Building Trustworthy Artificial Intelligence

Frameworks, Applications, and Self-Assessment for Readiness

September 2025

WORLD BANK GROUP

1818 H Street NW
Washington DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

This work is a product of the staff of The World Bank. The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank, its Board of Executive Directors, or the governments they represent.

The World Bank does not guarantee the accuracy, completeness, or currency of the data included in this work and does not assume responsibility for any errors, omissions, or discrepancies in the information, or liability with respect to the use of or failure to use the information, methods, processes, or conclusions set forth.

Nothing herein shall constitute or be construed or considered to be a limitation upon or waiver of the privileges and immunities of The World Bank, all of which are specifically reserved.

Rights and Permissions

This note’s findings, interpretations and implications are entirely those of the authors and do not necessarily represent the views of the World Bank Group. The authors may be contacted at jpark23@worldbankgroup.org and ahan@worldbank.org.

Cover illustration: Generated by Seelab, August 20, 2025, from the prompt “Hexagonal AI chip in middle, circuits extending to small hexagon bubbles containing icons; like a brainstorm flowchart; 2D.” Subsequent modifications made using Seelab and traditional design tools.

ACKNOWLEDGMENTS

This note is a product of East Asia and Pacific Region, Technology & Innovation Office, Information and Technology Solutions (ITS) Vice Presidency of the World Bank Group (WBG). This work is a collaborative effort of the World Bank Group with the Korea Institute of Science and Technology (KIST). The drafting team authoring the report comprised Jinhee Park (IT Officer, Business Solutions II, ITSTI) and Ahram Han (Consultant, ITSTI) at WBG, Professor Chansoo Kim at the University of Science and Technology and KIST, and Professor Haneol Cho at Sejong University.

The authors are grateful for valuable feedback received from Roman Kovalenko (Director, ITSDO), Srividya Sreeram (Practice Manager, ITSSG), Shrimant Tripathy (Practice Manager, ITSRM), Stuart Yikona (Program Manager, EEAF1), Ji Eun Choi (Senior Economist, DEPDU), and Yusaku Kawashima (Senior IT Officer, Business Solutions I, ITSTI).

The authors also appreciate the guidance and support received from Jason Allford, World Bank Group Country Manager to the Republic of Korea; Stela Mocan, Practice Manager, Digital Government, Digital Vice-Presidency; and Michelle Leonore Fortune, Senior IT Officer, ITS Technology & Innovation Office at the World Bank Group.

The authors are also thankful for the WBG Korea Office for its overall support and ITSTI colleagues for their thoughtful comments and contributions.

CONTENTS

ACRONYMS

AI	Artificial intelligence
BI	Business intelligence
CNN	Convolutional neural network
EU	European Union
KIST	Korea Institute of Science and Technology
LIME	Local interpretable model-agnostic explanations
ML	Machine learning
MPC	Multiparty computation
NIST	National Institute of Standards and Technology
NLP	Natural language processing
OECD	Organisation for Economic Co-operation and Development
PDPs	Partial dependence plots
PETs	Privacy-enhancing technologies
RL	Reinforcement learning
RNN	Recurrent neural network
SHAP	SHapley Additive exPlanations
TEE	Trusted executive environment
UN	United Nations
XAI	eXplainable AI

1. INTRODUCTION

The transformative potential of artificial intelligence (AI) in public governance is increasingly recognized across both developed and developing economies. Governments are exploring and adopting AI technologies to enhance service delivery, streamline administrative efficiency, and strengthen data-driven decision-making. However, the integration of AI into public systems also introduces ethical, technical, and institutional challenges—ranging from algorithmic bias and lack of transparency to data privacy concerns and regulatory fragmentation. These challenges are especially salient in public sector contexts, where trust, accountability, and equity are crucial.

This paper addresses a central question: How can public institutions adopt AI responsibly while safeguarding privacy, promoting fairness, and ensuring accountability? In particular, it focuses on the readiness of government agencies to implement AI technologies in a trustworthy and responsible manner.

The primary audience for this paper includes stakeholders engaged in the design, oversight, and implementation of AI in the public sector. This encompasses public sector leaders, technical policy makers, digital governance professionals, and development partners engaged in institutional capacity building and digital transformation. While global frameworks on trustworthy AI continue to evolve and ethical principles are widely discussed, many government teams still face challenges in translating these high-level concepts into operational decisions, particularly within resource-constrained or institutionally fragmented environments.

This paper responds to that gap by providing both conceptual grounding and practical tools to support implementation. First, it synthesizes key ethical considerations and international frameworks that underpin trustworthy AI governance. Second, it introduces relevant technical solutions, including explainability models, privacy-enhancing technologies, and algorithmic fairness approaches, that can mitigate emerging risks in AI deployment. Third, it presents a self-assessment toolkit for public institutions: a decision flowchart for AI application and a data privacy readiness checklist. These tools are designed to help public sector actors evaluate their preparedness, identify institutional gaps, and inform internal coordination processes prior to AI adoption.

By bridging theory and practice, this paper contributes to ongoing global efforts to build trustworthy AI that is lawful, ethical, inclusive, and institutionally grounded.

2. AI IN PUBLIC GOVERNANCE

AI holds significant promises for improving public governance by enhancing policy processes, delivering more personalized services, and promoting innovation. However, the adoption of AI by governments also introduces important ethical concerns around privacy, fairness, and accountability. This section explores how AI can be leveraged in the public sector, outlines key ethical considerations, and introduces foundational concepts necessary for building trustworthy AI systems.

2.1. AI'S POTENTIAL TO DRIVE INNOVATION AND EFFICIENCY IN PUBLIC SECTOR SERVICES

AI involves developing computer systems capable of performing tasks that typically require human intelligence, such as learning, problem solving, and decision-making. When discussing optimal policy formulation in the context of AI, it refers to leveraging these capabilities to significantly improve processes by efficiently and accurately analyzing vast amounts of data. This encompasses the following aspects:

1. Enhancing the efficiency of processes: AI can automate tasks, analyze data quickly, and identify patterns that might be missed by humans, leading to faster and more informed decision-making.
2. Providing better services: AI can personalize services, predict future needs, and improve the accessibility and quality of services.

By enhancing these core capabilities, AI can facilitate efficient and effective processes across various domains. The application fields of AI technologies are diverse and span across implementations such as data analysis, pattern recognition, and AI-based simulation. This section will examine specific cases and present a blueprint of how AI technologies are applied in practice, while also exploring the potential legal implications and challenges associated with their use.

2.1.1. Efficiency of policy-making process

AI can revolutionize the policy-making process by analyzing vast amounts of data (Deloitte 2017). This allows governments to more efficiently seek solutions to complex social problems and optimize resource allocation (Margetts and Dorobantu 2019). AI can process data much faster and more accurately than traditional human-centered analytical methods. This enables policy makers to swiftly grasp rapidly changing social and economic indicators and respond in a timely manner. AI can also minimize decision-making risks by conducting complex simulations and predictive modeling to analyze potential outcomes of various policy options in advance. For instance, utilizing natural language processing (NLP) to analyze public opinion or media reports can help understand social sentiments and reflect them in policy formulation. While AI can mitigate certain forms of human bias, such as inconsistent judgment, it also introduces new risks related to algorithmic bias and ethical use of data, which must be carefully managed.

2.1.2. Better public services

AI can be leveraged to grasp and predict citizens' needs in real-time by providing customized public services (Mehr 2017). Examples include chatbot-based civil complaint responses or improving traffic flow controls through predictive analytics, because these contribute to enhancing the quality of life for citizens (OECD 2019a). AI plays a pivotal role in increasing personalization and accessibility in public services. By analyzing individual citizens' service usage patterns through machine learning (ML) algorithms, a series of personalized information provisions becomes possible. Virtual assistant services utilizing voice recognition and NLP technologies can improve service accessibility for digitally marginalized groups such as people with disabilities or the elderly. In the transportation sector, real-time traffic data and AI analyses can predict congestion and suggest alternatives, enhancing mobility efficiency. In the health care sector, AI-based disease prediction and prevention services are feasible, boosting the effectiveness of public health policies. The expansion of such AI-based services can elevate satisfaction and strengthen trust. However, issues related to personal data protection and security still require careful examination.

2.2. ETHICAL CONSIDERATIONS IN AI AND GOVERNMENT

As AI becomes increasingly embedded in public sector operations, governments face a growing imperative to address the ethical and governance challenges that accompany its use. While AI offers transformative potential—enhancing service delivery, policy responsiveness, and operational efficiency—it also introduces complex risks (Floridi and Taddeo 2016). These risks include concerns about data privacy, algorithmic bias, lack of transparency, and accountability gaps. Misuse or malfunction of AI systems can lead to serious social, economic, and legal problems, necessitating governments to establish regulations and policies to manage and prevent these risks (Jobin, Ienca, and Vayena 2019). An AI framework should aim to promote the safe and ethical use of AI while maintaining a balance between technological innovation and social values (G20 2019). By establishing clear rules and standards, governments can safeguard citizens' rights and interests and foster a trustworthy AI ecosystem (Fjeld et al. 2020).

2.3. KEY CONCEPTS AND TERMINOLOGIES

In the context of AI, a shared understanding of core concepts and terminologies is essential to ensure responsible development, utilization, and governance. The definitions below provide a comprehensive overview of the most important concepts relevant to trustworthy AI, emphasizing the need for fairness, transparency, accountability, privacy, and collaboration efforts across stakeholders in both public and private sectors.

2.3.1. Main concepts of trustworthy AI governance

1. **Trustworthy AI** refers to AI systems that meet core requirements such as fairness, transparency, safety, and accountability to ensure that AI adheres to ethical and legal standards to gain public trust (Coeckelbergh 2020; Fjeld et al. 2020; G20 2019). It also involves having robust systems that can withstand cybersecurity threats such as model poisoning and prompt injections.
2. **AI governance** involves the overarching system or process that supervises, regulates, and coordinates AI research, development, and usage. It encompasses policy, ethics, and technology standards (G20 2019; Geyer, Klein, and Nabi 2017) and is the key to ensuring alignment of AI innovation with societal goals and values.
3. **Distributed responsibility** means a concept where responsibility for AI outcomes is shared across different stakeholders rather than assigned to one entity (Government of Japan 2016), to address the challenge of pinpointing accountability in complex AI systems. It is critical for transparent and fair public sector AI deployment.
4. **Ethical standards** are guidelines that uphold values like human rights, fairness, and safety throughout AI development and deployment (Coeckelbergh 2020; G20 2019; World Economic Forum 2020) that require collective understanding and compliance by government officials, developers, and citizens.
5. **Accountability** involves identifying clear lines of responsibility when AI-related errors or harm occur. It is essential in complex AI life cycles where various stakeholders such as developers, operators, and users have different roles (Floridi and Taddeo 2016; Strategic Council for AI Technology 2017). It is also closely associated with “distributed responsibility” in multiparty AI systems (Government of Japan 2016).
6. **Explainability** means the capacity of an AI system to provide understandable reasons for its decisions or predictions and seek to address the “black box” problem, particularly in deep learning. eXplainable AI (XAI) facilitates transparent policy decisions, as AI’s reasoning processes become reviewable by stakeholders (Gunning 2017; Jobin, Ienca, and Vayena 2019; White House 2019).
7. **Bias** means the systematic skew arising from data or algorithmic design, leading to unfair or discriminatory outcomes, which may manifest in many forms (e.g., data bias, algorithmic bias) and exacerbate social inequalities (European Data Protection Board 2018; O’Neil 2016). Minimizing bias is critical in public policy to ensure equitable services and decisions.
8. **Autonomy** applies to both human and AI; while human autonomy means the capacity for moral reasoning and free will in decision-making, AI autonomy refers to the ability to perform tasks independently based on data and algorithms (Information Commissioner’s Office 2020; Sweeney 2013). A balance between AI’s self-guided operations and human oversight in public governance is crucial.

