II of this report.

Americas: Northern America: United States

White House

The Blueprint for an AI Bill of Rights and the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Tool Type: Practical Guidance

The Blueprint for an AI Bill of Rights[281] was published by the White House Office of Science and Technology Policy in October 2022. Rather than providing prescriptive procedures, AI assessment steps, or technical method recommendations, the Blueprint for an AI Bill of Rights includes a wealth of practical guidance for design, use, and deployment of automated systems based on five principles for protecting people's rights.

---

278 *The Artificial Intelligence Governance Toolkit*, General Services Administration (2022), https://coe.gsa.gov/docs/AICoP-AI-GovernanceToolkit.pdf.

279 Robert Gellman, *From the filing cabinet to the cloud: Updating the Privacy Act of 1974*, World Privacy Forum (May 2021), https://www.worldprivacyforum.org/2021/05/from-the-filing-cabinet-to-the-cloud-updating-the-privacy-act-of-1974/.

280 *Artificial Intelligence Risk Management Framework*.

281 *The Blueprint for an AI Bill of Rights*, Office of Science and Technology Policy, https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

In particular, the voluntary guidance looks to protections mandated by the US Constitution or implemented under existing US laws. It describes protections that should be applied with respect to all automated systems that have the potential to meaningfully impact individuals' or communities' ability to exercise civil rights, privacy, and freedom of speech. It also ensures protections from discrimination and unlawful surveillance; and emphasizes access to education, housing, credit, employment, healthcare, and more.

In guidance related to protecting against algorithmic discrimination, the blueprint states that "for every instance where the deployed automated system leads to different treatment or impacts disfavoring the identified groups, the entity governing, implementing, or using the system should document the disparity and a justification for any continued use of the system."[282]

About a year after publishing the Blueprint for an AI Bill of Rights, on October 30, 2023, the White House unveiled its Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.[283] The sweeping order calls on multiple US federal agencies, including the Department of Commerce, NIST, the Department of Defense, the Department of Health and Human Services (HHS), the Department of Homeland Security, and others, to devise new standards, procedures, testing and measurement benchmarks, and reporting requirements related to AI safety and security, responsible innovation and competition, privacy and civil rights, worker and consumer protection, and health-related AI impacts.

For example, it calls on HHS to develop an AI assurance policy to evaluate important aspects of the performance of AI-enabled healthcare tools and infrastructure needs for enabling pre-market assessment and post-market oversight of AI-enabled healthcare-technology algorithmic system performance using real-world data.

The order calls for the launch of a pilot program implementing a National AI Research Resource (NAIRR); the program plan should address infrastructure, governance mechanisms, and user interfaces for initial integration of distributed computational, data, model, and training resources to be made available to the research community. In addition to use of private industry technology infrastructure,[284] plans for NAIRR also allow for private industry use of the resource.[285] [286]

The order also calls on federal agencies to designate a chief artificial intelligence officer to hold primary responsibility for coordinating their agency's use of AI, promoting AI innovation in their agency, and managing risks from their agency's use of AI.

Americas: Northern America: United States

US-Based Multistakeholder/Corporate Programs

Open Loop

---

282 *The Blueprint for an AI Bill of Rights*, *supra*, at 27.

283 *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, US White House (Oct. 30, 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

284 Kate Kaye, *Google's multicloud national AI research plan could cost $500M a year. It wants first crack at the data*, Protocol (Dec. 9, 2021), https://www.protocol.com/enterprise/google-ai-cloud-research.

285 *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource*, Nat'l Science Found. and White House Office of Science and Tech. Policy, 18, 22 (Jan. 2023), https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf.

286 Kate Kaye, *Startups are likely to get access to the national AI research cloud*, Protocol (May 25, 2022), https://www.protocol.com/bulletins/startups-national-ai-research-cloud.

Overseen by prominent corporate AI developer Meta, Open Loop is an experimental, multinational, multistakeholder tech governance program.[287] It involves private AI startups and representatives from academia, civil society, and governments. The program aims to "co-create and test new governance frameworks through policy prototyping programs, and to support the evaluation of existing legal frameworks through regulatory sandbox exercises."[288]

Meta's Open Loop team coordinated in 2022 and 2023 with AI startups, academia, and civil society and trade groups, as well as Uruguay's Agency for Electronic Government, the Information and Knowledge Society, and the Inter-American Development Bank, to test operational guidance for implementing Privacy Enhancing Technologies to reduce privacy risks of AI and other types of technical systems.[289]

Open Loop also recently led a policy prototyping effort on AI Transparency and Explainability in partnership with Singapore's Infocomm Media Development Authority and Personal Data Protection Commission. A July 2022 report about the initiative details a variety of algorithmic model types, the degree to which they can be interpreted for explainability purposes, and the limitations of their interpretability.[290] The report also addresses the limits of counterfactual explanatory strategies and SHapley Additive exPlanation, noting "several drawbacks of SHAP"[291]

The Meta-led Open Loop program also worked with several organizations in Europe to explore and test "alternative policy frameworks and regulatory pathways" for gauging the impacts of automated decision systems.[292] In part a response to EU legislative proposals calling for AI risk assessment, Open Loop's impact assessment "playbook" includes a step-by-step risk assessment methodology and examples of risk "mitigation" measures.[293] A detailed overview of the project is featured in Open Loop's January 2021 report, *AI Impact Assessment: A Policy Prototyping Experiment*.

The report presents steps for quantifying the severity of risks associated with automated systems, and suggests a "proper metric" for measuring AI system accuracy that takes into account the tradeoffs "between recall (no false negatives) and precision (no false positives)."[294] The report also highlights Facebook's previously stated desire for "self-assessment of AI risk" [295] and recommends that regulation requirements vary "in accordance with the specific AI application in question and the level and extent of the risks assessed, alongside the calculus of the benefits that application brings."[296]

---

287 Meta, https://about.meta.com/.

288 Open Loop (July 28, 2023, 8:11AM), https://openloop.org.

289 *Privacy Enhancing Technologies (PETs) (Uruguay),* Open Loop (July 28, 2023, 8:14AM), https://openloop.org/programs/open-loop-uruguay-program/.

290 Norberto Nuno Gomes de Andrade, *AI Transparency and Explainability - A Policy Prototyping Experiment* (2022), https://openloop.org/wp-content/uploads/2022/07/AI_Transparency_&_Explainability_A_Policy_Prototyping_Experiment.pdf.

291 Norberto Nuno Gomes de Andrade, *AI Transparency and Explainability - A Policy Prototyping Experiment* (2022), https://openloop.org/wp-content/uploads/2022/07/AI_Transparency_&_Explainability_A_Policy_Prototyping_Experiment.pdf. See p. 60. ("Of the several drawbacks of SHAP, the most practical one is that such a procedure is computationally burdensome and becomes intractable beyond a certain threshold").

292 Norberto Nuno Gomes de Andrade & Verena Kontschieder, *AI Impact Assessment: A Policy Prototyping Experiment* Open Loop, 15 (Jan. 2021), https://openloop.org/wp-content/uploads/2021/01/ AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf.

293 Norberto Nuno Gomes de Andrade, *supra,* at 43.

294 Norberto Nuno Gomes de Andrade, *supra,* at 90 (ADIA Prototype Law; Model data).

295 Norberto Nuno Gomes de Andrade, *supra,* at 13 (Automated Decision Impact Assessment).

296 Norberto Nuno Gomes de Andrade, *supra,* at 7 (Recommendations; Leverage a procedural risk assessment approach to determine what is the right set of regulatory requirements that apply to organisations deploying AI applications).

Open Loop's AI Impact Assessment report also presents a prototype for a law that would call for AI developers to consult supervisory authorities prior to deployment of automated decision-making systems if their assessments, possibly conducted by the AI developer itself, indicate a high risk.[297]

Partnership on AI

About ML Program

The Partnership on AI is a non-profit organization founded in 2016 by Amazon, DeepMind, Google, Microsoft, IBM, and Meta (at the time known as Facebook).[298]

As part of its About ML program, the organization evaluated 152 explainable AI, or XAI, tools intended to help explain why AI systems make decisions. The initial outcome of that effort, published in 2022, was a framework for documentation of XAI tools according to 22 tool dimensions such as tool type, intended users, technical compatibilities, and datasets used to build the tools.[299] In addition, the analysis identified dimensions of usability, such as the scope and formats of explanations the tools produced.

More recently in 2023, PAI published case studies of pilot projects involving ML life cycle documentation under its About ML initiative.[300]

The group also created a framework in 2021 to help guide decisions regarding AI data sourcing and related services.[301]

Figure 7: *Dimensions of Explainable AI Tools* Usability

DIMENSION	DEFINITION
Explanation Type	The technical (e.g. summary statistics) and non-technical formats (e.g. plain english explanations) in which the explanations are available to the users
Explainability Enhancing Features	Additional explainability metrics and attributes that make explanations more human interpretable
User Specifc Explanations	Ability to customize explanations based on the ML stakeholders' profle
Explanation Documentation	Capabilities to automatically provide documentation of the explanations
Use case	Information about the scope of tool in a particular application domain
Guidance for use	The support provided to choose the explanation algorithms

Source: Partnership on AI

297 Norberto Nuno Gomes de Andrade, *supra,* at 76 (ADIA Prototype Law; Risk management and governance).

298 Alex Hern, *Partnership on AI formed by Google, Facebook, Amazon, IBM, and Microsoft*, The Guardian, (Sept. 28, 2016), https://www.theguardian.com/technology/2016/sep/28/google-facebook-amazon-ibm-microsoft-partnership-on-ai-tech-firms.

