Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: www.tandfonline.com/journals/uaai20

AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development

Petar Radanliev

To cite this article: Petar Radanliev (2025) AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development, Applied Artificial Intelligence, 39:1, 2463722, DOI: 10.1080/08839514.2025.2463722

To link to this article: https://doi.org/10.1080/08839514.2025.2463722

APPLIED ARTIFICIAL INTELLIGENCE
2025, VOL. 39, NO. 1, e2463722 (41 pages)
https://doi.org/10.1080/08839514.2025.2463722

Taylor & Francis
Taylor & Francis Group

AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development

Petar Radanliev

Department of Computer Science, University of Oxford, Oxford, UK

ABSTRACT

The expansion of Artificial Intelligence in sectors such as healthcare, finance, and communication has raised critical ethical concerns surrounding transparency, fairness, and privacy. Addressing these issues is essential for the responsible development and deployment of AI systems. This research establishes a comprehensive ethical framework that mitigates biases and promotes accountability in AI technologies. A comparative analysis of international AI policy frameworks from regions including the European Union, United States, and China is conducted using analytical tools such as Venn diagrams and Cartesian graphs. These tools allow for a visual and systematic evaluation of the ethical principles guiding AI development across different jurisdictions. The results reveal significant variations in how global regions prioritize transparency, fairness, and privacy, with challenges in creating a unified ethical standard. To address these challenges, we propose technical strategies, including fairness-aware algorithms, routine audits, and the establishment of diverse development teams to ensure ethical AI practices. This paper provides actionable recommendations for integrating ethical oversight into the AI lifecycle, advocating for the creation of AI systems that are both technically sophisticated and aligned with societal values. The findings underscore the necessity of global collaboration in fostering ethical AI development.

ARTICLE HISTORY
Received 11 August 2024
Revised 5 September 2024
Accepted 2 February 2025

Introduction

In recent years, substantial advancements in AI ethics have emerged, with significant contributions addressing transparency, fairness, and privacy in AI development. Recent research studies (Bender et al. 2021) highlight the dangers of bias in large language models, raising concerns over the perpetuation of societal inequalities within AI systems. Similarly, Bommasani et al. (2023) critically evaluate compliance of foundation models with the draft EU AI Act, reflecting broader concerns over the accountability of AI systems at a foundational level. Moreover, Aldoseri, Al-Khalifa, and Hamouda (2023)

CONTACT Petar Radanliev petar.radanliev@cs.ox.ac.uk Department of Computer Science, University of Oxford, Oxford, UK

propose new data strategies for AI development, focusing on the integration of ethical principles across diverse datasets to mitigate bias. These works, alongside key regulatory frameworks from bodies such as the European Union (European Parliament 2023) and NIST (2024b), reinforce the imperative of developing AI systems that are not only transparent and fair but also respect data privacy and societal norms. By situating this study within the context of these contemporary advancements, the aim is to build upon these discussions by offering a comparative analysis of global AI policy frameworks, extending the dialog on how ethical AI can be systematically developed and maintained across diverse geopolitical contexts.

While this study references key AI policy frameworks from the European Union, the United States, and China, the focus of this study is a more granular examination of the challenges in comparing such diverse frameworks is crucial. AI ethics policies are deeply influenced by cultural, socio-political, and economic contexts, making cross-regional comparisons inherently complex. For instance, the European Union's AI Act places significant emphasis on safeguarding individual rights, prioritizing transparency and human oversight (European Parliament 2023), reflecting the EU's regulatory ethos aimed at protecting citizens from the potential harms of AI. In contrast, the United States' AI governance is more decentralized, with a focus on promoting innovation and maintaining global technological leadership, as seen in the National Institute of Standards and Technology (NIST) AI Risk Management Framework (2023a), which advocates for flexible, non-prescriptive guidelines that encourage industry-led solutions.

China's AI policy framework, meanwhile, is characterized by its focus on state security, social harmony, and the integration of AI into national economic strategies (Roberts et al. 2021). This framework aligns with China's broader governmental control over technology, where the state plays a central role in guiding AI development. These diverging priorities highlight the inherent challenges in creating universal standards for AI ethics. The task of comparing these frameworks, therefore, requires consideration of their distinct legal, cultural, and economic motivations, as well as the varying levels of public trust in AI technologies across these regions.