Box 1. Autonomy in humans vs. AI

Maintaining Human Autonomy

Philosophically, autonomy applies differently to humans and to AI. Human autonomy entails true free will and moral reasoning, which form the foundation for moral agency and responsibility (Montag, Nakov, and Ali 2024). AI, in contrast, lacks consciousness and moral intent; thus, it cannot possess true moral autonomy or be considered an independent moral agent. Instead, AI systems should be designed explicitly to augment rather than replace human judgment and decision-making. Recent studies highlight the importance of system design in either enhancing or undermining human autonomy: the modality of AI interaction and cultural context significantly influence whether users feel in control or coerced when engaging with AI systems (Montag, Nakov, and Ali 2024). For example, providing users with genuine alternatives—such as the option to interact with a human representative rather than being limited solely to AI—can significantly increase trust, acceptance, and perceived autonomy. An interactive and explainable AI respecting local norms further reinforces user autonomy, whereas opaque or authoritarian AI can diminish it. Thus, preserving human autonomy emerges not merely as a moral imperative but also as a pragmatic strategy for ensuring trustworthy AI. Maintaining a "human-in-the-loop" approach, which allows users to understand, contest, or override AI outputs, ensures moral responsibility remains firmly with humans, aligning AI system designs with Kantian autonomy principles and contemporary notions of distributed responsibility (Montag, Nakov, and Ali 2024).

Defining AI Autonomy and Accountability

In contrast, AI autonomy refers strictly to the capability to independently perform tasks according to preprogrammed algorithms and data inputs, without moral judgment or conscious experience (Bryson 2020; Floridi 2014). Information philosophy terms this "functional autonomy," emphasizing the complexity and adaptability of AI information-processing systems (Bostrom 2014). AI can react and adapt to environmental inputs, but these responses do not internalize human ethical judgments or intentions (Dreyfus 1992). Consequently, AI autonomy is fundamentally technical rather than moral.

As the operational autonomy of AI systems expands, accountability becomes a crucial issue since traditional responsibility frameworks assign moral accountability exclusively to human actors who possess intent and moral judgment (Matthias 2004). AI systems typically involve many stakeholders—such as data scientists, engineers, companies, and users—which complicates accountability. To address the complexity of assigning responsibility for AI system outcomes, new models such as distributed or shared responsibility have emerged (Righetti, Madhavan, and Chatila 2019). Distributed responsibility means accountability is collectively shared among developers, users, operators, and all involved stakeholders according to their roles and degree of influence in system design and implementation (European Parliament 2017). This approach recognizes the inherently collaborative nature of AI development and operation, ensuring clarity and fairness in addressing ethical concerns, safety risks, and unintended consequences.

2.3.2. Privacy protection techniques

1. **Differential privacy** refers to a technique that applies statistical noise to data, preventing the identification of individual data points, enabling data analysis or model training while safeguarding personal information (Floridi and Cowls 2019). It is highly relevant for maintaining a balance between privacy protection and AI utility.
2. **Federated learning** refers to a decentralized ML approach where local data remains on the originating device or organization, and only model parameters are shared. It may reduce risks of privacy breach and support collaborative AI development (European Commission 2020; UNESCO and COMEST 2019) and is often being adopted in fields handling sensitive data such as health care and finance.
3. **Homomorphic encryption** is an encryption scheme enabling computation (e.g., addition, multiplication) on encrypted data without decrypting it first, allowing secure AI model training or inference while preserving the confidentiality of the data (European Commission 2020; Floridi and Cowls 2019). It is useful for sensitive domains in public administration and medical data processing.
4. The **data minimization principle** recommends collecting and using only the data essential for AI operations, avoiding unnecessary personal data collection (Doshi-Velez and Kim 2017). It is particularly pertinent to public sector data usage for policy while preserving privacy.

3. BIG DATA-DRIVEN DECISION-MAKING

Data analytics, pattern recognition, and simulation enable enhanced decision-making processes. These enable policy makers to derive insights from large, complex datasets, predict future outcomes, and optimize policy interventions, thereby improving effectiveness, efficiency, and foresight of their decisions.

3.1. DATA ANALYTICS AND PATTERN RECOGNITION: AN INTEGRATED APPROACH

In the realm of AI, data analytics and pattern recognition are deeply intertwined, often working synergistically to extract valuable insights from data and enable intelligent decision-making. Both fields are crucial for how AI systems process and understand vast amounts of information.

Data analytics is the systematic computational process of transforming raw data sets into useful knowledge. In the context of AI, data analytics refers to leveraging AI techniques to analyze large, complex data sets, simplify processes, scale trend identification, and uncover actionable insights (Eggers, Schatsky, and Viechnicki 2017). It encompasses the entire life cycle of collecting, cleaning, transforming, and applying models to raw data to ultimately extract information relevant for decision-making (Eggers, Schatsky, and Viechnicki 2017). This includes the ability to handle structured, semistructured, and unstructured data (Margetts and Dorobantu 2019). Data analytics can be broadly categorized based on its objectives and approach (Deloitte 2017):

- **Descriptive analytics:** Answers “What happened?”—focusing on summarizing past data to understand basic characteristics. AI enhances this by rapidly processing vast amounts of structured and unstructured data to identify patterns, trends, and correlations. For example, a retailer might deploy AI algorithms to analyze customer data, uncovering insights into purchasing trends and preferences.
- **Diagnostic analytics:** Investigates “Why did it happen?”— identifying root causes and correlations within complex datasets. AI accelerates this, for instance, in health care by analyzing patient data, medical histories, and lab results to pinpoint underlying causes of diseases more accurately and faster than traditional methods.

- **Predictive analytics:** Forecasts "What might happen next?"—using historical data, statistical modeling, and AI/ML to project future trends and outcomes. Examples include predicting financial market trends or forecasting maintenance needs for manufacturing equipment.
- **Prescriptive analytics:** Determines "What should we do next?"—providing actionable recommendations to optimize future actions based on insights from previous stages. AI contributes by analyzing scenarios, such as in supply chain management, where it analyzes inventory levels, demand forecasts, and shipping conditions to recommend optimal order quantities and delivery schedules.

Pattern recognition is a fundamental branch of AI and ML concerned with the ability of machines to identify, analyze, and classify patterns (regularities, tendencies, structures, etc.) within data (Russell and Norvig 2016). This involves using algorithms to detect these patterns in various data types, including images, audio, text, and numerical data (Russell and Norvig 2016). The goal is to extract useful information and to understand, classify, and interpret data, enabling tasks like object recognition, anomaly detection, and prediction (Russell and Norvig 2016).

Pattern recognition utilizes diverse methodologies, including statistical methods, ML (e.g., supervised and unsupervised learning), and deep learning (e.g., convolutional neural networks [CNNs] for image patterns, transformers for text sequences) (Russell and Norvig 2016). These techniques are core components of AI used by data scientists to understand, classify, and interpret data across various domains (Russell and Norvig 2016).

3.1.1. The symbiotic relationship: How pattern recognition powers data analytics

Data analytics and pattern recognition are closely linked, with pattern recognition serving as a key engine for modern data analysis. It's not just a related field but a fundamental capability that underpins and enables advanced data analytics (Russell and Norvig 2016). Data scientists leverage AI pattern recognition techniques to understand, classify, and interpret data—core activities in the data analytics life cycle (Russell and Norvig 2016).

AI-driven pattern recognition techniques (ML, deep learning) allow data scientists to efficiently analyze vast, complex data sets, identifying subtle patterns, correlations, and trends that might be difficult or impossible to detect with traditional methods (OECD 2019a). This enables deeper understanding and more accurate data-driven decision-making (OECD 2019a).

The ability to recognize patterns in historical and current data is crucial for building predictive models (forecasting future events) and prescriptive models (recommending actions), which are key outputs of advanced data analytics (Deloitte 2017). For example, predicting equipment maintenance needs or forecasting shifts in consumer demand relies on recognizing relevant data patterns (Russell and Norvig 2016).

Many modern pattern recognition techniques, especially those based on deep learning (e.g., CNNs for images, recurrent neural networks [RNNs]/transformers for text), are specifically designed to extract patterns from unstructured data, significantly broadening the scope of data analytics (Eggers, Schatsky, and Viechnicki 2017). While data analytics provides the context for pattern recognition (cleaned, processed data) and interprets its findings, the insights derived from pattern recognition feed back into the analytical process, refining models and leading to more reliable predictions and solutions.

This relationship is not static but an evolving symbiosis. As data sets grow in volume, velocity, and variety (especially unstructured data), data analytics becomes critically reliant on sophisticated AI-based pattern recognition techniques (such as deep learning). Pattern recognition is transitioning from being a tool for data analysis to an essential engine driving modern, large-scale data analysis. As data complexity increases, so does the sophistication of pattern recognition required for effective data analysis, suggesting its growing, enabling role beyond being just a component. AI's ability to "analyze vast amounts of data quickly and efficiently" and identify patterns "difficult or impossible to detect using other technologies" signifies this shift (Russell and Norvig 2016). Advances in pattern recognition, therefore, directly fuel advances in the scope and depth of data analytics, particularly in complex domains.

As data analytics and pattern recognition increasingly rely on complex "black box" AI models (e.g., deep learning), a shared critical challenge emerges: the need for explainability and interpretability. The value of insights from data analytics and patterns from pattern recognition diminishes if the underlying processes are opaque, hindering trust and actionable decision-making. Complex AI models often lack inherent transparency in their internal decision logic (Adadi and Berrada 2018). "The effectiveness of these systems is limited by the machine's current inability to explain their decisions and actions to human users" (Gunning 2017). This lack of understanding leads to concerns about "reliability, fairness, biases and other ethical issues" (Adadi and Berrada 2018). For data analytics to lead to "informed decision-making" (Eggers, Schatsky, and Viechnicki 2017) and pattern recognition to provide "useful information" (G20 2019), the reasoning behind the outputs must be understood. Thus, the push toward more powerful data analysis and pattern recognition via complex AI necessitates a parallel drive for XAI to maintain trust, ensure fairness, and enable effective human oversight and intervention. This is a common thread linking the advancement of both fields. Table 1 compares the key characteristics of data analytics and pattern recognition in AI.