299 Surya Karunagaran, *Making it easier to compare the tools for explainable AI*, PAI (June 30, 2022) https://partnershiponai.org/making-it-easier-to-compare-the-tools-for-explainable-ai/

300 Jiyoo Chang, *Improving Documentation in Practice: Our First ABOUT ML Pilot*, Partnership on AI (Oct. 11, 2022), https://partnershiponai.org/improving-documentation-in-practice-our-first-about-ml-pilot/; *see also ABOUT ML in Practice An Example from the Humanitarian Sector*, Partnership on AI (Jan. 18, 2023), https://partnershiponai.org/wp-content/uploads/2023/01/PAI_about-ml-in-practice-UNOCHA.pdf.

301 *Responsible sourcing of data enrichment services*, PAI (June 17, 2021), http://partnershiponai.org/wp-content/uploads/2021/08/PAI-Responsible-Sourcing-of-Data-Enrichment-Services.pdf.

Asia: Eastern Asia: Japan

Government of Japan: Governance Guidelines for Implementation of AI Principles

Japan has been very active in the area of AI tools and governance, with notable AI governance efforts becoming public as early as 2016. In 2019, Japan published its Social Principles of Human-centric AI, which played a role in shaping the OECD AI Principles.[302] Version 1.1 of its detailed principles for implementing AI in society, Governance Guidelines for Implementation of AI Principles, published in 2022,[303] address risk analysis in relation to social aspects of AI and gap analysis in system design, as well as management of AI systems. Unique to the Japanese approach are the segments regarding evaluation, verification, and re-analysis of conditions and risks—which are classic, advanced governance principles.

Appendix 1 of the Guidelines includes specific action targets and practical examples, as well as a section on how to implement agile governance in AI systems. This is a unique contribution to AI governance tools.

Government of Japan: Contract Guidelines on Utilization of AI and Data

Japan's Ministry of Economy, Trade, and Industry in 2019 created non-binding guidelines for how risk mitigation—including consumer and social protections, transparency, and safety thresholds—should be included in contracts for the development or utilization of AI software.[304]

Monetary Authority of Singapore (MAS)

Veritas Initiative

Tool Type: Practical Guidance and Process Framework with Self-assessment Questions and Technical Code

The Monetary Authority of Singapore (MAS), along with the financial industry, co-created what is now known as the MAS 2018 Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT)[305] in late 2018. Work by MAS to develop the FEAT principles as applicable to the financial sector is already in use regionally through the Asia Development Bank.

---

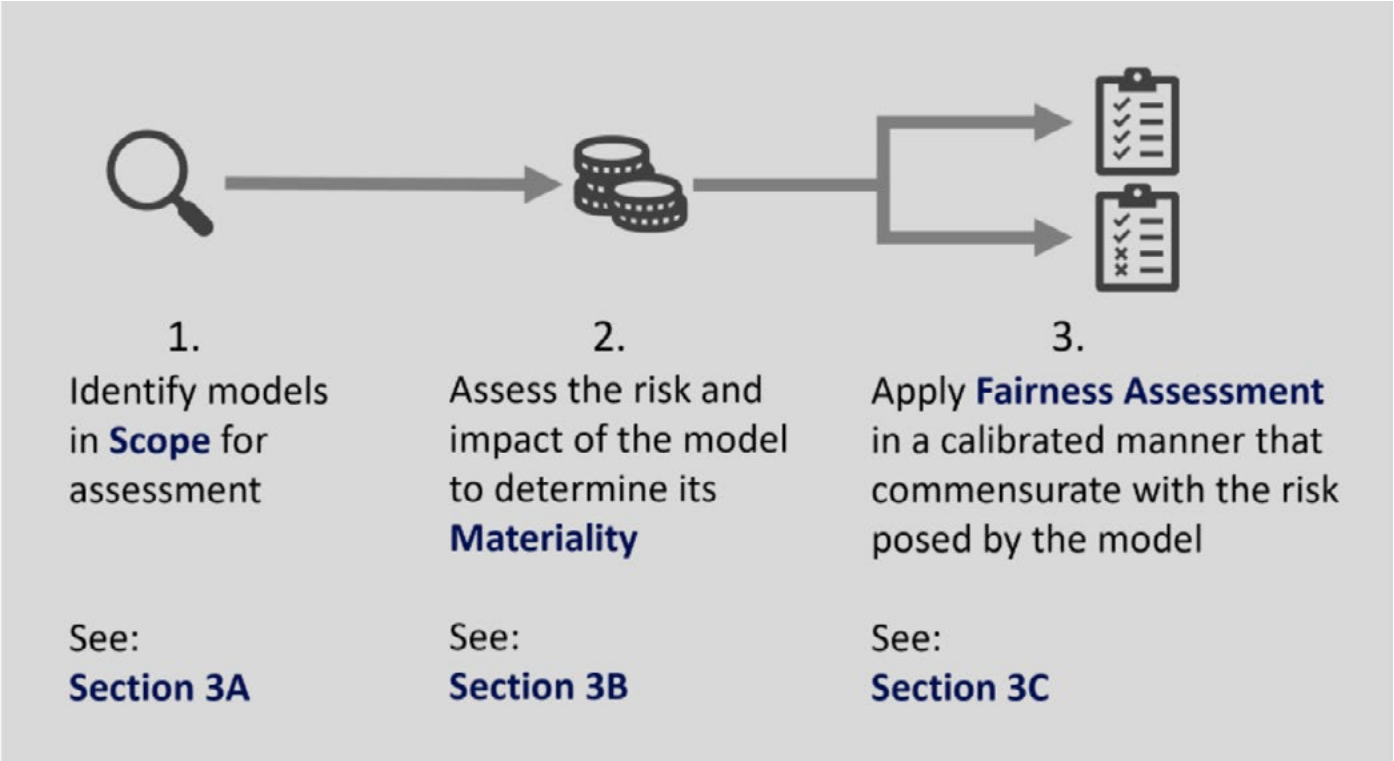
302 Japan was an important negotiator of the OECD AI Principles, and utilized its draft principles in discussions, as observed by WPF Executive Director Pam Dixon in 2018 and 2019, who was part of that process.

303 AI Governance Guidelines WG, *Governance Guidelines for Implementation of AI Principles*, Ver. 1.1, Ministry of Econ., Trade and Indus. (Jan. 28, 2022), https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf#page16.

304 *"Practical Guidebook on Data Provision for Fostering Human Resources of Experts in AI and Data Science" Formulated*, Ministry of Economy, Trade and Industry (METI), (Nov. 14, 2023), https://www.meti.go.jp/english/press/2021/0301_003.html.

305 *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector*, Monetary Authority of Singapore (Nov. 12, 2018), https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/feat.

Figure 8: ***Financial AI and Data Analytics Model Management***



Source: Monetary Authority of Singapore

To provide further concrete guidance regarding how to implement these principles in the financial sector, MAS worked with a consortium of financial institutions and technology companies. This consortium, called the Veritas consortium, was launched in November 2019 with the specific goal of cooperating to provide financial institutions verifiable ways of incorporating FEAT principles, including tools.[306] The initiative aims to devise a framework for financial institutions to promote adoption of responsible AI and data analytics.

The Veritas initiative began work on Phase 2 of the effort in January 2021, specifically addressing use cases regarding credit risk scoring, customer marketing, predictive underwriting, and fraud detection. These use cases were evaluated in relation to methods used to promote the FEAT principles.[307] Notably, the MAS implementation report, published in 2022,[308] provides a discussion of observations of the implementation of fairness principles by financial institutions in selected use cases.

The ongoing Veritas effort includes technical code for executing a fairness and transparency diagnosis of specific use cases. It also includes a question-based and procedural method for assessing fairness risks, along with practical guidance regarding the implementation of fairness principles for use of AI and machine learning by financial institutions.

306 *The Veritas Initiative*, MAS (Oct. 26, 2023), https://www.mas.gov.sg/schemes-and-initiatives/veritas.

307 *Id.*, at Phase 2.

308 *Implementation of fairness principles in financial institutions' use of Artificial Intelligence/Machine Learning: Observations from a thematic review,* MAS (June 2022), https://www.mas.gov.sg/-/media/mas-media-library/publications/monographs-or-information-paper/imd/2022/info-paper-on-implementation-of-fairness-principles-in-fis-use-of-aiml-final.pdf.

In addition, FEAT Fairness Principles Assessment Methodology[309] guidance published in 2022 mentions IBM's AI Fairness 360[310] and Microsoft's Fairlearn,[311] an open-source set of fairness metrics, algorithms, and other resources originally created by Microsoft Research.[312] The guidance also references use of the Four-Fifths Rule to measure disparate impact and lists several examples of tools that employ it. As noted in the Findings section of this report, AI fairness methods based on the rule have drawn sharp criticism in scholarly literature reviewed here in Part I.

An earlier FEAT Fairness Principles Assessment Methodology published in 2020 acknowledges that "all fairness measures capture different notions of fairness, each with their own limitations. Which of these notions of fairness are important for a particular system, and whether any are sufficient, is a context-dependent ethical question with no objective or universally accepted answer."[313]

Asia: South-Eastern Asia: Singapore

Singapore Infocomm Media Development Authority

AI Verify

Tool Type: Practical Guidance and Technical Framework with Technical Software

Compared to other tools reviewed in this report, Singapore has taken what might be considered the most prescriptive approach to implementing AI principles. What stands out is Singapore's development of a detailed technical testing framework and use of software designed for its program.

Singapore's Infocomm Media Development Authority in 2022 launched the international pilot of AI Verify,[314] an AI governance testing framework that includes software. In addition to process guidance, AI Verify includes open-source plugin tools addressing AI robustness, fairness, and explainability. The AI Verify program is intended for companies' use to validate the performance of their AI systems through standardized self-testing.

The program also features a process checklist addressing data governance, explainability, fairness, human agency and oversight, inclusive growth and societal and environmental well-being, reproducibility, robustness, safety, security, and transparency.