This study addresses these complexities by identifying common ethical principles such as fairness, transparency, and privacy, and by analyzing how each region interprets and prioritizes these principles. By employing comparative tools such as Venn diagrams and Cartesian graphs, the article visually and analytically demonstrates the differences, but also the common points in these frameworks. The aim of this study was not to look only for the differences, but to find a solution for global AI governance and to promote the potential for harmonization across jurisdictions.

This paper explores the pressing need for ethical considerations in the rapidly evolving domain of Artificial Intelligence (AI) (Meissner 2020). This

technology has significantly impacted various sectors, including healthcare, finance, and communication. This study aims to establish a robust ethical framework for AI development by addressing complex issues such as data privacy, algorithmic transparency, and fairness. Our objectives include analyzing fundamental ethical principles, comparing international AI policy frameworks (Helbing et al. 2018), proposing strategies for bias mitigation, and contributing to academic and practical discussions in AI ethics. This paper is structured to systematically dissect these topics, providing an in-depth exploration of AI ethics and its implications for future AI development and governance (de Fine Licht and de Fine Licht 2020).

The Imperative of Ethical Considerations in AI

The advent of Artificial Intelligence (AI) has inaugurated a new epoch in technological evolution, profoundly influencing diverse sectors, including healthcare, finance, transportation, and communication (Hosny et al. 2018; NIST 2023b; Yu, Beam, and Kohane 2018). This unprecedented integration of AI into the societal fabric necessitates the urgent formulation of robust ethical frameworks. These frameworks must address the complexities inherent in AI technologies, such as data privacy, algorithmic opacity, equity in decision-making, and broader societal impacts.

Ethical considerations in AI transcend academic discourse, bearing significant real-world repercussions. Paramount among these are issues related to data privacy and the need for informed consent, where personal information often powers AI algorithms. Equally critical is the transparency and explicability of these algorithms, which are essential for sustaining public trust, especially in high-stakes scenarios like legal adjudication or medical diagnostics. Moreover, the challenge of ensuring equity and circumventing ingrained biases in AI systems is a pivotal ethical imperative, given these systems' propensity to mirror and perpetuate existing societal disparities.

The need for ethical AI is driven by the imperatives of harm prevention and justice but also by the strategic objective of nurturing sustainable, socially beneficial, and universally accepted innovation.

Aims and Objectives of the Study

The primary goals of this academic study are threefold. First, it seeks to explore the ethical considerations inherent in the development of AI. This involves thoroughly examining the fundamental ethical principles of transparency, equity, and privacy within AI systems and understanding how they relate to each other and their significance in isolation. Second, the research aims to critically analyze various global AI policy frameworks, focusing on those from the EU, the US, and China. The goal is to discern their similarities and

differences and what they mean for international AI governance. Third, the paper intends to provide a synthesis of approaches and practices to recognize and mitigate bias in AI systems, ensuring their fairness and dependability. This study aims to contribute to the ongoing academic and practitioner dialog on AI ethics by offering relevant insights and recommendations to both groups. The overarching goal is to promote a more ethical and responsible path for AI development.

Structure and Content of the Paper

The paper has been methodically structured to explore AI ethics across various dimensions systematically. Section 2, expands into the foundational ethical principles in AI: transparency, equity, and privacy. The section uses a Venn diagram to demonstrate the interaction between these principles, emphasizing their interconnectedness and how they relate to AI.

Moving on to section 3, the focus is on integrating global AI policy frameworks within AI development and deployment processes. The section presents a flowchart outlining the critical stages in AI projects and the influence of various international frameworks. This section examines the role of policies in ensuring responsible AI development and the importance of incorporating international frameworks to achieve this goal.

Section 4, provides a comparative analysis of AI Ethics Policy Frameworks from different nations, using a Cartesian graph for evaluation based on transparency, accountability, equity, and privacy. This seciton highlights the varying approaches different countries take to AI ethics policies and how they compare.