Table 1. Key characteristics of data analytics and pattern recognition in AI

Feature	Data analytics	Pattern recognition
Primary goal	Extract comprehensive insights and support decision-making	Identify and classify regularities/structures in data
Scope	Broad end-to-end process from data collection to interpretation	More focused on the "identification" and "classification" steps
Key methodologies	Statistical analysis, business intelligence (BI), descriptive/diagnostic/predictive/prescriptive modeling, data mining	ML algorithms (clustering, classification), deep learning (CNNs, RNNs), statistical pattern matching, syntactic pattern recognition
Typical input	Raw or preprocessed data, business questions	Specific data sets for pattern extraction (images, signals, text)
Typical output	Reports, dashboards, predictions, recommendations	Identified patterns, classified data, feature representations
Interrelationship	Data analytics utilizes pattern recognition as a key enabling technique, especially for complex data and predictive/prescriptive tasks. Pattern recognition provides the "recognition" upon which data analytics builds broader understanding and action	Pattern recognition is used to perform specific tasks within the data analytics process, enhancing its accuracy and efficiency

Source: Original to this publication.

3.1.2. Key techniques and applications

Data analytics techniques are amplified by AI and pattern recognition. AI significantly enhances traditional data analysis capabilities, such as automated data imputation, synthetic data generation, and insight explanation (White House 2019). Generative AI transforms data discovery through conversational interfaces (Mehr 2017). Pattern recognition deepens specific data analyses:

- **Image and video analysis:** CNNs facilitate medical image analysis and security applications (Eggers, Schatsky, and Viechnicki 2017).
- **NLP:** Transformers support text classification and sentiment analysis (Fjeld et al. 2020).
- **Speech recognition:** Enables voice identification and virtual assistants (Mehr 2017).
- **Anomaly detection:** Detects unusual patterns for fraud prevention and disease outbreak prediction (Deloitte 2017).
- **Time series analysis:** Predicts trends in financial markets and consumer behavior (Gunning 2017).

Data analytics and pattern recognition continue evolving together, with expanded impacts anticipated alongside AI advancements. They synergistically drive innovation across various fields, and some of the examples are listed below.

- **Health care:** Disease pattern recognition informs treatment and early disease detection (Deloitte 2017; Floridi and Cowls 2019).
- **Finance:** Transaction anomaly detection and market trend prediction (Deloitte 2017; G20 2019).
- **Retail:** Predicting demand, optimizing inventory, and personalized recommendations (Deloitte 2017; Doshi-Velez and Kim 2017).
- **Manufacturing:** Predictive maintenance based on the Internet of Things sensor data patterns (Deloitte 2017).
- **Policy making:** Market predictions and refining policies via NLP and pattern identification (Information Commissioner's Office 2020; Wirtz, Weyerer, and Geyer 2019).

3.2. AI-BASED SIMULATION

Simulations and modeling using AI are extensively used because they allow for the forecasting and evaluation of policy outcomes in advance (OECD n.d.). They play a significant role in enhancing policy effectiveness and minimizing adverse effects. In the modern era, AI-based simulations and models are useful for estimating future states of complex systems. Simulations can predict how traffic flow will change with transportation policy adjustments or assess the impact of environmental regulations on industries. Simulations also enable testing various scenarios to select optimal policy alternatives. Reinforcement learning (RL) algorithms can develop response strategies for diverse situations that may arise after policy implementation. Results from RL applications reduce policy failure risks and minimize negative impacts, which are targets of the computations. We also remember that the accuracy of models heavily depends on input data and assumptions, necessitating continuous verification and updates.

4. ETHICAL CHALLENGES AND RISKS IN AI APPLICATION

While the potential benefits of AI for governments are substantial, it is equally important to address the ethical risks and governance challenges that accompany its use. As AI systems increasingly influence decisions that affect public welfare, concerns arise around issues such as explainability, the illusion of objectivity, and the protection of personal data. This section explores these challenges and risks in AI applications, which call for a comprehensive approach to trustworthy AI use.

4.1. EXPLAINABILITY

Complex AI models like deep learning may be opaque, making it difficult to explain results (Lipton 2018). This can be an obstacle in ensuring transparency and accountability in policy decisions. If the basis of AI used in policy making is not understood, trust in the outcomes may diminish, negatively affecting communication and trust-building with the public. Additionally, when errors occur in predictions or decisions, it becomes challenging to clarify responsibility. To solve these issues, developing and adopting XAI technology is necessary. XAI aims to provide the operational principles and decision-making processes of AI systems in a form understandable by humans. This allows policy makers to review AI's judgment basis and adjust as needed. However, since XAI technology is still in its early stages, technological development alongside institutional support is needed. More details of XAI are discussed in the next section.

4.2. ILLUSION OF OBJECTIVITY IN AI

AI systems are often regarded as tools that support data-driven decision-making and reduce certain types of human error. By systematically processing large and diverse datasets, AI has the potential to help mitigate the impact of cognitive limitations—such as overreliance on recent or emotionally salient information.

Nevertheless, the application of AI does not ensure objectivity. AI systems can exhibit bias in various ways, often related to the methods and environments in which data is collected, as well as the choices made during model development. One emerging concern is the proliferation of "AI slop," referred to as low-quality, misleading, or irrelevant content generated by AI, which may distort information environments and undermine trust. Addressing these risks requires careful attention throughout the entire process of AI model development and

Box 2. Types of bias

A. Computation-based biases

1. Data bias
   - Errors arising from biases inherent in the data that AI models learn from
   - Predictions distorted as a result of overrepresentation or underrepresentation of specific groups or features
2. Algorithmic bias
   - Errors resulting from biases introduced during the design or selection process of algorithms
   - Distortion of specific results or predictions, undermining fairness
3. Sampling bias
   - Errors occurring when the sample in data collection does not represent the entire population
   - Difficulty in generalizing results, reducing prediction accuracy
4. Selection bias
   - Overestimation or underestimation of specific elements during the data or case selection process
   - Skewing of analysis results, leading to incorrect conclusions
5. Representation bias
   - Data sets that do not properly reflect specific groups or characteristics
   - Models making inaccurate predictions for certain groups
6. Omission, self-selection, and imbalanced data bias
   - Important features or variables missing from data, leading to inaccurate model predictions
   - Reduced reliability of results because data are incomplete
   - Reduced representativeness of samples in research or surveys because participants self-select
   - Results skewed toward opinions or characteristics of specific groups
   - Models making biased predictions about specific classes owing to imbalance in data quantities between classes
   - Decreased prediction accuracy for rare classes
7. Overfitting bias
   - Models overly specialized in training data, reducing their ability to generalize to new data
   - Performance decreases in actual application
8. Privacy bias
   - Bias that occurs data sets when data collection is limited to protect personal information
   - Reduced predictive power because models lack some information
9. Biased loss function
   - Design of the loss function that favors specific results, leading the model to learn biasedly
   - Damage to the model’s fairness and accuracy

Box 2. Types of bias (continued)

B. Social biases

10. Sociocultural and socioeconomic bias
    - Social or cultural stereotypes or prejudices reflected in data or algorithms
    - Discrimination or unfair outcomes against specific groups
    - Prejudices or stereotypes against specific genders inherent in AI systems
    - Discriminatory outcomes based on gender.
    - Prejudices against specific races or ethnicities reflected in AI systems
    - Inequalities or discrimination based on socioeconomic status inherent in AI systems
    - Unfair outcomes based on income or education levels.
11. Age and language bias
    - Bias that includes discrimination or preconceptions based on age groups in AI models
    - Inaccurate predictions or discriminatory results based on age
    - Adverse effects for some users from bias toward specific languages or dialects
    - Failure to reflect linguistic diversity, reducing accuracy
12. Geographical bias
    - Data excessively reflecting specific regions or places
    - Failure to consider regional characteristics, lowering the model's generalization ability
13. Stereotype, herd, and imbalanced data bias
    - Socially formed stereotypes reflected in AI systems
    - Reinforcement of discriminatory results or predictions
    - Tendency to follow the opinions or behaviors of the majority that is reflected in data or decision-making
    - Difficulty making independent judgments, leading to biased conclusions
14. Cognitive bias
    - Bias inherent in the AI team of developers or data scientists
15. Confirmation bias
    - Interpreting or collecting information or data selectively to confirm existing beliefs or hypotheses
    - Difficulty making objective judgments, leading to biased conclusions
16. Automation bias
    - Humans overly trusting automated systems or AI judgments
    - Users that follow incorrect decisions without recognizing system errors or limitations
17. Economic bias
    - AI firms voluntarily manipulating data and algorithms to maximize profits

Source: Original to this publication.

usage, from data selection to algorithm design and system implementation. Box 2 outlines common types of bias that can affect the objectivity and reliability of AI applications.

4.3. DATA PRIVACY

AI extensively utilizes personal data, raising concerns about personal information leakage and misuse (European Union 2016). This is an area requiring legal regulations and ethical considerations. The data collected and processed by AI systems may include sensitive personal information, which could violate laws such as the General Data Protection Regulation. Improper deidentification or anonymization processes can lead to the identification of individuals, while risks like data misuse or hacking persist. These issues threaten individual privacy and can undermine social trust.

Strict management and control for personal information protection are necessary for AI-driven policy decisions. Governments and organizations must establish legal and institutional mechanisms to balance data utilization and personal data protection. Methods such as the data minimization principle, consent-based data collection, and personal data impact assessments should be implemented. Additionally, enhancing public awareness and education on personal information protection is crucial.

4.4. ACCOUNTABILITY

As AI systems are increasingly used in high-stakes public domains, accountability becomes a foundational pillar of trustworthy AI. However, accountability in AI, involving a clear assignment and responsibility trace, is particularly challenging owing to the complexity of AI systems and the multistakeholder environment that includes developers, data providers, platform operators, government agencies, and end users. Because these actors may contribute to system design, training, deployment, and oversight, it is difficult to isolate responsibility when outcomes are problematic.

This complexity gives rise to the concept of distributed responsibility, where multiple stakeholders jointly influence the behavior of AI systems throughout their life cycle (Floridi and Taddeo 2016; Government of Japan 2016; Strategic Council for AI Technology 2017). Without well-defined governance structures, this diffusion of responsibility can lead to accountability gaps, situations where no single actor is held liable for harm, eroding public trust and legal clarity.

To address this, governments must ensure that clear accountability mechanisms are embedded into AI governance frameworks. This includes defining the roles and obligations of each stakeholder, setting standards for documentation and traceability, and establishing oversight bodies capable of auditing and evaluating AI systems. Furthermore, policies should support redress mechanisms, enabling individuals affected by AI decisions to seek explanations, corrections, or legal remedies. Embedding accountability from the outset not only mitigates risk but also fosters transparency and trust in AI-enabled public services.

4.5. COMPREHENSIVE ETHICS

Concerns related to trustworthy AI, including fairness, explainability, privacy, and accountability, can be comprehensively addressed under the common framework of AI ethics. This concept broadly deals with various ethical issues arising in the development and application process of AI technology, and ethical considerations are essential for effectively utilizing AI in policy decisions (Moor 2006).

AI ethics provide fundamental principles for the responsible use of technology. First, fairness must be secured to ensure AI systems do not disadvantage specific groups or individuals. This includes minimizing data biases and preventing discriminatory algorithmic outcomes (Whittlestone et al. 2019). Second, transparency and explainability should be achieved so that the operational principles and decision-making processes of AI are understandable. This is important for building trust with policy makers and the public alike by clearly presenting the basis of AI judgments (Doshi-Velez and Kim 2017). Third, privacy and data protection should be ensured so that personal information is not unjustly collected or misused. This means applying strict security and ethical standards throughout the entire process from data collection to processing, storage, and disposal (Taddeo and Floridi 2018). Fourth, accountability of AI systems should be secured to clarify responsible parties when errors or side effects occur, ensuring appropriate responses and corrections are made (Danks and London 2017).