---

309 *Veritas Document 3A FEAT Fairness Principles Assessment Methodology,* MAS, 60 (July 28, 2023, 8:08AM), https://www.mas.gov.sg/-/media/mas-media-library/news/media-releases/2022/veritas-document-3a---feat-fairness-principles-assessment-methodology.pdf.

310 *AI Fairness 360 (AIF360), Catalogue of Tools and Metrics for Trustworthy AI*, OECD.AI (Sept. 9, 2022), https://oecd.ai/en/catalogue/tools/aif360 (IBM created AI Fairness 360 to address bias in machine learning algorithms and donated it to the Linux Foundation for open-source use in 2020. *See also*: Todd Moore et al. *IBM and LFAI move forward on trustworthy and responsible AI*, Open Source @ IBM (June 29, 2020), https://www.ibm.com/opensource/blogs/lfai-move-forward-trustworthy-ai/.Although AI Fairness 360 features several "bias mitigation" algorithms, its Disparate Impact Remover algorithm references as its key source research published in 2015 that adopts "a generalization of the 80 percent rule," also known as the Four-Fifths Rule. As noted in the Findings section of this report, this approach has drawn sharp criticism in scholarly literature reviewed here in Part I). *See also*: AIF360, GitHub, Trusted AI, Supported Bias Mitigation Algorithms, "Disparate Impact Remover." (https://github.com/Trusted-AI/AIF360/tree/master.)

311 Roman Lutz, Fairlearn: Assessing and Improving Fairness of AI Systems, J. Mach. Learn. Res. 24 (2023): 257:1-257:8.

312 *Veritas Document 3A FEAT Fairness Principles Assessment Methodology*.

313 *Veritas Document 1, FEAT Fairness Principles Assessment Methodology,* MAS, 75 (Dec. 2020), https://www.mas.gov.sg/-/media/mas/news/media-releases/2021/veritas-document-1-feat-fairness-principles-assessment-methodology.pdf.

314 *AI Verify AI Governance Testing Framework & Toolkit*, Singapore Infocomm Media Development Authority (2022), https://www.imda.gov.sg/content-and-news/press-releases-and-speeches/press-releases/2022/singapore-launches-worlds-first-ai-testing-framework-and-toolkit-to-promote-transparency-invites-companies-to-pilot-and-contribute-to-international-standards-development. See also: https://file.go.gov.sg/aiverify.pdf.

AI Verify was developed based on internationally-recognized AI principles from OECD as well as principles from the European Union.[315] AI Verify also integrates Singapore's own Model AI Governance Framework, established by its Personal Data Protection Commission in 2019, and updated in 2020.[316]

Thus far, companies in the financial, healthcare, human resources, and technology sectors have tried the AI Verify toolkit.[317]

In June 2023, Singapore's IMDA established the not-for-profit AI Verify Foundation[318] as a forum where people from the international open-source community could contribute to ongoing development of AI Verify testing frameworks, codebase, standards, and best practices. AI Verify developer tools are available to the open-source software community.[319]

The foundation's "premier" members tasked with setting strategic direction and a development plan for AI Verify include Aicadium, Google, IBM, IMDA, Microsoft, Red Hat, and Salesforce. The Foundation also has more than 60 general members.[320]

Technical components of AI Verify

Unlike any other tool reviewed in this report, Singapore's AI Verify features specifically designed and piloted software. The AI Verify software is intended for companies to download and use in their own enterprise environments by importing their AI models and running technical tests locally.

Following installation, the software walks AI developers and evaluators through a series of 11 process checks and technical tests featuring 85 testable criteria including for fairness classification, explainability, robustness, and image corruption.[321] To assess AI system fairness, AI Verify requires developers to evaluate testing and ground truth datasets used in AI model training.

Other features of AI Verify:

- A technical plugin that uses an algorithm to compute and display a list of fairness metrics to measure how correctly a model predicts among a given set of sensitive features (such as how it allocates job opportunities, loans, or medical assistance among demographic groups reflected in the data)
- A technical plugin that intentionally clutters a dataset with unwanted data "noise" to test the robustness of an AI model

---

315 *Ethics Guidelines for Trustworthy AI*, The European Commission (June 2018), https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

316 *Model AI Governance Framework, Second Edition,* Singapore Personal Data Protection Comm'n (Jan. 2020), https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

317 *AI Verify toolkit announcement*, 2022. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2022/sg-launches-worlds-first-ai-testing-framework-and-toolkit-to-promote-transparency

318 *Singapore launches AI Verify Foundation to shape the future of international AI standards through collaboration*, Infocomm Media Development Auth. of Singapore and Informa Tech (June 7, 2023), https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/singapore-launches-ai-verify-foundation-to-shape-the-future-of-international-ai-standards-through-collaboration.

319 *IMDA-BTG/aiverify-developer-tools*, GitHub (November 5, 2023), https://github.com/IMDA-BTG/aiverify-developer-tools.

320 *Foundation Members,* AI Verify Found.(July 28, 2023) https://aiverifyfoundation.sg/foundation-members/.

321 *How It Works*, AI Verify User Guide (Oct. 17, 2023 10:01AM) https://imda-btg.github.io/aiverify/introduction/how-it-works/#process-checks (process checks).

- A technical plugin that uses SHAP to explain how AI system features affect overall predictions
- Reports based on results of the AI Verify technical and process tests and checks, offering details on ways companies can fine-tune their AI models for improvement[322]

A 2022 AI Verify project pilot document[323] stated that AI Verify included the Adversarial Robustness Toolbox and AI Fairness 360, both of which were originally developed by IBM and donated in 2020 to the Linux Foundation AI Foundation.[324] It also included Fairlearn, an open-source set of metrics, algorithms, and other resources originally created by Microsoft Research.[325]

As noted in the Findings section of this report, use of SHAP for AI explainability, as well as research that formed the foundation of the AI Fairness 360 Disparate Impact Remover algorithm, have drawn sharp criticism in scholarly literature reviewed here in Part I.

According to AI Verify documentation, AI Verify "does not define ethical standards" and "does not guarantee that any AI system tested under this Framework will be free from risks or biases or is completely safe."[326]

Asia: South-Eastern Asia: Singapore

Singapore Infocomm Media Development Authority

Generative AI (Gen AI) Evaluation Sandbox Evaluation Catalogue

Tool Type: Practical Guidance

Singapore's Infocomm Media Development Authority is among the first government agencies to recommend specific approaches to evaluating generative AI systems. On October 31, 2023, along with the AI Verify Foundation, the authority introduced its Generative AI (Gen AI) Evaluation Sandbox,[327] which features an Evaluation Catalogue including baseline methods and recommendations for Large Language Models (LLMs).[328]

Although the Evaluation Catalogue notes the limitations and early stage of development of evaluation methods for LLMs, it nonetheless recommends specific evaluations including approaches that produce scores. The recommended evaluation methods directly reflect the 11 AI ethics principles listed in the AI Verify Framework, which include explainability, reproducibility, robustness, fairness, data governance, human agency and oversight, and security. In some cases the principles are translated to concepts more applicable to LLMs, such as factuality, bias, and toxicity generation.

---

322 *AI Verify Governance Testing Framework and Toolkit*, AI Verify (June 6, 2023), https://aiverifyfoundation.sg/downloads/AI_Verify_Sample_Report.pdf.

323 *Invitation to Pilot AI Governance Testing Framework and Toolkit*, Singapore Infocomm Media Development Auth., 7 (May 25, 2022), https://file.go.gov.sg/aiverify.pdf.

324 Todd Moore et al., *IBM and LFAI move forward on trustworthy and responsible AI*, Open Source @ IBM (June 29, 2020), https://www.ibm.com/opensource/blogs/lfai-move-forward-trustworthy-ai/.

325 *Frequently Asked Questions*, Fairlearn, 6 (Aug. 24, 2023, 7:02AM), https://fairlearn.org/v0.9/faq.html (regarding the relationship between Fairlearn and Microsoft).

326 *Invitation to Pilot AI Governance Testing Framework and Toolkit, supra,* at 6.

327 *First of its kind Generative AI Evaluation Sandbox for Trusted AI by AI Verify Foundation and IMDA*, Infocomm Media Development Auth., (Oct. 31, 2023).

328 *Cataloguing LLM Evaluations*, Infocomm Media Development Auth. and AI Verify Found. (Oct. 2023), https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf.

Specific LLM evaluation and benchmarking methods mentioned in the Evaluation Catalogue include TruthfulQA, a benchmark developed by researchers at OpenAI and the University of Oxford[329] that measures whether a language model is truthful in generating answers to questions; Bias Benchmark for QA (BBQ), a dataset of question sets intended to highlight social biases against people belonging to protected classes in Natural Language Processing model outputs;[330] and Perspective, which uses machine learning models to produce numerical scores assessing levels of toxicity, insults, threats, and sexual explicitness. Perspective was created by Google's Jigsaw unit and Google's Counter Abuse Technology team.[331]

Notably, the Evaluation Catalogue states that LLM evaluation techniques tend to be Western-centric, and should consider user demographics and cultural sensitivities. It also notes the risks of "imprecise or faulty benchmarks" that "could potentially lead to developers being blindsided to critical areas of deficiency."[332] In addition, it states that evaluation recommendations "should not be taken as an endorsement of the reliability and validity of the identified evaluation and testing approaches."[333]

The document also presents a taxonomy of evaluations assessing the capabilities of LLMs, such as evaluations used for Natural Language Understanding, LLM reasoning, evaluations related to LLMs for specific domains such as finance and healthcare, and evaluations related to LLM bias, robustness, and data governance.

As part of its taxonomy, the Evaluation Catalogue references work on extreme risks of LLMs, featured in a May 2023 paper focused on risks of so-called "frontier" AI models;[334] the paper was authored by researchers from private sector corporations building LLMs, including Anthropic, Google's DeepMind, and OpenAI, as well as other organizations focused on researching the existential risks (or "x-risks") of AI, including the Centre for Long-Term Resilience and the Alignment Research Center.