Section 5, proposes a range of strategies for addressing and reducing bias in AI systems. It emphasizes the significance of data diversity, rigorous audits, ethical training, and algorithmic clarity in reducing bias in AI systems.

Finally, the paper concludes with section 6, which summarizes the key findings, discusses their implications for the future trajectory of AI development, and suggests avenues for further scholarly inquiry. The paper provides a comprehensive, multifaceted examination of AI ethics through this structured approach, contributing substantive insights to the ongoing scholarly dialog in this critically pivotal domain.

Ethical Considerations in AI Development

Ethical considerations are of the utmost importance in the development of AI to ensure that these systems are safe, fair, and transparent. A Venn diagram illustrates the interplay between three key aspects: transparency, fairness, and privacy.

Transparency, Explainability, and Clarity refer to making AI systems clear and understandable to users. Fairnessrequires that AI systems be designed equitably and without biases. Privacy, or "Data Protection and Consent," focuses on protecting personal data and obtaining informed consent for its usage.

These aspects are interconnected, and the intersections of the Venn diagram illustrate how they overlap. For example, the intersection of transparency and fairness is called accountability, which emphasizes the importance of transparent and fair AI decision-making. The intersection of transparency and privacy, called user trust, emphasizes the need for transparency in data use and protection (Aldoseri, Al-Khalifa, and Hamouda 2023; Bécue, Praça, and Gama 2021; Malhotra 2018; Mijwil, Aljanabi, and ChatGPT 2023). The intersection of fairness and privacy, known as nondiscriminatory data practices, highlights the need for privacy considerations to align with fairness to avoid discrimination.

The center intersection of the Venn diagram represents the ideal of responsible AI use that balances transparency, fairness, and privacy. A flowchart provides a clear, step-by-step guide to embed ethical considerations into AI development.

The process starts with a commitment to ethical AI development and defining ethical principles such as transparency, fairness, and privacy. Data use and AI training guidelines should then be implemented to adhere to these principles. Regular audits should be conducted to identify and correct biases and ensure compliance with ethical standards.

Clear lines of responsibility and accountability should be established in AI-driven decisions. Continuous improvement based on feedback from users and stakeholders should be a regular practice. The goal is the realization of AI systems that fully embody ethical principles and ensure the safety and well-being of all users.

The framework from Figure 1 is discussed in more detail in the next section.

Transparency, Explainability, and Clarity

In developing artificial intelligence, it is essential to prioritize transparency, explainability, and clarity to ensure ethical development and deployment. Transparency refers to the accessibility of AI systems and their workings to users and stakeholders. Explainability, closely linked to transparency, pertains to the ability of AI systems to be understood and interpreted by human beings, ideally in non-technical language. Meanwhile, clarity ensures that AI systems' purposes and outcomes are communicated in a straightforward and understandable manner.

The importance of these elements cannot be overstated. They are crucial in building and maintaining user trust, ensuring that AI systems operate