Implementing responsible AI involves navigating trade-offs between ethical principles. These trade-offs arise because optimizing one aspect can detract from another, requiring careful balance. Common trade-offs include fairness, transparency, privacy and accountability. Table 2 illustrates principle pairs and their trade-offs (Sanderson, Douglas, and Lu 2024).

These ethical principles are also emphasized in the international community. The Organisation for Economic Co-operation and Development (OECD) presented ethical principles including human-centered values, fairness, transparency, and safety in its recommendations on AI (OECD 2019b), and the European Union (EU) introduced strict ethical standards and supervision for high-risk AI systems through the AI Act (European Commission 2021). Therefore, AI users for governance must reflect these international trends and establish legal and institutional devices centered on ethics in domestic AI policies.

Utilizing AI based on ethics is essential for increasing public trust and achieving sustainable technological development. To this end, AI users for governance should establish ethical AI guidelines and build a governance structure involving developers, policy makers, and citizens (Floridi and Cowls 2019). Efforts should also be made to enhance understanding of

Table 2. **Trade-offs between AI principles**

Principle pair	Nature of trade-off	Example scenario
Fairness vs. Transparency	Fairness may require obscuring sensitive attributes to prevent discrimination, while transparency demands openness about model features and logic.	An AI hiring tool hides gender and ethnicity to ensure fairness, but this limits transparency.
Privacy vs. Utility	Strong privacy protections (e.g., anonymization, differential privacy) can reduce the accuracy or usefulness of AI models.	A health prediction model trained on anonymized data performs worse from loss of granularity.
Transparency vs. Accountability	Highly transparent models may expose internal logic that complicates assigning responsibility, while accountability requires clear traceability.	A transparent AI system reveals decision logic, but it's unclear who is responsible for errors.
Fairness vs. Accountability	Ensuring fairness may involve complex pre- or post-processing that obscures who is responsible for final outcomes.	A fairness-enhanced credit scoring model adjusts outputs post hoc, making responsibility unclear.

Source: Adapted from Sanderson, Douglas, and Lu (2024).

ethical issues through education and awareness-raising activities and to establish a culture of trustworthy AI use throughout society (Mittelstadt 2019).

5. TECHNOLOGICAL SOLUTIONS TO ADDRESS AI CHALLENGES

While ethical concerns challenge the trustworthy use of AI, a range of technological solutions has emerged to help mitigate these risks. This section explores how innovations in AI design and technologies can be leveraged to enhance transparency, reduce bias, and strengthen data protection. Although these innovations cannot eliminate all risks, they offer valuable tools to support more trustworthy AI systems.

5.1. TECHNOLOGICAL APPROACHES TO EXPLAINABILITY

XAI refers to a set of techniques and methods aimed at making the decision-making processes of ML models more transparent and understandable to humans. This is essential for enabling users to grasp the reasoning behind a model's output, which is particularly important for fostering trust in AI systems (Adadi and Berrada 2018). The global effort to solve explainability issues is multifaceted, involving research and development of new methodologies and the introduction of interpretability frameworks across various sectors.

5.1.1. Explainability by model design

These approaches focus on designing inherently transparent models or models whose architectures facilitate interpretability.

- **Linear models (linear regression and logistic regression):** These simpler models' decision-making processes can be easily traced and understood. They are often used when transparency is more critical than predictive power.
- **Decision trees:** These models break down decision-making into a series of yes/no questions, facilitating easy tracing of the decision-making process.
- **Rule-based models:** These models use predefined rules, making their decision-making processes explicitly transparent.
- **Transparent neural networks:** Advances in neural architecture aim at building interpretable neural networks by reducing complexity or applying constraints on layers and parameters (Gilpin et al. 2018).
- **Causal inference models:** These models explicitly encode cause-and-effect relationships, making them intuitive and transparent by clearly showing how outputs relate to inputs.

- **Local interpretable model-agnostic explanations (LIME):** LIME generates interpretable approximations of black-box model predictions for specific instances, identifying features critical to the decision-making process through perturbations of input data (Bellamy et al. 2019).
- **SHapley Additive exPlanations (SHAP):** SHAP assigns an importance value to each feature based on its contribution to the model's predictions using cooperative game theory, applicable across various complex models (Lundberg and Lee 2017).
- **Surrogate models:** Simpler, interpretable models like decision trees or linear models that approximate the predictions of more complex models, providing insight into how the complex model makes decisions (Molnar 2020).
- **Counterfactual explanations:** These explanations describe conditions under which specific inputs would lead to different predictions, enabling actionable insights (Wachter, Mittelstadt, and Floridi 2017).
- **Partial dependence plots (PDPs):** Visual tools that show how the model's predictions vary with changes in one or two features while holding other features constant, providing clear interpretability of feature influence (Friedman 2001).

5.1.2. Model-specific explanation methods

These techniques are tailored specifically to interpret certain complex models, particularly neural networks, and are common in computer vision and NLP tasks.

- **Saliency maps:** Common in image classification tasks, saliency maps visually highlight influential pixels or regions in the input image contributing to the model's decision (Simonyan, Vedaldi, and Zisserman 2013).
- **Attention mechanisms:** Used prominently in transformer models within NLP, these mechanisms highlight parts of the input (e.g., specific words or phrases) the model considers important for making its predictions, clarifying the internal reasoning of the model (Vaswani et al. 2017).

5.2. TOOLS AND TECHNIQUES FOR ADDRESSING AI BIAS AND THE ILLUSION OF OBJECTIVITY

Addressing AI bias requires leveraging technological advancements alongside ethical and regulatory frameworks.

5.2.1. Bias detection and fairness audits

- **Bias metrics and evaluation tools:** Metrics such as disparate impact analysis, statistical parity, and equalized odds are employed to assess bias in AI models. These tools help quantify disparities in model outcomes across different demographic groups.

- **AI fairness toolkits:** Open-source tools such as IBM's AI Fairness 360 (Bellamy et al. 2019), Google's What-If tool, and Microsoft's Fairlearn (RBC Borealis 2022) provide automated bias detection and mitigation techniques.
- **Automated fairness audits:** Implementing AI-driven auditing systems that continuously assess model fairness throughout the AI life cycle.

5.2.2. Data preprocessing techniques

- **Resampling and reweighting:** Adjusting datasets by oversampling underrepresented groups or applying reweighting techniques helps balance the training data, leading to more equitable model performance (Hardt, Price, and Srebo 2016).
- **Synthetic data generation:** Utilizing generative models, such as generative adversarial networks, to create synthetic yet representative data can fill gaps in biased data sets, enhancing diversity and fairness.
- **Bias-aware feature engineering:** Identifying and modifying features that contribute to biased decision-making ensures that models focus on relevant attributes, reducing unintended bias.
- **Privacy-preserving data augmentation:** Techniques like differential privacy and federated learning enhance data diversity while safeguarding sensitive information, promoting both fairness and privacy (Hardt, Price, and Srebro 2016).

5.2.3. Algorithmic adjustments for fairness

- **Fairness-constrained optimization:** Incorporating fairness constraints into the objective functions of ML models ensures equitable performance across different demographic groups (Arya et al. 2019).
- **Adversarial debiasing:** Training models with adversarial techniques encourages the learning of unbiased representations, effectively reducing discriminatory patterns.
- **eXplainable AI (XAI) for bias identification:** Employing interpretable ML methods, such as SHAP and LIME, aids in understanding and rectifying biased model behaviors.
- **Adaptive learning models:** Implementing reinforcement learning approaches that dynamically adjust to biases detected during deployment ensures continuous improvement in fairness (Arya et al. 2019).

5.2.4. Post hoc bias mitigation

- **Output calibration and adjustment:** Applying corrective measures after model training, such as equalizing error rates across demographic groups, helps rectify biased outcomes (Hardt, Price, and Srebro 2016).
- **Bias-correcting neural networks:** Developing models capable of self-monitoring and adjusting their decision boundaries promotes sustained fairness in predictions (Bellamy et al. 2019).

- **Human-in-the-loop systems:** Incorporating human oversight into AI decision-making processes allows for the identification and correction of biases that automated systems might overlook.

5.3. PRIVACY-ENHANCING TECHNOLOGIES (PETs)

Technologies such as data deidentification, anonymization, encryption, and differential privacy are fundamental tools for safeguarding personal data within AI-based systems. They allow the use of personal data while minimizing risks of privacy infringement (El Emam and Arbuckle 2013).

- **Data deidentification:** Protects individuals' privacy by removing or transforming information that can identify them (European Union 2016; Stallings 2017). It still maintains the necessary information for data analysis and ML while reducing privacy infringement risks. Anonymization involves completely removing personal identifiers, making individuals unidentifiable. Therefore, it provides only the level of protection required by personal data protection laws (International Organization for Standardization 2011).
- **Basic encryption:** Transforms raw data into an unreadable format, ensuring that if information is intercepted during transmission or storage, it remains protected (Tavani 2007).
- **Homomorphic encryption:** Enables computations to be performed on encrypted data without decrypting it, thus maintaining confidentiality while facilitating AI model training and inference (Gentry 2009). Data owners can perform encrypted operations locally while the server orchestrates the training process. Decryption keys remain strictly with data owners, reducing exposure risk.
- **Secure multiparty computation (MPC):** Allows multiple participants to jointly compute a function over their inputs without revealing those inputs to one another (Solove 2006; Tavani 2007). Key applications include
  - Cross-institutional collaboration: Multiple hospitals can collaboratively train disease detection models
  - Banking and finance: Joint fraud analysis without disclosing transaction details
  - Implementation benefits: Recent optimizations have improved performance through precomputation, batched operations, and hardware acceleration
- **Differential privacy:** Adds statistical noise to hide individual contributions while preserving overall data utility (Abadi et al. 2016). Features include
  - Centralized noise application by server or local noise addition by clients
  - Privacy budget tracking using moments accountant
  - Gradient clipping and randomized perturbations for enhanced protection
- **Federated learning:** Enables model training on local devices while sharing model updates with a central server only (Rawls 1971). This distributed approach greatly reduces the risk of exposing sensitive information, balancing privacy protection and data utilization (National Artificial Intelligence Initiative Office 2021; Rawls 1971). Federated learning holds great potential, especially in fields with sensitive data such as health care, finance, and smart cities (Li et al. 2020). For example, hospitals can develop a joint AI model without collecting patient

Box 3. The case of Korea: How Korea leveraged trustworthy AI to combat COVID-19 while ensuring privacy protection

As one of the leading science and technology research institutes in the Republic of Korea, Korea Institute of Science and Technology (KIST) took charge of estimating the spread of COVID-19 and formulating response strategies during the pandemic (KIST 2020). KIST integrated big data analysis and AI modeling to track infection routes and optimize quarantine measures. By analyzing nationwide infection case with telecommunications data, KIST assessed regional risk levels and efficiently allocated quarantine resources. Using AI-based predictive models to simulate future infection trends provided scientific grounds for the Korean government's adjustments to social distancing. KIST contributed to enhancing effectiveness of the pandemic responses by developing diagnostic kits and conducting therapeutic research utilizing AI technology, driving innovation in the medical field.