Singapore's Catalogue suggests that a baseline set of evaluations should be used to define a minimal level of LLM safety and trustworthiness for LLMs before deployment. However, it also notes that only advanced, large-scale general purpose machine learning models need to undergo evaluations in its Extreme Risks category; here, the document acknowledges the influence of the Frontier Model Forum, a private sector group comprised of Anthropic, Google, Microsoft, and OpenAI.[335]

Several technology corporations joined the Sandbox, such as Amazon Web Services, Anthropic, Google, IBM, Microsoft, and IBM, as well as AI app developers and third-party AI testers including consulting firms Deloitte and EY.

Asia: Southern Asia: India

NITI Aayog

The Responsible AI #AIFORALL Approach Document for India Part 1 – Principles for Responsible AI

---

329 Stephanie Lin et al., *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, in 1 Proc. of the 60th Annual Meeting of the Ass'n for Computational Linguistics, 3214–3252 (May 8, 2022), https://arxiv.org/abs/2109.07958.

330 Alicia Parrish et al., *BBQ: A hand-built bias benchmark for question answering*, in Findings of the Ass'n for Computational Linguistics: ACL 2022, 2086–2105 (2022).

331 Perspective, https://perspectiveapi.com/.

332 *Cataloguing LLM Evaluations*, *supra*, at 18.

333 *Cataloguing LLM Evaluations*, *supra,* at 27.

334 T. Shevlane et al., *Model evaluation for extreme risks,* ArXiv (Sept. 22, 2023), https://arxiv.org/abs/2305.15324.

335 *Frontier Model Forum*, OpenAI (July 26, 2023), https://openai.com/blog/frontier-model-forum.

Tool Type: Practical Guidance with Self-assessment Questions

India's NITI Aayog is a government body formed in 2015 to provide a platform to discuss intersectoral, interdepartmental, and federal issues to accelerate the implementation of India's national development agenda.[336]

NITI Aayog in 2021 published its two-part *Responsible AI #AIFORALL Approach* papers, which identified principles for responsible design, development, and deployment of artificial intelligence in India. These papers also included enforcement mechanisms for the operationalization of India's Responsible AI principles.

Part 1, published in February 2021, includes a Self-Assessment Guide for AI Usage.[337] The self-assessment guide features a series of consideration-based questions organized according to various stages of the AI development life cycle: Problem Definition and Scoping, Data Collection, Bias in Data Labelling, Model Selection, Training, Evaluation, Deployment and Ongoing Monitoring.

Each consideration set features one or more general approaches to reducing negative impacts of an AI system. For instance, the description of a "mitigation strategy" for identifying a way to address errors in decisions by an AI system states, "If the potential degree of harm for a decision is expected to be high, have appropriate mechanisms in place so stakeholders can contest and humans can get involved in the decision making process."[338]

Some proposed strategies are more specific. A strategy intended to address privacy in relation to data collection states, "Create and document a process to continually scan for and identify new sources of personal and/ or sensitive data."[339]

The self-assessment process also provides guidance for monitoring AI systems after deployment. It suggests tracking system performance and changes over time, and ensuring that mechanisms are in place to allow third-party agencies to review system behavior. [340]

In addition to briefly mentioning LIME and SHAP in relation to AI explainability, [341] NITI Aayog's document also refers to specific methods for model transparency, including Google's Model Card Toolkit, IBM's Fact Sheet for AI governance, and Datasheets for Datasets.[342]

The second portion of the paper, published in August 2021, addresses policy approaches for establishing Responsible AI use in India.[343]

---

336 *Meeting of Governing Council, National Portal of India*, NITI Aayog Constitution (July 28, 2023, 10:49AM), https://www.niti.gov.in/content/niti-governing-council-meetings#:~:text=NITI.

337 *Responsible AI #AIForAll Approach Document for India, Part 1 – Principles for Responsible AI,* NITI Aayog (Feb. 2021), https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf.

338 *Id.* at 45.

339 *Id.* at 46.

340 *Id.* at 49.

341 *Id.* at 35.

342 *Id.* at 53-54.

343 *Responsible AI #AIForAll Approach Document for India, Part 2* – Operationalising Principles for Responsible AI, NITI Aayog (Aug. 2021), https://www.niti.gov.in/sites/default/filehttps://niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf.

India's Proposed Independent Council for Ethics and Technology

In addition to recommending mandatory adherence to the principles for high-risk AI and AI procured by the government,[344] India's Part 2 Operationalizing Principles for Responsible AI paper proposes that an independent Council for Ethics and Technology comprised of lawyers and experts in computer science and AI, as well as civil society, private sector and standards bodies representatives, should be established. [345]

The paper also recommends use of responsible AI tools and techniques and acknowledges that although "the perspectives on the ethics of AI are mostly dominated by western concerns and philosophies," India's socio-economic context "could represent the concerns of 40% of the world."[346]

India: Tamil Nadu State

Tamil Nadu State Policy for Safe and Ethical AI

Tool Type: Practical Guidance with Self-assessment Questions and Scoring Output

India's Tamil Nadu State Policy for Safe and Ethical AI, published by the Tamil Nadu Information Technology Department in 2020,[347] provides a plan for adoption of AI-based systems for Tamil Nadu's policymakers. In particular, the policy recommends use of a scoring system and features a framework for evaluation of AI-based systems addressing transparency, accountability and legal issues, misuse protection, data deficiencies, and fairness and equity.

The policy proposes use of the "DEEP-MAX Scorecard," described as "a transparent point-based rating system for AI Systems." Scoring is based on a set of parameters that form the DEEP-MAX acronym: Diversity, Equity and Fairness, Ethics, Privacy and Data Protection, Misuse Protection, Audit and Transparency, and Cross Geography and Society.

344 *Id.* at 32.

345 *Id.* at 22.

346 *Id.* at 18.

347 *India's Tamil Nadu State Policy for Safe and Ethical AI*, Tamil Nadu Info. Tech. Department (2020), https://it.tn.gov.in/sites/default./files/2021-06/TN_Safe_Ethical_AI_policy_2020.pdf.

Figure 9: ***DEEP MAX Scorecard***

D	Diversity :	Diversity Score - how well the AI System is trained for diversity in race, gender, religion, language, color, features, food habits,accent etc.?
E	Equity & Fairness :	Equity Score - Does the system promotes equity and treats everyone fairly?
E	Ethics	Ethics Score - how well the AI System preservers human values of dignity, fairness, respect, compassion and kindness for a fellow human being
P	Privacy & Data Protection :	Privacy & Data Protection Score - how well the AI System protects privacy of individuals? Does it have data protection features built in?
M	Misuse - Protection :	Misuse Pervention Score - Has the system been designed to incorporate features that inhibit or discourage the possible misuse?
A	Audit & Transparency :	Auditability Score - how good in auditability of decisions made by the autonomous system?
X	Cross Geography & Society :	Cross Geography & Cross Society Score - How well the AI System works across geographies and across societies especially for the disadvantaged societies?

Source: Tamil Nadu Information Technology Department

The "Cross" in Cross Geography is represented by the "X" in DEEP-MAX, and is also referred to in the document as Digital Divide and Data Deficit. This Cross Geography criteria addresses how well a system performs across geographies and societies. This is a concept that matters everywhere: particularly in India, where the world's largest population represents several distinct cultures, languages, and customs.

Details of the process used to determine scores are not included in the policy; however, it does note that testing with "suitably designed test data sets" might be used.[348] The policy also proposes use of a blockchain-based system for storing DEEP-MAX scores for all AI systems used in the public domain, to be designed by Tamil Nadu's e-Governance Agency (TNeGA).[349]

The Tamil Nadu State Policy for Safe and Ethical AI is applicable to Tamil Nadu government authorities, government-controlled organizations and partnerships, and joint venture companies of the government.[350] It states that

348 *Id.* at 30.

349 *Id.* at 29.

350 *Id.* at 21.

implementation of the guidelines would be overseen by a Safe and Ethical AI monitoring committee headed by a chief secretary, with members consisting of Secretary and Senior Officers of select government departments along with AI and policy experts from academic and research institutions. [351]

Asia: Western Asia: United Arab Emirates: Dubai

Digital Dubai

AI System Ethics Self-Assessment Tool

Tool Type: Practical Guidance with Self-assessment Questions and Scoring Output

The AI System Ethics Self-Assessment Tool[352] from the government of Dubai[353] is intended to assist organizations developing or using AI systems to evaluate the "ethics level" of those systems and reduce potential problems that run afoul of the country's AI Ethics Guidelines.[354]

The tool, from Dubai's Digital Authority, Digital Dubai, incorporates a series of questions intended to identify whether the AI system under evaluation is used to make insignificant, significant, or critical decisions.

The self-assessment relies on the assessor to answer "yes" or "no" to statements such as, "Mitigating measures have been pursued to ensure individuals in the same circumstances receive equal treatment."

Depending on the assessor's responses, the tool suggests ways to address potential problems. When the questionnaire is complete, the tool generates a set of scores intended to assess the performance levels of the system under evaluation in relation to fairness, accountability, and transparency.

In its current beta stage, Dubai's AI System Ethics Self-Assessment Tool is for self-assessment purposes only and will not be audited, checked, or regulated.

---

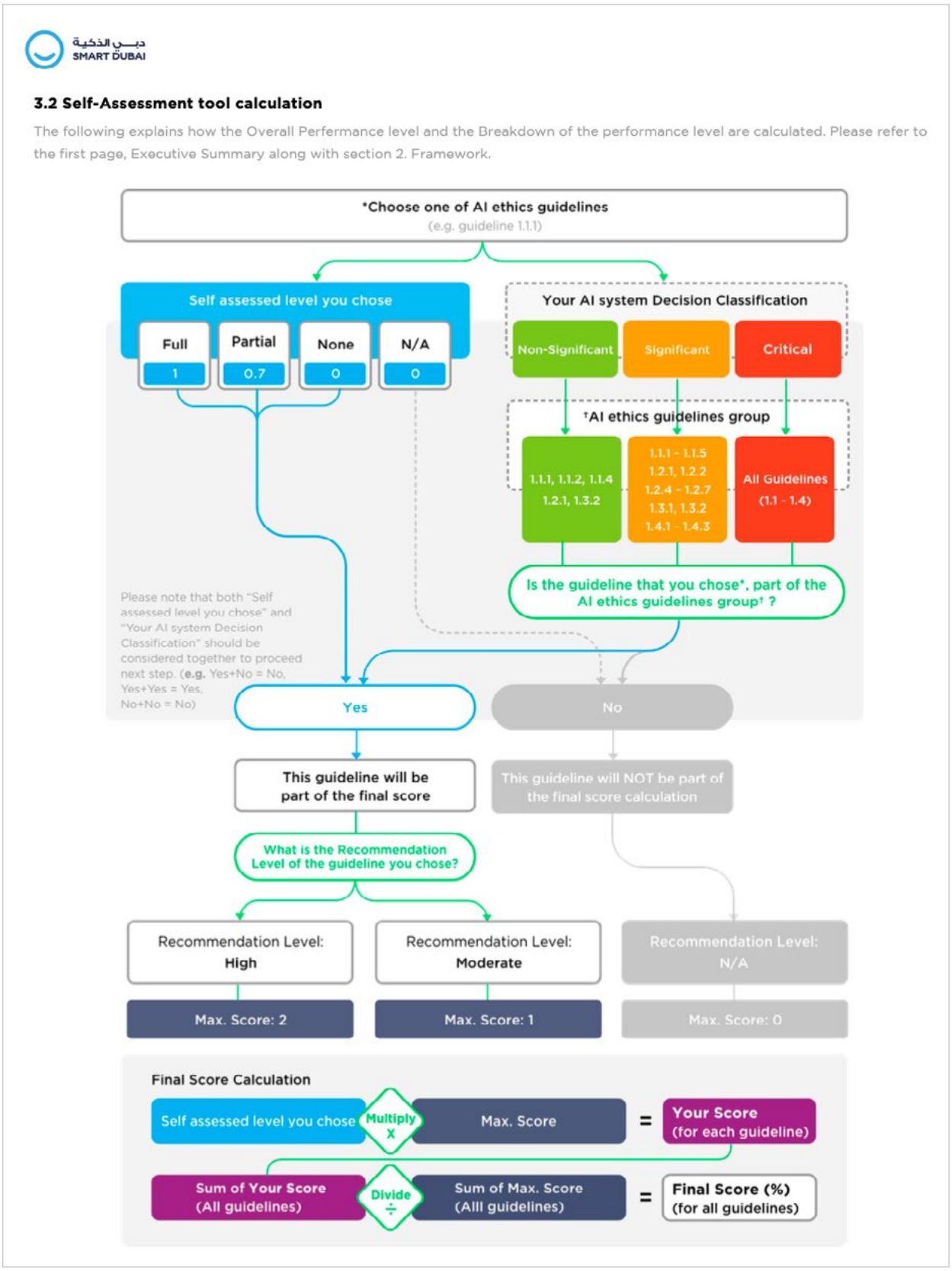
351 *Id.* at 37.

352 *AI System Ethics Self-Assessment Tool,* Digital Dubai (Nov. 14, 2023), https://www.digitaldubai.ae/self-assessment.

353 *The UAE Government*, UAE (May 2023), https://u.ae/en/about-the-uae/the-uae-government; *see also Subnational Governments Around the World: Structure and finance, A first contribution to the Global Observatory on Local Finances*, OECD (2016), https://www.oecd.org/regional/regional-policy/Subnational-Governments-Around-the-World- Part-I.pdf.

354 *Artificial Intelligence Ethics Guidelines,* Digital Dubai (July 27, 2023, 6:12PM), https://www.digitaldubai.ae/self-assessment.

Figure 10: ***AI System Ethics Self-Assessment Tool Report***



Source: Government of Dubai

Europe: Northern Europe: United Kingdom of Great Britain and Northern Ireland

UK Information Commissioner's Office

AI and Data Protection Risk Toolkit

Tool Type: Practical Guidance & Process Framework with Scoring Output

The AI and Data Protection Risk Toolkit from the UK Information Commissioner's Office (ICO) is intended to help AI developers and others assessing AI systems to gauge and reduce the risks of those systems to individual rights and freedoms. It is also intended to complement data protection impact assessments legally required where data processing is likely to result in high risk to individuals.

Originally launched in 2021[355] and updated in March 2023,[356] the AI and Data Protection Risk Toolkit[357] takes the form of a spreadsheet that organizes steps in the AI life cycle: from design, to data acquisition and preparation, to training and testing, followed by deployment and monitoring. Each life cycle category includes risk areas related to issues such as accountability, security, and purpose limitation, all directly referencing specific articles in the UK General Data Protection Regulation.[358]

For each risk area, the process flow detailed in the spreadsheet outlines a control mechanism. Examples of control mechanisms include "Conduct a data protection impact assessment (DPIA)" or "Document clear audit trails of how personal data is moved and stored from one location to another during the training and testing phase." The suggested controls are accompanied by practical steps for implementation.

The toolkit is not mandatory for AI developers or users, and the ICO makes a point of noting that it "is not designed to be 'one size fits all.'" It adds, "There may also be additional risks that apply to your context that are not included in this toolkit."[359]

Europe: Northern Europe: United Kingdom of Great Britain and Northern Ireland

Centre for Data Ethics and Innovation

CDEI portfolio of AI assurance techniques

The UK's Centre for Data Ethics and Innovation (CDEI), part of its Department for Science, Innovation, and Technology, unveiled in June 2023 the CDEI portfolio of AI assurance techniques.[360] Intended to provide guidance to people involved in designing, developing, deploying, or procuring AI-enabled systems, the portfolio features

---

355 Alister Pearson, *New toolkit launched to help organisations using AI to process personal data understand the associated risks and ways of complying with data protection law*, UK Info. Comm'r's Office (July 20, 2021), https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2021/07/new-toolkit-launched-to-help-organisations-using-ai/.

356 *AI and data protection risk toolkit*, UK Info. Comm'r's Office (Aug. 19, 2023, 10:08AM) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/.

357 *Id.*

358 Data Protection Act 2018, UK Public General Acts, 2018, c. 12. https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted.

359 *Id.* at User Guide.

360 *CDEI portfolio of AI assurance techniques*, UK Centre for Data Ethics and Innovation (June 2023), https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques.

descriptions and case studies highlighting various AI assessment methods and products addressing fairness, explainability, robustness, redress, and other AI considerations.

The collection of AI governance tools was developed in conjunction with TechUK,[361]a tech industry trade association. Some products and techniques sold by TechUK member companies are featured in the CDEI portfolio.[362] The portfolio spotlights impact assessments, bias and compliance audits, and performance testing services.[363]

Some techniques featured in the CDEI portfolio mention use of specific quantifiable measures, including SHAP.[364] As noted in the Findings section of this report, use of SHAP in AI measurement has drawn sharp criticism in scholarly literature reviewed here in Part I.

According to the CDEI's description of its portfolio, "the inclusion of a case study in the portfolio does not represent a government endorsement of the technique or the organisation, rather we are aiming to demonstrate the range of possible options that currently exist." The agency plans to publish subsequent editions of its portfolio in the future.

Oceania: Australia and New Zealand: Australia

Australia Commonwealth Ombudsman's Office

Automated Decision-making Better Practice Guide

Tool Type: Practical Guidance with Self-assessment Questions

Australia's Automated Decision-making Better Practice Guide[365] was originally published by the Commonwealth Ombudsman's Office in 2007, long before automation software and systems incorporated machine learning and AI capabilities.

In part due to recognition of AI's impact, the guide was updated in 2019.[366] The updated version considers big data analytics, AI, and machine learning to be features of automated systems.

The guide includes a lengthy checklist of considerations that should be addressed when government agencies implement or update use of an automated system for administrative decision-making.[367] Checklist topics address administrative law, privacy, system design and governance, system maintenance and upgrades, quality assurance assessment, and audit trails for transparency and accountability.

---

361 TechUK, https://www.techuk.org/

362 TechUK members selling products and services also featured in the CDEI portfolio include Deloitte, Holistic AI, Mind Foundry, and Nvidia.

363 *Find out about artificial intelligence (AI) assurance techniques*, UK Centre for Data Ethics and Innovation (Sept. 2023), https://www.gov.uk/ai-assurance-techniques.

364 *Nvidia: Explainable AI for credit risk management: applying accelerated computing to enable explainability at scale for AI-powered credit risk management using Shapley values and SHAP*, UK Centre for Data Ethics and Innovation (June 2023), https://www.gov.uk/ai-assurance-techniques/nvidia-explainable-ai-for-credit-risk-management-applying-accelerated-computing-to-enable-explainability-at-scale-for-ai-powered-credit-risk-management-using-shapley-values-and-shap.

365 *Automated decision-making better practice guide, Appendix A: Better practice checklist,* Commonwealth Ombudsman, (2019), https://www.ombudsman.gov.au/__data/assets/pdf_file/0029/288236/OMB1188-Automated-Decision-Making-Report_Final-A1898885.pdf.

366 *Id.* at 3.

367 *Id.* at 28-34.

Oceania: Australia and New Zealand: New Zealand

The Ministry for Social Development of New Zealand

Model Development Lifecycle and Privacy, Human Rights and Ethics framework (PHRaE)

Tool Type: Practical Guidance with Self-assessment Questions and Scoring Output

New Zealand's Ministry of Social Development created the Model Development Lifecycle (MDL) as a practical guide in March 2022[368] to help manage operational algorithms such as AI-powered services and automated decision systems. Its operational algorithm governance structure, a guide for which was published in 2021,[369] includes three approval stages or "gates" designed to identify and reduce risk throughout the life cycle of an operational algorithm. According to that guide, titled *Governance Guide, Model Development Lifecycle*, in order to progress from a peer review phase to deployment of an algorithmic model, project managers and coordinators must receive approval for sign-off at each gate based on how and whether identified risks have been addressed.