To ensure responsible data use and manage personal data, KIST implemented a range of privacy-preserving measures while achieving its research objectives by adhering to key principles and methods:

- **Personal data collection and use:** KIST adhered to the principle of collecting only the minimum necessary personal data for the research objectives and ensured that data anonymization was implemented whenever possible to prevent privacy breaches. The purpose of data use was clearly defined to avoid unnecessary data collection.
- **Data protection measures:** A range of technical, administrative, and physical safeguards, including data encryption, access control, and security training, were employed to minimize the risk of data breaches and misuse.
- **Data subject rights:** KIST ensured that data subjects' rights, including access, rectification, deletion, and processing restrictions and established procedures to facilitate these rights.
- **Data protection impact assessment:** This assessment was conducted during the project planning phase to assess potential privacy risks and implement necessary mitigation measures.

To balance research objectives with privacy protection, KIST implemented a combination of technical and procedural safeguards, including deidentification, anonymization, and strict data access controls. Specifically, anonymized data sets were used to prevent the identification of individuals, thereby reducing the risk of privacy breaches. Access to data unrelated to the research objectives was restricted, and permissions were granted only on a need-to-know basis. The expertise gained through this process not only contributed to the refinement of privacy-preserving techniques but also informed the development of Korea's broader AI governance strategies.

Source: Original to this publication.

data centrally, improving diagnostic accuracy and enhancing medical services (McMahan et al. 2017). It also offers benefits such as reducing the load on communication networks and saving data transmission costs (Yang et al. 2019). Each participant trains the model using local data on their devices or servers and transmits only the resulting model parameters to a central server for integration (UNHCR 2018). Through this, the original personal data is securely stored on each device, minimizing the risk of data leakage or privacy infringement (Mantelero 2018).

- *Core privacy features:* The original data never leaves user devices. Only model parameters are transmitted, and updates are temporarily stored for aggregation only.
- *Training process:* Involves random client selection, model parameter broadcasting, local computation on client devices, and secure parameter aggregation at the server.
- *Enhanced protection:* Employs multiple complementary techniques:
  - Differential privacy with calibrated noise addition
  - Secure aggregation using cryptographic protocols
  - Hardware-based isolation through trusted execution environments (TEEs)
- *Privacy in depth:* Combines multiple protective layers including local data confinement, differential privacy, secure aggregation, and TEEs. This multilayered approach is crucial for maintaining privacy standards in sensitive sectors such as health care and finance.

6. AI GOVERNANCE

As AI technologies become increasingly embedded in public and private decision-making, the demand for robust governance frameworks has gained global attention. Ensuring the safe, ethical, and effective use of AI requires not only technical solutions but also comprehensive regulatory and institutional responses. This section examines global efforts to establish shared principles and norms, along with country-level approaches.

6.1. GLOBAL FRAMEWORKS

With the rapid improvement in AI research and development, international organizations and governments have established regulations and policies for AI (OECD 2019b; Torrey and Goertzel 2016). Each aims to promote the safe and ethical use of AI, and they eventually allow people to manage potential risks in AI. Many governments are developing frameworks tailored to their specific contexts, while also collaborating internationally to establish consensus-driven principles and standards.

6.1.1. United Nations

The United Nations (UN) Secretary-General established a high-level Advisory Body on Artificial Intelligence to promote a globally inclusive approach in making recommendations for AI governance. Consisting of AI experts around the globe, the body delivered a report, “Governing AI for Humanity” with the UN General Assembly resolution on trustworthy AI. Finalized in September 2024, the report has established a comprehensive framework for AI research, development, and application (United Nations 2024). This framework emphasizes aligning AI technologies with human rights, transparency, and accountability. It advocates for international cooperation to establish global norms and standards for AI governance, addressing both opportunities and risks. The UN’s approach focuses on inclusive, ethical AI that benefits all people, while mitigating the potential for harm through robust regulatory and policy measures.

6.1.2. World Bank

In 2024, The World Bank published the “Global Trends in AI Governance” report to explore the emerging landscape of AI governance and guide policy makers in developing and deploying AI in an ethical, transparent, and accountable manner. The report discusses four governance tools that countries can adapt to their specific needs, highlighting their advantages and disadvantages with examples (World Bank Group 2024):

1. **Industry self-governance:** Companies such as Google and Microsoft develop and adopt voluntary ethical business standards. These can influence practices but lack enforcement and may lead to “ethics-washing” risks.
2. **Soft law:** Nonbinding principles and technical standards offer flexibility but may not clearly define rights or responsibilities.
3. **Regulatory sandboxes:** These controlled environments allow for the testing of innovative regulatory approaches but can be very resource-intensive to manage.
4. **Hard law:** Binding frameworks, such as the EU Artificial Intelligence Act or national legislation, provide consistency and legal certainty. However, they need to be tailored to the local context, considering existing capacity and resources.

The World Bank has also developed an Artificial Intelligence Risk Management Framework that is geared toward managing risks, promoting responsible AI practices within the organization.

6.1.3. Organisation for Economic Co-operation and Development

In 2019, OECD introduced the first intergovernmental standard on AI, known as the OECD AI Principles. These principles aim to encourage innovation and build trust in AI by promoting the responsible management of trustworthy AI, while ensuring respect for human rights and democratic values. Updated in 2024, the AI Principles are composed of five value-based principles and five recommendations, offering practical and adaptable guidance (OECD 2024) (table 3).

Table 3. OECD AI principles and recommendations

Principles	Recommendations
Inclusive growth, sustainable development, and well-being	Investing in AI research and development: Long-term public investment in research and development, open-source tools, and open data sets
Respect for the rule of law, human rights, and democratic values, including fairness and privacy	Fostering a digital ecosystem for AI: Development of and access to inclusive, dynamic, sustainable, and interoperable digital ecosystem for trustworthy AI
Transparency and explainability	Shaping an enabling policy environment for AI: Agile policy environment that supports transitioning from research and development stage to the deployment and operation stage for trustworthy AI systems, along with assessment mechanisms
Robustness, security, and safety	Building human capacity and preparing for labor market transformation: Cooperation with stakeholders and promotion of social dialogue to prepare for the fair transition and responsible use of AI
Accountability	International cooperation for trustworthy AI: Promotion of cooperation among governments and stakeholders and development of consensus-driven indicators and technical standards

Source: OECD (2024).

6.2. GOVERNMENTS

Table 4 provides a comparative overview of the regulatory framework for AI in selected governments. Each government has adopted strategies to address the opportunities and risks associated with AI, reflecting their priorities in fostering innovation while safeguarding ethical principles, public safety, and human rights.

In addition, governments have adopted legal and technical measures to regulate personal data protection, with a particular focus on data de-identification, anonymization, encryption, and privacy-enhanced technologies. Each government listed in table 5 has developed frameworks to balance data utilization with privacy protection, ensuring both innovation and public trust in AI systems.

Table 4. Regulatory frameworks for AI development and utilization

Governments	Key initiatives	Regulations
European Union	AI Act: The first government-level act that regulates AI systems based on their risk levels	EU's AI Act regulates AI systems based on their risk levels, including strict regulations for high-risk AI systems and self-regulation for lower-risk systems (European Commission 2021). The risks are categorized into four levels: unacceptable risk, high risk, limited risk, and minimal risk. For example, real-time facial recognition technology is classified as high-risk and is subject to strict regulations. The EU mandates technical and administrative measures to ensure AI's transparency, safety, and respect for human rights (European Commission 2020).
Japan	Society 5.0: Promotes human-centered AI utilization (Strategic Council for AI Technology 2017)	The government established the AI Technology Strategy to enhance industrial application and global competitiveness in AI (White House 2019). The AI Ethics Guidelines were also enacted to clarify ethical principles and responsibilities in AI development and utilization (International Telecommunication Union 2018).
Korea	AI Basic Act: Establishes a comprehensive regulatory framework for AI	In December 2024, the government passed the Basic Act on the Development of AI and Establishment of Trust, aimed at fostering AI innovation while addressing ethical, safety, and societal concerns with a focus on transparency requirements, establishment of ethical guidelines for the development and application of AI, and a classification framework to identify high-impact AI systems (Ministry of Science and ICT 2024). Previously, the government established its AI Ethics Guidelines to recommend adherence to personal data protection and ethical principles in AI system development and utilization (Ministry of Science and ICT 2019).
United States	National AI Initiative Act: Establishes long-term strategies and coordination bodies in the AI sector (Strategic Council for AI Technology 2017)	Through the National Institute of Standards and Technology (NIST), the government is developing technical standards and guidelines to ensure the reliability, fairness, and transparency of AI systems (NIST 2020). In January 2025, the government introduced a regulatory framework to manage the responsible development of advanced AI, focusing on controlling advanced computing chips and model weights for powerful AI systems (US Bureau of Industry and Security 2025). The rules are designed to prevent AI misuse for military or harmful uses while encouraging innovation, with exemptions for allied countries, certain supply chains, and low-volume applications.

Source: Compiled on the basis of the cited references.

Table 5. Data protection measures and privacy regulations

Governments	Key measures
European Union	● The General Data Protection Regulation specifies technical measures (European Data Protection Board 2018; European Union 2016) for personal data protection, active data deidentification, anonymization, and encryption ● Privacy-enhancing technologies promote data minimization, anonymization, pseudonymization, and encryption (ENISA 2018) mandating firms and institutions use encryption to prevent data exposure in case of breaches (ENISA 2018)
Japan	● Act on the Protection of Personal Information legalizes data deidentification and anonymization for analysis and AI development (Garfinkel 2015; NIST 2020) ● Guidelines for Balancing Data Utilization and Personal Data Protection propose technical measures for data protection, emphasizing encryption and secure data management (NIST 2020)
Korea	● Personal Information Protection Act and Credit Information Use and Protection Act institutionalizes personal data deidentification and pseudonymized data utilization (NIST 2019) ● The 2020 amendment expands pseudonymized data utilization to promote the data economy ● AI Ethics Guidelines recommend compliance with personal data protection and ethical principles in AI development and utilization (Ministry of Science and ICT 2019)
United States	● The National Institute of Standards and Technology provides technical standards and guidelines for personal data protection and supports the development and application of data deidentification and encryption technologies (Kairouz et al. 2021) ● The privacy framework guides organizations in managing personal data risks and implementing protection measures (European Commission 2018) ● The Federal Trade Commission encourages companies to use encryption and deidentification technologies for personal data protection. It aims to balance data utilization and privacy (ENISA 2018)

Source: Compiled on the basis of the cited references.

7. AI AND SOCIETY

The widespread adoption of AI technologies is reshaping not only how the governments and businesses operate but also how individuals engage with society. While AI offers opportunities to improve service responsiveness and enhance the quality of life, it also raises social concerns. Key issues include the digital divide, potential job displacement resulting from automation, and the risk of marginalizing vulnerable populations. This section examines the societal impacts of AI adoption and strategies to ensure that the benefits of AI are equitably shared across all segments of populations.

7.1. DIGITAL DIVIDE OR JOB DISPLACEMENT

While AI offers immense potential for societal advancement, it also poses significant risks of exacerbating existing inequalities. The digital divide, a persistent challenge, can be further deepened by AI because access to technology and digital literacy are crucial for reaping its benefits. Job displacement is another pressing concern because AI-powered automation has the potential to render certain jobs obsolete (Federal Trade Commission 2012).

To mitigate these risks, policy makers must prioritize digital inclusion initiatives. This includes expanding access to affordable, high-quality internet, promoting digital literacy programs, and ensuring that AI technologies are designed with accessibility in mind. Additionally, robust social safety nets and retraining programs can help workers adapt to the changing job market and acquire the skills necessary to thrive in the AI era.

Government or public policy designers should promote digital inclusion policies to ensure all citizens equally benefit from AI technologies (Government of Japan 2015). Strengthening social safety nets and providing retraining programs can alleviate workers' anxieties in the face of job transitions (PPC Japan 2016).

7.2. RESPONSIVENESS VIA AI

AI offers a powerful tool for enhancing the responsiveness and personalization of public services. By analyzing vast amounts of data, governments can gain valuable insights into the needs and preferences of their citizens. AI-powered chatbots can provide 24/7 support, while predictive analytics can help anticipate and address emerging challenges (Mehr 2017; UNESCO 2022). However, it is crucial to ensure that AI-driven public services are ethical, transparent, and accountable. Governments must establish clear guidelines for the use of AI, including data privacy protections and algorithmic bias mitigation. Moreover, it is essential to involve citizens in the development and deployment of AI systems to ensure that their needs and concerns are adequately addressed.

7.3. MEASURES TO SUPPORT MARGINALIZED GROUPS

AI policies must be designed to benefit all members of society, including vulnerable and marginalized groups. This requires a nuanced understanding of the specific challenges faced by these populations and the potential impact of AI on their lives (OECD 2018). People with disabilities can benefit from AI-powered assistive technologies, but it is essential to ensure that these technologies are accessible and inclusive. Elderly individuals can leverage AI to maintain their independence and quality of life, but they may require additional support to navigate the complexities of the digital world. Low-income individuals may be disproportionately affected by job displacement and economic inequality, so it is crucial to provide them with the skills and opportunities they need to thrive in the AI era (National Police Agency of Japan 2019). By adopting a human-centered approach to AI, policy makers can harness its potential to create a more equitable and prosperous future for all.

8. SELF-ASSESSMENT FOR AI USE

To support the adoption and utilization of trustworthy AI, it is important for governments and organizations to evaluate their readiness and performance. This section offers self-assessment tools that can serve as practical benchmarks for tracking progress over time, identifying gaps, and adjusting strategies accordingly.

8.1. BACKGROUND

As AI technologies continue to evolve, governments and organizations across sectors are increasingly exploring their potential for enhancing public governance and service delivery. While some entities may be prepared to swiftly adopt AI, others may need to address foundational gaps before moving forward. Key factors influencing these decisions include the availability of technical expertise, data management infrastructure, and adherence to legal and ethical frameworks for emerging technologies.

A self-assessment for AI use offers a structured approach to evaluate readiness, helping identify areas for growth and improvement. By focusing on critical aspects such as data privacy, security, and technical infrastructure, organizations can better understand their current capabilities and pinpoint gaps. This process enables informed decision-making and ensures that AI adoption is responsible, efficient, and aligned with relevant policies and international standards.

Given the varying levels of preparedness, particularly in developing regions, self-assessment serves as a valuable tool for guiding AI adoption in a thoughtful and efficient manner. It empowers organizations to evaluate their unique needs and tailor AI strategies accordingly, fostering informed decision-making and ensuring alignment with broader goals and priorities. Governments, particularly developing countries, can benefit from checklists and toolkits to assess the suitability and readiness of AI systems independently. Such tools enable key stakeholders and organizations to proactively review the ethical, legal, and technical aspects of AI technology, identify potential risks, and develop strategies for successful implementation (OECD 2019b; Sheller et al. 2020). The expected benefits of using the checklists include:

- **Preemptive risk management:** Identifying potential risks and issues before AI introduction to be able to respond in advance (Bonawitz et al. 2019).
- **Efficient resource utilization:** Recognizing areas with low readiness to allocate resources efficiently and set priorities (Bughin et al. 2017).

- **Securing public trust:** Ensuring transparency and systematic evaluation, which helps build public trust and increase the acceptability of AI policies (European Commission 2020).
- **Enhancing international competitiveness:** Improving competitiveness in the global market by aligning with international standards and best practices (G20 2019).

8.2. AI READINESS CHECKLISTS

AI systems are inherently data-driven. The quality, integrity, and ethical handling of data throughout the AI life cycle—from collection and processing to model training and deployment—are paramount. “AI cannot exist without data, and governed AI cannot exist without governed data” (Collibra 2024) underscores data’s fundamental role. Poor data quality hinders AI deployment; even the most advanced algorithms produce flawed results if the underlying data are poor (Ataman 2025). Ethical AI development intrinsically depends on responsible data practices. Issues like bias, fairness, accountability, and transparency are deeply intertwined with how data are managed (Berryhill et al. 2019). Data privacy and governance act as enablers for broader AI ethical principles:

- **Fairness and nondiscrimination:** Biased data used to train AI models can lead to discriminatory outcomes. Ethical data collection, meticulous preprocessing, and bias audits—all aspects of data governance—are critical for mitigating AI bias and promoting fairness. Checklist items focusing on data quality and bias assessment directly support this.
- **Transparency and explainability:** Understanding the data used to train an AI model (its sources, characteristics, limitations) is a prerequisite for explaining the model’s behavior (IBM 2025). Clear data governance policies contribute to overall system transparency (Collibra 2024). Checklist items on data documentation and provenance are essential here.
- **Accountability:** Robust data governance establishes clear responsibilities for data handling, which is vital for AI system accountability (Collibra 2024). If an AI system produces adverse outcomes, tracing data lineage and processing steps (ensured by good governance) is crucial for identifying liability. Checklist items on data ownership and access control support accountability.
- **Security and robustness:** Protecting data from unauthorized access, corruption, or misuse is fundamental to the security and robustness of AI systems. Checklist items on secure data storage and access control are directly relevant. To facilitate the process, we provide two checklists (figure 1 and table 6). Both serve a purpose to establish critical standards for data handling necessary for realizing comprehensive AI ethics, with the first centering on data collection and processing and the second focusing on data privacy considerations specific to AI use.

While AI ethics is multifaceted, the checklists’ focus on data privacy and governance is not a limitation but a strategic choice. Ethical data management is an essential first step toward achieving broader AI ethical principles such as fairness, accountability, and transparency. Data ethics is not merely one of many ethical concerns; rather, it forms a foundational element supporting all other AI ethical principles. Without robust data practices such as privacy protection, security, quality assurance, and bias mitigation, the overall trustworthiness and ethical stability of an AI system can be compromised.

Also, this data-centric approach and ethical data handling (e.g., consent, minimization, security, and bias checks (National Assembly Research Service 2021) aligns with major international and globally recognized standards to address fundamental prerequisites for ethical AI; the EU AI Act, OECD AI Principles, and the UNESCO Recommendation on the Ethics of AI all treat data governance and data protection as core requirements for trustworthy AI.[1]

These tools can be incorporated into a comprehensive AI application evaluation toolkit for developing countries, including the following features:

- **Automated evaluation system:** An online platform, where entering answers for each item automatically calculates a readiness score.
- **Feedback and recommendations:** Provides feedback on deficiencies and suggests improvements based on evaluation results.
- **Resources and materials:** Offers educational materials, guidelines, and best practices related to each item to support capacity building.
- **Community and networking:** Includes community features to promote experience sharing and collaboration with other countries or organizations.

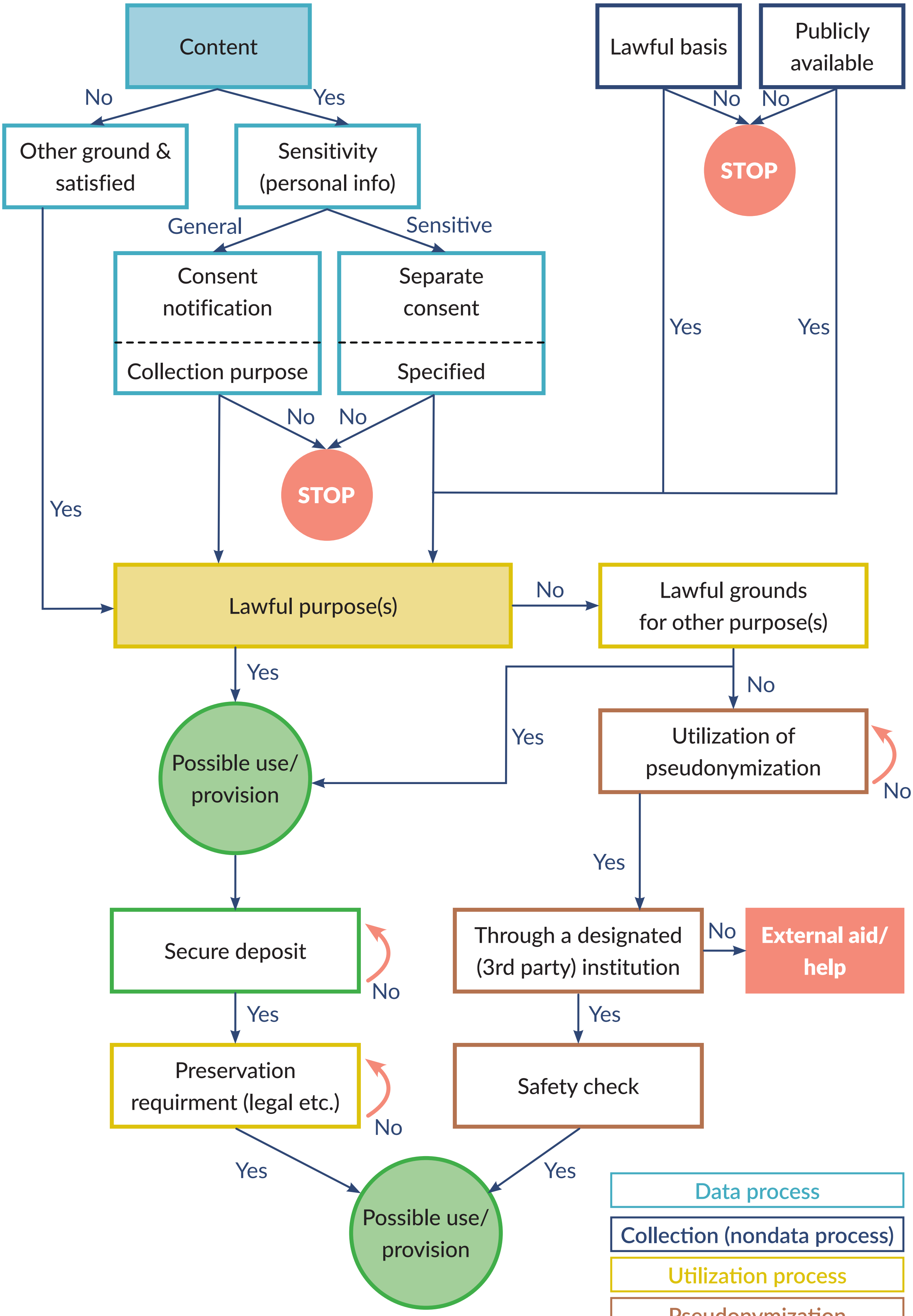
Figure 1 presents a step-by-step checklist designed to help stakeholders assess the appropriateness of applying AI technologies to data collection and processing. Each step is framed as a binary question (yes or no), and responses determine the next relevant step, enabling a customized and context-sensitive evaluation. To support intuitive navigation, the checklist uses color coding to distinguish the stages of the process:

- **Teal box**: Key data processing decisions (e.g., consent and legal basis)
- **Blue box:** Data collection and availability
- **Yellow box**: Data utilization stage—reuse, share, dispose
- **Brown box**: Pseudonymization and safety measures

This flowchart helps identify legal, ethical, and operational conditions that must be met before deploying AI technologies in data environments.