The MDL process for managing operational algorithms features a risk classification matrix intended to determine the risk level of a project. The matrix involves a self-assessment to gauge the severity of impact should a risk occur, and likelihood of the risk occurring. While it does not produce a quantifiable score, the MDL process establishes the risk classification level of a project.

Once risks are identified and classified, controls may be put in place to reduce them, and further external technical, legal, or ethical reviews by a Technical Advisory Group may be recommended.

The MDL process also incorporates the Ministry of Social Development's Privacy, Human Rights and Ethics (PHRaE) framework, a tool for all operational algorithms required throughout all stages of the operational algorithm governance framework. According to the PHRaE framework,[370] projects must engage with the PHRaE process as soon as a proposal to use personal information moves beyond the idea stage. The PHRaE is based on legislative and ethical considerations relating to personal data collection, use, or disclosure.[371]

Results of that process could prompt assignment of a PHRaE Lead to provide guidance for the project throughout the design and development cycle.

The Government of New Zealand partnered with World Economic Forum's Centre for the Fourth Industrial Revolution in 2019 to devise AI governance policy approaches.[372] That partnership project included further assessment of the PHRaE framework, still considered in 2020 to be a pilot project.[373]

---

368 According to a September 2023 email exchange between WPF and New Zealand's Ministry of Social Development.

369 *Governance Guide Model Development Lifecycle*, New Zealand Ministry of Social Development (Oct. 2021), https://www.msd.govt.nz/documents/about-msd-and-our-work/work-programmes/initiatives/phrae/mdl-governance-guide-for-effective-operational-algorithm-decision-making.pdf.

370 *The Privacy, Human Rights and Ethics (PHRaE) Framework*, New Zealand Ministry of Social Development (Nov.r 5, 2023), https://www.data.govt.nz/assets/data-ethics/algorithm/phrae-on-a-page.pdf.

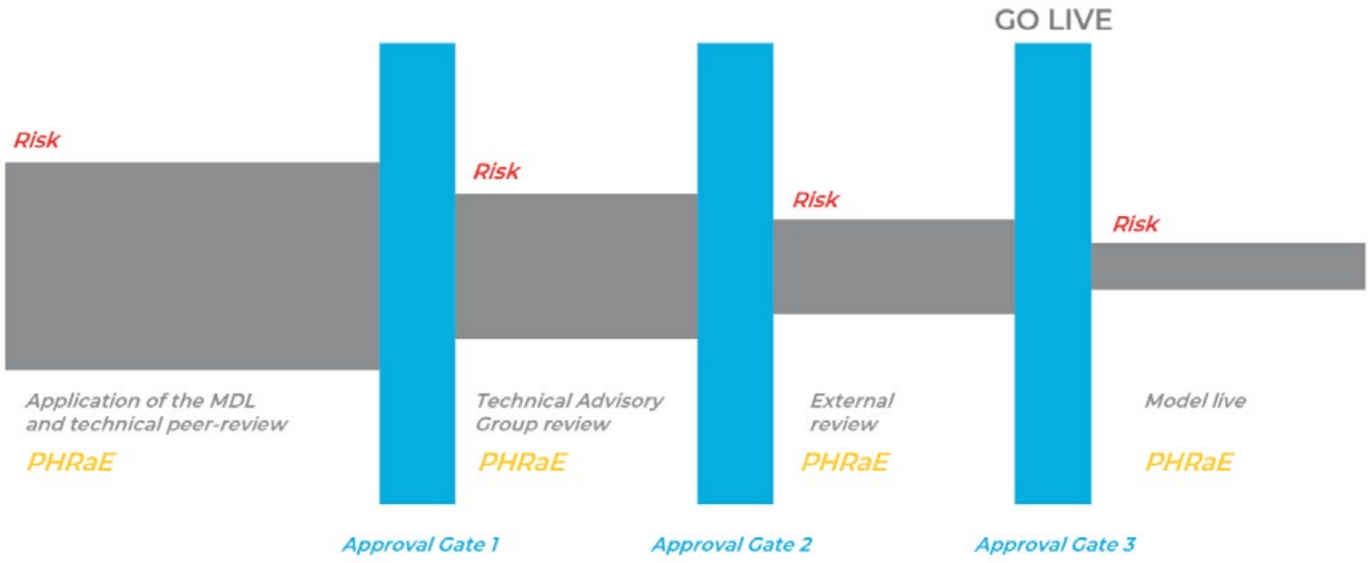
371 *Using personal information responsibly: The Privacy, Human Rights and Ethics Framework*, New Zealand Ministry of Social Development, https://www.msd.govt.nz/about-msd-and-our-work/work-programmes/initiatives/phrae/index.html.

372 Alastair Farr, *Reimagining regulation in the age of artificial intelligence (AI),* New Zealand Gov't (Nov. 11, 2019) https://www.digital.govt.nz/blog/reimagining-regulation-in-the-age-of-artificial-intelligence/.

373 *Reimagining Regulation for the Age of AI: New Zealand Pilot Project,* World Economic Forum (June 29, 2020), https://www.weforum.org/publications/reimagining-regulation-for-the-age-of-ai-new-zealand-pilot-project.

This recent work follows New Zealand's Algorithm Assessment Report of 2018. For the report, New Zealand's chief data steward and the government chief digital officer assessed existing algorithms and their uses across government agencies. The report provides a summary of self-reported information submitted by 14 government agencies about the algorithms that they use to deliver their functions.[374]

Figure 11: ***Model Development Lifecycle Operational Algorithm Governance Structure***



Source: New Zealand Ministry of Social Development

374 *Algorithm assessment report 2018*, New Zealand Government (June 15, 2023), https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-assessment-report/.

Appendix A:

AI Governance Tool Types Lexicon

This report uses the term AI Governance Tools, defined here as:

> AI Governance Tools - Socio-technical Tools for mapping, measuring, or managing AI systems and their risks in a manner that operationalizes or implements trustworthy AI.

This definition encompasses the wide array of formats and methods reviewed here in Part II. This lexicon of AI Governance Tool Types further distinguishes differences among the various types of these formats and methods.

These AI governance tool types were developed on the basis of evidence gathered through the research conducted for this report. These tool types reflect the commonly found components of the AI governance tools identified and named in the Part II review of tools. For more insight into this process, please refer to the *AI Governance Tools and Features Comparison Chart* (See Appendix B). Readers will also find a tool type listed alongside selected relevant tool review entries in Part II of this report.

In general, all of these tool types are designed to improve or measure AI systems, particularly in relation to AI principles including fairness, explainability, and robustness. When applying these types to specific AI governance tools, they may be combined to form hybrid types depending on the components of each tool in question.

AI Governance Tool Types:

- **Practical Guidance** - Includes general educational information, practical guidance, or other consideration factors
- **Self-assessment Questions** - Includes assessment questions or detailed questionnaire
- **Procedural Framework** - Includes process steps or suggested workflow for AI system assessments and/or improvements
- **Technical Framework** - Includes technical methods or detailed technical process guidance or steps
- **Technical Code or Software** - Includes technical methods such as use of specific code or software
- **Scoring or Classification Output** - Includes criteria for determining a classification, or a mechanism for producing a quantifiable score or rating reflecting a particular aspect of an AI system
- **Catalog** - A collection of multiple AI governance tools and types

Appendix B:

AI Governance Tools and Features Comparison Chart

Our AI Governance Tools Comparison chart spotlights key features of select AI Governance Tools from national governments and multilateral organizations. The features indicated here directly map to the tool types we assign to each tool, and reflect our AI Governance Tool Lexicon featured in Appendix A.

		Tool Features					
		Practical guidance	Assessment questions	Process steps	Technical process guidance, code or software	Score or classification output	Mentions specific metrics, code or software
Australia 2019	Automated Decision-making Better Practice Guide TYPE: Practical Guidance with Self-assessment Questions	✓	✓				
Canada 2019	Algorithmic Impact Assessment tool TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓	✓			✓	
Chile 2022	AI Procurement Directorate TYPE: Practical Guidance	✓					✓
Dubai 2019	AI System Ethics Self-Assessment Tool TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓	✓			✓	
Ghana 2023	FACETS Framework TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓	✓			✓	
India 2021	The Responsible AI Approach Document for India Part 1 TYPE: Practical Guidance with Self-assessment Questions	✓	✓				✓
Tamil Nadu, India 2020	Policy for Safe and Ethical AI TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓				✓	
Latin America & Caribbean/IDB 2019	fAIr LAC in a box TYPE: Catalog	✓	✓	✓	✓		✓
Global/OECD 2023	Catalogue of AI Tools and Metrics to Promote Trustworthy AI TYPE: Catalog	✓	✓	✓	✓	✓	✓

New Zealand 2020	Model Development Lifecycle TYPE: Practical Guidance with Self-assessment Questions & Scoring Output	✓	✓	✓		✓	
Singapore 2022	AI Verify TYPE: Practical Guidance with Technical Framework & Software	✓	✓	✓	✓		✓
Singapore 2022	Veritas Initiative TYPE: Practical Guidance & Process Framework with Self-assessment Questions & Technical Code	✓	✓	✓	✓		✓
Singapore 2023	Generative AI Evaluation Catalogue TYPE: Practical Guidance	✓					✓
UK 2021	AI and Data Protection Risk Toolkit TYPE: Practical Guidance & Process Framework with Scoring Output	✓		✓		✓	
US 2021	Artificial Intelligence: An Accountability Framework TYPE: Practical Guidance & Process Framework with Self-Assessment Questions	✓	✓	✓			✓
US 2022	Artificial Intelligence Governance Toolkit TYPE: Practical Guidance & Process Framework with Self-Assessment Questions	✓	✓	✓			
US 2022	Blueprint for an AI Bill of Rights TYPE: Practical Guidance	✓					
US 2023	Artificial Intelligence Risk Management Framework TYPE: Practical Guidance & Process Framework with Self-Assessment Questions	✓		✓			✓

Appendix C:

Some AI Governance Tools Feature Off-label, Unsuitable, or Out-of-context Uses of Measurement Methods

A detailed explanation of Finding 2.

As noted in the Findings section of this report, 7 of 18—or more than 38%—of select AI governance tools reviewed in detail in Part II either mention, recommend, or incorporate at least one of three measurement methods shown in scholarly literature to be off-label, unsuitable, or out-of-context when applied to measure AI systems.

This finding demonstrates the connections between Parts I and II of this report. Analysis gathered through scholarly literature reviewed in Part I informed some criteria used in reviews of AI governance tools surveyed in Part II.

In particular, the criteria used in reviews of AI governance tools drew from the literature presented in the two use cases featured in Part I. These use cases expose and analyze problems associated with off-label use of the US Four-Fifths or 80% Rule to measure disparate impact in AI, and the use of SHAP and LIME in AI explainability measures.

Here, the term "off-label" refers to applications of methods that fall outside of the scope of the original intended applications, such as with "off-label" use of prescription drugs in clinical settings.[375]

Explaining the review sample

More than 30 AI governance tools and AI governance tool-adjacent items were reviewed for Part II of this report. Of those, 18 fully satisfied the definition of AI governance tools introduced in this report:

> AI Governance Tools - Socio-technical tools for mapping, measuring, or managing AI systems and their risks in a manner that operationalizes or implements trustworthy AI.

As featured in Appendix A, the AI Governance Tool Type Lexicon created for this report introduces tool types present in the 18 tools reviewed in depth: Practical Guidance, Self-assessment Questionnaires, Process Frameworks, Technical Frameworks, Technical Code and Software. Also see the AI Governance Tools Types and Features Chart in Appendix B for more detail.

Explaining the Calculations

375 The term "off label use" originally stemmed from the practice in clinical settings of repurposing prescription drugs in a way that differs from what is approved by the FDA and printed on the original prescription label. In the AI context, "off-label" refers to the practice of taking a technology that was created for one context, and using it in another outside of the original use case. NIST mentions "off label use" in its AI Risk Management Framework: "...existing frameworks and guidance are unable to.... consider risks associated with third-party AI technologies, transfer learning, and off-label use where AI systems may be trained for decision-making outside an organization's security controls or trained in one domain and then 'fine-tuned' for another." *NIST AI Risk Management Framework,* National Institute of Standards and Technology, Feb. 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. p. 39. In a study of off-label use of imaging databases, a National Academy of Sciences study found that the practice could lead to bias in AI algorithms. See: Efrat Shimron, Jonathan I. Tamir, Ke Wang, and Michael Lustig, *Implicit data crimes: Machine learning bias arising from misuse of public data*, March 21, 2022. PNAS, https://doi.org/10.1073/pnas.2117203119. And finally, increased risk is also associated with the term as used in its clinical context; see: Rebecca Dresser and Joel Trader, *Off-label prescribing: A call for heightened professional and governmental oversight*, Journal of Law and Medical Ethics, 2009 Fall: 37(3) 476-396. doi: 10.1111/j.1748-720X.2009.00408.x. "The potential for harm is greatest when an off-label use lacks a solid evidentiary basis. A 2006 study examining prescribing practices for 169 commonly prescribed drugs found high rates of off-label use with little or no scientific support."

The review of the select 18 AI governance tools found that 7 — or more than 38% — of those select tools mention, recommend, or incorporate use of at least one of the three problematic measures reviewed in the Part I use cases: SHAP, LIME, and fairness tools that base disparate impact measures on the US Four-Fifths Rule (or mention specifically the US Four-Fifths Rule and its 80% threshold). Seven represents just over 38.8% of 18.

Discovering off-label fairness measures hidden in AI governance tools

Off-label applications in AI governance tools are not always obvious. While it appears that only one of the 18 AI governance tools directly names the Four-Fifths Rule, several of them mention, recommend, or incorporate brand-name AI fairness measures that base disparate impact measures on the US Four-Fifths Rule. In particular, the Part II review found three such brand-name AI fairness measurement methods: Aequitas, AI Fairness 360, and BlackBoxAuditing.

- **Aequitas**

  Aequitas[376] is an AI fairness measure referenced in AI governance tools reviewed in Part II of this report. The open-source bias measure developed by the Center for Data Science and Public Policy at University of Chicago is based on 2018 research[377] that suggests use of the Four-Fifths Rule's 80% benchmark to measure disparity.

- **AI Fairness 360**

  AI Fairness 360, or AIF360, was originally developed by IBM and donated in 2020 to the Linux Foundation AI Foundation.[378] AIF360 supports several "bias mitigation algorithms," including a Prejudice Remover Regularizer and Adversarial Debiasing algorithm. At issue here, the documentation for its Disparate Impact Remover algorithm[379] [380] specifically cites 2015 research introducing a disparate impact measurement based on the Four-Fifths Rule's 80% benchmark.[381]

- **BlackBoxAuditing**

  BlackBoxAuditing is another AI fairness measure referenced in one of the 18 AI governance tools. The BlackBoxAuditing repository[382] bases its disparate impact "repair process" on the aforementioned 2015 research that incorporates the Four-Fifths Rule's 80% benchmark.[383]

---

376 *Aequitas*, Univ. of Chicago Center for Data Science and Public Policy and Carnegie Mellon Univ. Data Science and Public Policy (2018), http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/.

377 P. Saleiro et al., *Aequitas: A Bias and Fairness Audit Toolkit*, Arxiv (2018), https://arxiv.org/abs/1811.05577.

378 Todd Moore et al. *IBM and LFAI move forward on trustworthy and responsible AI*, Open Source @ IBM (June 29, 2020), https://www.ibm.com/opensource/blogs/lfai-move-forward-trustworthy-ai/.

379 AIF360, GitHub, Trusted AI, Supported Bias Mitigation Algorithms, "Disparate Impact Remover." (November 11, 2023), https://github.com/Trusted-AI/AIF360/tree/master.

380 *This notebook demonstrates the ability of the DisparateImpactRemover*, GitHub, https://github.com/Trusted-AI/AIF360/blob/2572fcbfd0267c80cacdec7babf0c32c9c75ba5f/examples/demo_disparate_impact_remover.ipynb.

381 M. Feldman et al., *Certifying and removing disparate impact*, ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (2015), https://arxiv.org/abs/1412.3756 ("While the Supreme Court has resisted a 'rigid mathematical formula' defining disparate impact, we will adopt a generalization of the 80 percent rule advocated by the US Equal Employment Opportunity Commission").

382 *BlackBoxAuditing*, GitHub, https://github.com/algofairness/BlackBoxAuditing.

383 M. Feldman et al.

Seven specific AI governance tools featuring off-label measures found in the Part II survey

A total of 7 of the select 18 tools reviewed in depth for Part II of this report mention, recommend, or incorporate the use of SHAP and/or LIME for AI explainability, use of the US Four-Fifths Rule's 80% threshold, or specifically name or recommend AI Fairness 360, Aequitas or BlackBoxAuditing.

Figure 12: ***AI Governance Tools Including Off-Label Measures***

Government or Organization	Name of AI Governance Tool	Off-Label Measurement Methods Mentioned
Chile (ChileCompra)	AI Procurement Directorate	LIME
Inter-American Development Bank	fAIr LAC in a box/Responsible Use of AI for Public Policy: Data Science Toolkit	Shapley Values
India (NITI Aayog)	The Responsible AI Approach Document for India Part 1	SHAP, LIME
Monetary Authority of Singapore	Veritas Initiative FEAT Fairness Principles Assessment Methodology	Aequitas, AI Fairness 360, SHAP, LIME
Singapore Infocomm Media Development Authority	AI Verify	AI Fairness 360, SHAP
US Government Accountability Office	Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities in 2021	AI Fairness 360
Organization of Economic Cooperation and Development	Catalogue of Tools and Metrics for Trustworthy AI	15 entries in the catalog include off-label measures including Aequitas, AI Fairness 360, BlackBoxAuditing, and several entries using SHAP and/or LIME, including InterpretML

Source: World Privacy Forum, Research: Kate Kaye, Pam Dixon.

1. Chile's Standard Bidding Terms for Data Science and AI Projects mentions that LIME could be used as an explanation tool.[384]

2. IDB FairLAC's Responsible Use of AI for Public Policy Data Science Handbook mentions use of Shapley Values as a quantitative explainability method for deep neural networks, and includes them in a detailed workbook section.[385]

384 *Direccion de Compras y Contratación Publica Aprueba Formato Tipo de Bases Administrativas Para la Adquisición de Proyectos de Ciencia de Datos e Inteligencia Artificial, Resolución N°60, supra*, at 57.