[1] The EU AI Act emphasizes high-quality, representative training data to avoid discrimination and specifies data governance requirements for high-risk AI systems (European Commission 2021). The OECD AI Principles include "[h]uman rights and democratic values, including fairness and privacy" as a core value and recommend policies for an "AI-enabling ecosystem," implying strong data governance. UNESCO explicitly states that "[p]rivacy must be protected and promoted throughout the AI lifecycle. Adequate data protection frameworks should also be established," and its AI Readiness Assessment Methodology includes data privacy (UNESCO 2022).

Figure 1. **Checklist for AI data collection and processing**



Source: Original to this publication.

8.2.1. Decision flow for AI data collection and processing

The process begins by examining whether the data in question contain information that requires user consent or falls under other lawful grounds for processing. If personal information is involved, the flowchart differentiates between general and sensitive data types, prompting either a standard consent notification or a separate, explicit consent.

Once a legal basis or valid consent is established, users are asked to verify whether the intended AI use remains consistent with the original data collection purpose. If the purpose has changed, the checklist guides users to either establish lawful justification for secondary use or apply pseudonymization techniques to minimize risk.

When pseudonymization is employed, the checklist introduces additional safeguards. Specifically, the data must be processed through an authorized third-party institution and undergo a formal safety assessment to verify that adequate protective measures are in place. If these conditions are not satisfied, the AI application is deemed noncompliant, and the process is halted to prevent inappropriate use.

For data that remain within the original purpose, the checklist then considers whether the data can be reused or shared with others. If permitted, users are guided to verify conditions for secure disposal. In cases where disposal is not immediately feasible, legal or institutional data preservation requirements must be confirmed.

The process is considered complete only when all required legal, ethical, and technical safeguards have been addressed. This structured flow ensures that AI technologies are deployed responsibly and in compliance with applicable governance standards.

8.3. CHECKLIST FOR DATA PRIVACY IN AI

To complement the decision flow described above, table 6 provides a detailed checklist to assess data privacy readiness in the context of AI applications. Each item reflects a key consideration drawn from data protection principles, ranging from lawful collection and consent to data retention, reuse, and pseudonymization.

The checklist is organized into five categories:

1. Collection of data directly from the data subject
2. Collection from third-party or public sources
3. Use and provision of personal information
4. Retention and disposal
5. Processing of pseudonymized information

Table 6. Checklist for data privacy in AI

Step	Question	Yes	No
1. Collection from the data subject			
1-1	Did you obtain consent from the data subject?		
1-2	Did you determine whether the information is sensitive personal information (racial or ethnic origin, political opinions, religious beliefs, health data, genetic data, biometric data, sexual orientation)?		
1-3	Have you obtained informed, specific, unambiguous, and freely given consent by clearly informing the purpose and collecting minimum necessary data?		
1-4	If collecting sensitive or unique identification information (e.g., Social Security number, passport number), have you obtained separate explicit consent?		
1-5	Does collection meet another lawful basis (legal obligation, contract performance, vital interests, public task, legitimate interests)?		
2. Collection from other than the data subject			
2-1	When collecting personal information from a third party, do you have a lawful basis (consent, legitimate interests, legal obligation)?		
2-2	For publicly available information, is collection proportionate, within a reasonable scope, and for a legitimate purpose that respects data subject rights?		
3. Use and provision of personal information			
3-1	Are you limiting the use and provision of personal information strictly to the original stated purpose?		
3-2	Is there a separate lawful basis (new consent, legal obligation) for using or providing data beyond the original purpose?		
4. Retention and disposal of personal information			
4-1	Do you securely dispose (shredding, secure deletion) of personal information when it's no longer needed for the stated purpose or required by law?		
4-2	Are you segregating and securely protecting personal information requiring longer retention (legal/regulatory compliance)?		
5. Processing pseudonymized information			
5-1	Is pseudonymized information processed only for specific, legitimate purposes permitted by law (e.g., statistics, scientific research, public record preservation)?		
5-2	If combining pseudonymized data, is this handled securely by a designated, trusted entity with technical and organizational safeguards?		
5-3	Can anonymization be safely achieved without losing data utility (preferred over pseudonymization)?		

Source: Original to this publication.

Each question is answered using a conservative binary (yes or no) format reflecting the strict and nonnegotiable nature of data privacy requirements. In areas such as consent, legal basis, or data disposal, partial implementation does not meet compliance thresholds. A “no” response is not a judgment of failure but a signal that follow-up action is needed to meet minimum standards.

The current checklist design emphasizes clarity, usability, and legal defensibility—particularly for early-stage or resource-constrained environments. A greater number of “yes” responses indicates stronger alignment with privacy principles and higher readiness for responsible AI deployment. This format provides a practical tool for implementers and policy makers to identify gaps, track progress, and prioritize actions toward building privacy-compliant and trustworthy AI systems.

8.4. CONSIDERATIONS IN USING CHECKLISTS

The AI readiness and data privacy checklists serve as helpful tools for organizations seeking to assess their preparedness for AI adoption. By systematically evaluating key areas such as data infrastructure, ethical guidelines, and security considerations, these checklists provide a structured approach that supports informed decision-making. They help organizations understand their strengths and identify areas that may require additional focus, enabling more targeted and effective planning for AI implementation.

While the checklists provide essential guidance, they should be viewed as a starting point rather than a definitive solution. AI adoption is complex and context dependent, with varying needs across organizations. These checklists can help highlight crucial factors for consideration, but each organization may face unique challenges or regulatory requirements that require tailored solutions beyond what is covered in the checklist. Therefore, the checklists should be used in conjunction with other tools and expert assessments to ensure a comprehensive and nuanced approach.

It is also important to note that these checklists promote an iterative process of reflection and improvement. As technologies and regulatory environments evolve, these tools can help organizations revisit key considerations and track incremental progress. While they are not designed as comprehensive governance frameworks, they offer a starting point for identifying gaps and prompting internal dialogue. Used periodically and in combination with broader planning efforts, the checklists can contribute to more informed, responsible, and context-aware AI adoption.

9. IMPLICATIONS

The increasing adoption of AI across the public sector calls for a shift from aspirational principles to actionable practices. While global frameworks and ethical standards provide valuable guidance, translating them into context-sensitive, operational actions remain a challenge for many governments. This report responds to that gap by outlining the core dimensions of trustworthy AI and introducing practical, low-barrier tools: an AI readiness flowchart and a data privacy checklist. These instruments are designed to help governments and development actors assess foundational safeguards before deploying AI, by prompting structured reflection on legal, ethical, and organizational preconditions.

Importantly, these tools are not intended as prescriptive checklists or rigid audits but rather as adaptive aids to support internal decision-making. In this capacity, they help foster more deliberate and well-grounded decisions across public sector institutions engaged in the adoption and use of trustworthy AI. By translating complex legal and ethical considerations into a structured set of binary prompts, the tools encourage early-stage discussions when adjustments are still feasible and cost effective. When used judiciously and in conjunction with broader governance processes, they can help clarify roles, surface potential risks, and strengthen collaboration across legal, technical, and policy functions.

Trustworthy AI governance must be approached not as a one-time compliance exercise but as a continuous process of reflection, coordination, and adaptation. As AI systems become more integrated into public institutions, advancing their responsible use will require not only technical safeguards but also sustained institutional engagement. Looking ahead, further efforts may be needed to refine and contextualize tools such as those proposed in this report, strengthen internal capacity, and promote knowledge sharing across governments and sectors. While modest in scope, this contribution aims to support ongoing efforts to align the adoption of AI with public sector values of accountability, inclusion, and transparency.

REFERENCES

Abadi, Martín, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talway, and Li Zhang. 2016. "Deep Learning with Differential Privacy." In *CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–18. New York: Association for Computing Machinery.

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–60.

Arya, Vijay, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques." arXiv: 1909.03012.

Ataman, Altay. 2025. "Data Quality in AI: Challenges, Importance & Best Practices." *AIMultiple*, July 9. https://research.aimultiple.com/data-quality-ai/AIMultiple+1AIMultiple+1.

Bellamy, Rachel K. E., Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias." *IBM Journal of Research and Development* 63 (4–5): 4:1–4:15.

Berryhill, Jamie, Kévin Kok Heang, Rob Clogher, and Keegan McBride. 2019. "Hello, World: Artificial Intelligence and its Use in the Public Sector." OECD Working Paper on Public Governance 36, Organisation for Economic Co-operation and Development, Paris.

Bonawitz, Keith, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloé Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, and Jason Roselander. 2019. "Towards Federated Learning at Scale: System Design." In *Proceedings of the 2nd SysML Conference*, edited by Ameet Talwalkar, Virginia Smith, and Matei Zaharia, 374–88.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.

Bryson, Joanna J. 2020. "The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 2–25. Oxford, UK: Oxford University Press.

Bughin, Jacques, Eric Hazan, Sree Ramaswamy, Michael Chiu, Tera Allas, Peter Dahlström, Nicolaus Henke, and Monica Trench. 2017. *Artificial Intelligence: The Next Digital Frontier?* McKinsey Global Institute.

Coeckelbergh, Mark. 2020. *AI Ethics*. Cambridge, MA: MIT Press.

Collibra. 2024. "Understanding the Importance of Data Governance in the Age of AI." *Collibra Blog*, August 19. https://www.collibra.com/blog/understanding-the-importance-of-data-governance-in-the-age-of-ai.

Danks, David, and Alex John London. 2017. "Algorithmic Bias in Autonomous Systems." In *IJCAI '17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*, edited by Carles Sierra, 4691–97. Melbourne: AAAI Press.

Deloitte. 2017. "The Power of Artificial Intelligence for Government." *Deloitte Insights*.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv: 1702.08608.

Dreyfus, Hubert L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press.

Eggers, William D., David Schatsky, and Peter Viechnicki. 2017. "AI-Augmented Government: Using Cognitive Technologies to Redesign Public Sector Work." *Deloitte Insights*, April.

El Emam, Khaled, and Luk Arbuckle. 2013. *Anonymizing Health Data*. Sebastopol, CA: O'Reilly Media.

ENISA (European Union Agency for Network and Information Security). 2018. *Recommendations on Shaping Technology According to GDPR Provisions*. Attiki, Greece: European Union. https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions.

European Commission. 2018. *Artificial Intelligence for Europe*. COM (2018) 237. Brussels: European Commission.