385 *Responsible use of AI for public policy data science handbook, supra.*

3. India's Responsible AI #AIFORALL Approach Document for India Part 1 – Principles for Responsible AI report mentions LIME and SHAP in relation to AI explainability.[386]

4. Monetary Authority of Singapore's FEAT Fairness Principles Assessment Methodology, part of its Veritas Initiative, mentions AI Fairness 360 and Aequitas as commonly-used open-source AI fairness tools, and references use of the "four-fifths rule" to measure disparate impact.[387] A related, earlier Veritas document states that LIME or SHAP could be used for AI explainability, but cautions that "these and other explainability techniques may themselves introduce bias ."[388]

5. Singapore's AI Verify software features a technical plugin that uses SHAP to explain how AI system features affect overall predictions.[389] The 2022 AI Verify project pilot document stated its inclusion of SHAP and AI Fairness 360 in its tool package.[390]

6. The US Government Accountability Office's report, Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities in 2021, mentions AI Fairness 360 as an example of guidance that might be considered when defining responsible AI goals and objectives.[391]

7. The OECD's Catalogue of Tools and Metrics for Trustworthy AI is counted as a single tool in the Finding tally of 7 select tools that mention the use of off-label AI measures; however, this tool catalog itself includes 15 entries featuring off-label measures. These include the following AI measurement methods::

    - Aequitas[392]
    - AI Fairness 360[393]
    - BlackBoxAuditing[394]
    - InterpretML[395]
    - Shapley Additive Explanation (SHAP)[396]

386 *Responsible AI #AIForAll Approach Document for India Part 1 – Principles for Responsible AI*, *supra.*

387 *Veritas Document 3A FEAT Fairness Principles Assessment Methodology, supra.*

388 *Veritas Document 1 FEAT Fairness Principles Assessment Methodology, supra.*

389 *SHAP Toolbox*, GitHub, https://github.com/IMDA-BTG/aiverify/tree/main/stock-plugins/aiverify.stock.shap-toolbox.

390 *Invitation to Pilot AI Governance Testing Framework and Toolkit supra.*

391 *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, *supra.*

392 *Aequitas: Bias and Fairness Audit Toolkit*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI (Feb 23, 2022), https://oecd.ai/en/catalogue/tools/ aequitas:bias-and-fairness-audit-toolkit.

393 *IBM AI Fairness 360*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI (Feb. 22, 2022), https://oecd.ai/en/catalogue/tools/ibm-ai-fairness-360. *See Also: AI Fairness 360 (AIF360)*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI (Sept. 9, 2022), https://oecd.ai/en/catalogue/tools/aif360.

394 *Black Box Auditing and Certifying and Removing Disparate Impact*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI (May 23, 2023), https://oecd.ai/en/catalogue/ tools/black-box-auditing-and-certifying-and-removing-disparate-impact.

395 *Microsoft InterpretML*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI (Mar. 1, 2022), https://oecd.ai/en/catalogue/tools/microsoft-interpretml. *See also: InterpretML*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI (Sep. 9, 2022), https://oecd.ai/en/catalogue/tools/interpret. *See also: Shapley Additive Explanations*, InterpretML, Blackbox Explainers, https://interpret.ml/docs/shap.html. *See also: Local Interpretable Model-agnostic Explanations*, InterpretML, Blackbox Explainers, https://interpret.ml/docs/lime.html. See also: InterpretML, Types of Models Supported, Black- Box, (Nov. 14, 2023), https://interpret.ml/.

396 *Shapley Additive Explanation (SHAP)*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI Policy Observatory, OECD, (Oct. 16, 2023) https://oecd.ai/en/catalogue/ metrics/shapley-additive-explanation-(shap).

- Shapley Variable Importance Cloud (ShapleyVIC)[397]
- Beta Shapley[398]
- Data Shapley[399]
- SHAP[400]
- Shapley Explanation Networks[401]
- Fastshap[402]
- Shapley[403]
- Shapley Additive Global Importance (SAGE)[404]
- Local Interpretable Model-agnostic Explanation (LIME)[405]
- Lime[406]

397 *Shapley Variable Importance Cloud (ShapleyVIC)*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI (Oct. 16, 2023) https://oecd.ai/en/catalogue/metrics/shapley- variable-importance-cloud-(shapleyvic).

398 *Beta Shapley*, Catalogue of Tools and Metrics for Trustworthy AI, OECD.AI (Oct.16, 2023), https://oecd.ai/en/catalogue/metrics/beta-shapley.

399 *Data Shapley*, *Catalogue of Tools and Metrics for Trustworthy AI*, OECD.AI (Oct.16, 2023,) https://oecd.ai/en/catalogue/metrics/data-shapley.

400 *SHAP, Catalogue of Tools and Metrics for Trustworthy AI*, OECD.AI (June 8, 2023), https://oecd.ai/en/catalogue/tools/shap.

401 *Shapley Explanation Networks, Catalogue of Tools and Metrics for Trustworthy AI,* OECD.AI (Sept. 20, 2022), https://oecd.ai/en/catalogue/tools/shapleyexplanationnetworks.

402 *Fastshap*, *Catalogue of Tools and Metrics for Trustworthy AI*, OECD.AI (Sept. 20, 2022), https://oecd.ai/en/catalogue/tools/fastshap.

403 *Shapley, Catalogue of Tools and Metrics for Trustworthy AI*, OECD.AI (Sept. 9, 2022), https://oecd.ai/en/catalogue/tools/shapley.

404 *Shapley Additive Global Importance (SAGE), Catalogue of Tools and Metrics for Trustworthy AI,* OECD.AI (Sept. 9, 2022), https://oecd.ai/en/catalogue/tools/sage.

405 *Local Interpretable Model-agnostic Explanation (LIME), Catalogue of Tools and Metrics for Trustworthy AI*, OECD.AI, https://oecd.ai/en/catalogue/metrics/local-interpretable-model-agnostic-explanation-(lime).

406 *Lime, Catalogue of Tools and Metrics for Trustworthy AI,* OECD.AI (June 8, 2023), https://oecd.ai/en/catalogue/tools/artificial-life-simulator.

Appendix D: OECD Catalog of Tools and Metrics Framework

Figure 13. ***Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems*****, Page 15.**

Type	Field	Definition	Options (if applicable)
Tool description	Name	The name of the tool	
	Link	A link to an up-to-date document	
	Description	A brief summary of the tool and its purpose	
Tool origin	Organisation	The organisation that developed the tool	
	Stakeholder group	The stakeholder group from which the initiative originates	Academia; Trade union/worker representative; Private sector; Civil society; Technical community; Public sector; International governmental organisation; Other
	Country	The country or region where the initiative originated	International; OECD countries; List of regions; List of countries; Other
	Date of publication	Date the tool was published in its first version	
	Contact email	Email of the contact person for the tool (not for public use)	
Tool categorisation	Type of Approach	High-level category of the tool	Process-related approach; Technical approach; Educational approach; Other
	Type of Tool	Category of the tool	Toolkits/toolboxes/software tools; Technical documentation; Technical certification; Technical standards; Product development/lifecycle tools; Technical validation tools; Guidelines; Governance frameworks; Risk management tools; Sector-specific codes of conduct; Collective agreements; Certification; Process-related documentation; Process standards; Change management processes; Capacity/awareness building tools; Inclusive design guidance; Educational materials/training programmes; Other
Scope	Technology platform	The technology platform(s) that the tool can be used for	Platform neutral; Platform specific; Multi-platform; Other
	Target stakeholder group	The stakeholder group where the tool is expected to be implemented	Academia; Trade union/worker representative; Private sector; Civil society; Technical community; Public sector; International governmental organisation; Other
	Primary and secondary policy area	The policy area(s) where the tool is expected to be implemented	Agriculture; Competition; Corporate governance; Development; Digital Economy; Economy; Education; Employment; Environment; Finance and insurance; Health; Industry and entrepreneurship; Innovation; Investment; Public governance; Science and technology; Social and welfare issues; Tax; Trade; Transport; All of the above; Not applicable; Other
	Geographical scope	The country or region that the initiative targets	International; OECD countries; List of regions; List of countries
	Target users of the tool	Users who are expected to use the tool to implement a project	AI system business leader; AI system technical developers; IT specialists; Researchers; AI system operators; Executive management; Government agencies; Data scientists; Project managers; HR managers; All employees; Other
	Impacted stakeholders	Groups of people that will be impacted by the implementation of the tool	Employees; Specific policy communities; Consumers; Regulators; Management; Other
	AI system lifecycle stage(s) covered	The stages of the AI system lifecycle that the tool helps to implement	Planning & design; Data collection & processing; Model building & interpretation; Verification & validation; Deployment; Operation & monitoring; All stages
Alignment with international AI Principles	Relevance to international AI Principles	Grade relevance to international AI Principles	Values-based Principles: Socio-economic and environmental impacts; Human-centred values & fairness; Transparency & explainability; Robustness, security, safety; Accountability; Human agency and oversight. Recommendations for policy makers: Investing in research; Data, compute, technologies; Enabling policy environment; Jobs, skills, transitions; International co-operation
Potential for adoption	Maturity of the tool	Project phase the tool is currently in	Project stage; In development; Running code; Implemented in one project; Implemented in multiple projects; Not relevant anymore; Other
	Degree tool is kept up to date	How the tool is kept up to date with evolving standards, requirements, etc.	No update mechanism planned; Periodic review; Always up to date; Other
	Degree of free use of the tool	Legal conditions for using the tool	Subscription fee; One-time license fee; Free-to-use (creative commons); Open source; Other
	Required resources to implement	The extent to which certain resources are needed to implement/use the tool	IT skills; Domain expertise; Data; IT infrastructure; Operational infrastructure; Financing
	Stakeholders involved	Stakeholders who will be involved in the implementation and operation of the tool	IT employees; Operations employees; All employees; Business unions; Trade unions/worker representatives; Clients; Suppliers; Government agencies; Other
Implementation incentives	Expected benefits	Expected benefits from using the tool	Reduction in risk of AI system failure; Reduction in cost of AI system implementation; Faster implementation of an AI system; Increased quality of AI system results; Improved ability of AI system's implementation to scale; Responsible implementation of AI system; Other
	Enforcement mechanisms	Enforcement mechanisms attached with the usage of this tool	Internal mediation (ombudsman); Ethics board; Certification; Enforcement body; Governmental regulation; Log registrars; Reporting frameworks; Collective agreements; N/A; Other

Source: OECD, June 2021, https://www.oecd-ilibrary.org/docserver/008232ec-en.pdf