European Commission. 2020. "White Paper on Artificial Intelligence: A European Approach to Excellence and Trust." COM (2020) 65, European Commission, Brussels.

European Commission. 2021. *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. COM (2021) 206. Brussels: European Commission.

European Data Protection Board. 2018. "Guidelines on Personal Data Breach Notification under Regulation 2016/679." European Commission, Brussels.

European Parliament. 2017. "Civil Law Rules on Robotics." Luxembourg City: European Parliament.

European Union. 2016. General Data Protection Regulation (GDPR). Regulation (EU) 2016/679. *Official Journal of the European Union*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Federal Trade Commission. 2012. *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers*. Washington, DC: Federal Trade Commission.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. "Principled Artificial Intelligence.: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." Research Publication 2020-1, Berkman Klein Center for Internet and Technology, Harvard University, Cambridge, MA.

Floridi, Luciano. 2014. *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford, UK: Oxford University Press.

Floridi, Luciano, and Josh Cowls. 2019. "A Unified Framework of Five Principles for AI in Society." *Harvard Data Science Review* 1 (1).

Floridi, Luciano, and Mariarosaria Taddeo. 2016. "What Is Data Ethics?" *Philosophical Transactions of the Royal Society A* 374 (2083): 20160360.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–232.

G20. 2019. "G20 Ministerial Statement on Trade and Digital Economy." https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf.

Garfinkel, Simson L. 2015. *De-identification of Personal Information*. NIST IR 8053, National Institute of Standards and Technology.

Gentry, Craig. 2009. "A Fully Homomorphic Encryption Scheme." PhD thesis, Stanford University, Stanford, CA.

Geyer, Robin C., Tassilo Klein, and Moin Nabi. 2017. "Differentially Private Federated Learning: A Client Level Perspective." arXiv: 1712.07557.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. arXiv: 1806.00069.

Government of Japan. 2015. Act on the Protection of Personal Information (APPI). https://www.ppc.go.jp/en/legal/.

Government of Japan. 2016. *Society 5.0*. Cabinet decision of January 22.

Gunning, David. 2017. "Explainable Artificial Intelligence (XAI)." DARPA I/20 Program Update. https://nsarchive.gwu.edu/sites/default/files/documents/5794867/National-Security-Archive-David-Gunning-DARPA.pdf.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." NIPS '16: *Proceedings of the 30th International Convention on Neural Information Processing Systems*, edited by Daniel Lee, Ulrike von Luxburg, Roman Garnett, Masashi Sugayama, and Isabelle Guyon, 3315–23. Red Hook, NY: Curran Associates.

IBM. 2025. "AI Ethics." https://www.ibm.com/artificial-intelligence/ai-ethics.

Information Commissioner's Office. 2020. "Guidance on AI and Data Protection." https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/.

International Organization for Standardization. 2011. "Information Technology–Security Techniques–Privacy Framework." ISO/IEC 29100:2011, International Organization for Standardization, Geneva.

International Telecommunication Union. 2018. *AI for Good Global Summit Report*. Geneva: International Telecommunication Union.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–99.

Kairouz, Peter, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Ben Hutchinson, Justin Hau, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tacrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. "Advances and Open Problems in Federated Learning." *Foundations and Trends in Machine Learning* 14 (1–2): 1–210.

KIST (Korea Institute of Standards and Technology). 2020. "AI Research on COVID-19 Response." https://www.kist.re.kr/.

Li, Tian, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. "Federated Learning: Challenges, Methods, and Future Directions." *IEEE Signal Processing Magazine* 37 (3): 50–60.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability." *Communications of the ACM* 61 (10): 36–43.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, edited by Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus, 4765–74. Red Hook, NY: Curran Associates.

Mantelero, Alessandro. 2018. "AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment." Computer Law and Security Review 34 (4): 754–72.

Margetts, Helen, and Cosmina Dorobantu. 2019. "Rethinking Public Policy in the Era of Big Data and AI." *Policy and Internet* 11 (1): 1–4.

Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–83.

McMahan, H. Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. "Communication-Efficient Learning of Deep Networks from Decentralized Data." In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 1273–82. MLR Press.

Mehr, Hila. 2017. "Artificial Intelligence for Citizen Services and Government." Ash Center for Democratic Governance and Innovation, Harvard Kennedy School, Cambridge, MA. August. https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf.

Ministry of Internal Affairs and Communications of Japan. 2019. *AI Utilization Guidelines*. Tokyo: Ministry of Internal Affairs and Communications of Japan. August. https://www.soumu.go.jp/main_content/000658284.pdf.

Ministry of Science and ICT of the Republic of Korea. 2019. *National Strategy for Artificial Intelligence*. Seoul: Government of the Republic of Korea.

Ministry of Science and ICT of the Republic of Korea. 2024. "A New Chapter in the Age of AI: Basic Act on AI Passed at the National Assembly's Plenary Session." Press release, December 26. https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=1071&searchOpt=ALL&searchTxt=.

Mittelstadt, Brent. 2019. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1 (11): 501–7.

Molnar, Christopher. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* Leanpub.

Montag, Christian, Preslac Nakov, and Raian Ali. 2024. "Considering the IMPACT Framework." *Telematics and Informatics Reports* 13:100112.

Moor, James H. 2006. "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years." *AI Magazine* 27 (4): 87–91.

National Artificial Intelligence Initiative Office. 2021. "National AI Initiative Act of 2020." Washington, DC.

National Assembly Research Service of the Republic of Korea. 2021. "Necessity and Legislative Tasks for Enacting the Basic Law on Artificial Intelligence." Seoul.

National Police Agency of Japan. 2019. *Introduction of AI in Crime Prevention*. Tokyo: National Polic Agency of Japan.

NIST (National Institute of Standards and Technology). 2019. *U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools*. Washington, DC: NIST.

NIST (National Institute of Standards and Technology). 2020. *NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management*. Washington, DC: NIST.

OECD (Organisation for Economic Co-operation and Development). 2018. *Bridging the Digital Gender Divide*. Paris: OECD Publishing.

OECD (Organisation for Economic Co-operation and Development). 2019a. *Digital Government Review of Sweden: Towards a Data-Driven Public Sector*. Paris: OECD Publishing.

OECD (Organisation for Economic Co-operation and Development). 2019b. *OECD Principles on Artificial Intelligence*. Paris: OECD Publishing.

OECD (Organisation for Economic Co-operation and Development). 2024. "Recommendation of the Council on Artificial Intelligence." OECD/LEGAL/0449, as amended on May 3, 2024. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OECD (Organisation for Economic Co-operation and Development). n.d. "Artificial Intelligence in the Public Sector." Observatory of Public Sector Innovation (OPSI). Accessed August 7, 2025. https://oecd-opsi.org/work-areas/ai/.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

PPC (Personal Information Protection Commission) Japan. 2016. *Guidelines on the Act on the Protection of Personal Information (General Rules)*. PPC Notification No. 6 of 2016. Tokyo: PPC. https://www.ppc.go.jp/en/legal/guidelines/general/.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

RBC Borealis. 2022. "Industry Analysis: AI Fairness Toolkits Landscape." *Computer Vision* (blog), May 6.

Righetti, Luca, Raj Madhavan, and Raja Chatila. 2019. "Unintended Consequences of Biased Robotic and Artificial Intelligence Systems." *IEEE Robotics and Automation Magazine* 26 (3): 11–13.

Russell, Stuart, and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. 3rd ed. Essex, UK: Pearson Education.

Sanderson, Conrad, David Douglas, and Qinghua Lu. 2024. "Implementing Responsible AI: Tensions and Trade-Offs between Ethics Aspects." arXiv: 2304.08275.

Sheller, Micah J., Brandon Edwards, G. Anthony Reina, Jason Martin, Sarthak Pati, Aikaterini Kotrotsou, Mikhail Milchenko, Weilin Xu, Daniel Marcus, Rivka R. Cohen, and Spyridon Baskas. 2020. "Federated Learning in Medicine: Facilitating Multi-institutional Collaborations without Sharing Patient Data." *Scientific Reports* 10 (1): 1–12.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2013. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." arXiv: 1312.6034

Solove, Daniel J. 2006. "A Taxonomy of Privacy." *University of Pennsylvania Law Review* 154 (3): 477–560.

Stallings, William. 2017. *Cryptography and Network Security: Principles and Practice*. 7th ed. Essex, UK: Pearson Education.

Strategic Council for AI Technology. 2017. "Artificial Intelligence Technology Strategy." Government of Japan, Tokyo.

Sweeney, Latanya. 2013. “Discrimination in Online Ad Delivery.” *Communications of the ACM* 56 (5): 44–54.

Taddeo, Mariarosaria, and Luciano Floridi. 2018. “How AI Can Be a Force for Good.” *Science* 361 (6404): 751–52.

Tavani, Herman T. 2007. “Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy.” *Metaphilosophy* 38 (1): 1–22.

Torrey, Lance, and Ben Goertzel. 2016. *The AGI Revolution: An Inside View of the Rise of Artificial General Intelligence*. Dallas, TX: BenBella Books.

UNESCO (United Nations Educational, Scientific, and Cultural Organization). 2022. “Recommendation on the Ethics of Artificial Intelligence, Adopted 23 November 2021.” Document SHS/BIO/PI/2021/1, UNESCO, Paris.

UNESCO (United Nations Educational, Scientific, and Cultural Organization) and COMEST (World Commission on the Ethics of Scientific Knowledge and Technology). 2019. “Preliminary Study on the Ethics of Artificial Intelligence.” Document SHS/COMEST/EXTWG-ETHICS-AI/2019/1, UNESCO, Paris.

UNHCR (United Nations Human Rights Council). 2018. “Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression.” Document A/73/348, Geneva.

United Nations. 2024. *Governing AI for Humanity: Final Report*. New York: United Nations. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

US Bureau of Industry and Security. 2025. “Biden–Harris Administration Announces Regulatory Framework for the Responsible Diffusion of Advanced Artificial Intelligence Technology.” Press release, January 23. https://www.bis.gov/press-release/biden-harris-administration-announces-regulatory-framework-responsible-diffusion-advanced-artificial.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. 2017. “Attention Is All You Need.” In *NIPS ’17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, edited by Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus, 6000–10. Red Hook, NY: Curran Associates.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology* 31 (2): 841–87.

White House. 2019. “Executive Order on Maintaining American Leadership in Artificial Intelligence.” https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/.

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. “The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions.” In *AIES ’19: Proceedings of* *the AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. New York: Association for Computing Machinery.

Wirtz, Bernd W., Jan C. Weyerer, and Carolin Geyer. 2019. “Artificial Intelligence and the Public Sector-Applications and Challenges.” *International Journal of Public Administration* 42 (7): 596–615.

World Bank Group. 2024. *Global Trends in AI Governance: Evolving Country Approaches*. Washington, DC: World Bank. https://documents1.worldbank.org/curated/en/099120224205026271/pdf/P1786161ad76ca0ae1ba3b1558ca4ff88ba.pdf

World Economic Forum. 2020. *The Future of Jobs Report 2020*. Geneva: World Economic Forum.

Yang, Qiang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. “Federated Machine Learning: Concept and Applications.” In *ACM Transactions on Intelligent Systems and Technology* 10 (2): art. 12.

WORLD BANK GROUP