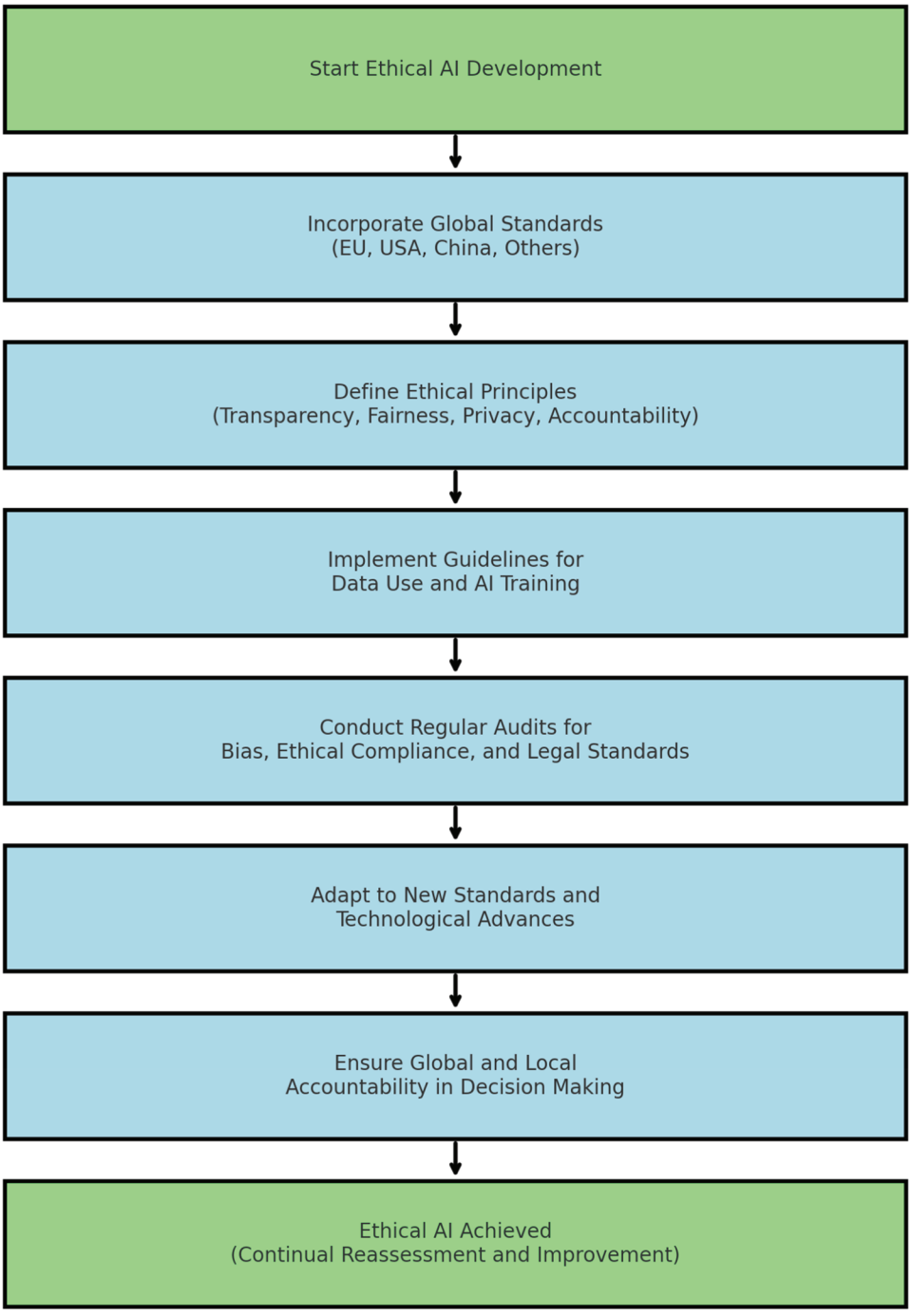

Figure 1. Transparency, Explainability, and Clarity: A framework for developing AI systems that embody ethical principles and ensure the safety and well-being of all users.

understandably and predictably. Furthermore, transparency and explainability play a pivotal role in establishing accountability, ensuring that AI developers and users are held responsible for the outcomes of AI systems (European Parliament 2023; ISO 2023; McCorduck and Cfe 2004; MeitY 2023; NIST 2023b; Office for Artificial Intelligence 2023).

Fairness and Bias Prevention

Ensuring fairness in AI systems requires creating programs that make decisions without prejudice or partiality (Bender et al. 2021). This necessitates a conscientious effort to design AI systems that do not perpetuate existing biases or create new ones (Shu, Zhang, and Yu 2021). However, achieving fairness in AI poses significant challenges, as these systems often learn from real-world data, which can be inherently biased.

The intersection of fairness, privacy, and accountability is a complex but essential consideration. Ensuring fairness often involves careful handling of sensitive data while also maintaining transparency and accountability in decision-making processes. This balancing act is critical in mitigating biases and ensuring that AI systems are equitable and just.

Privacy and Data Protection

Privacy and data protection are critical ethical considerations that must be considered during AI development. It involves protecting personal and sensitive information from unauthorized access and ensuring that data is used responsibly. Regulations and standards, such as the General Data Protection Regulation (GDPR) (GDPR 2018; ICO 2018) in the European Union, play a significant role in shaping AI ethics by setting strict guidelines for data use. Privacy, fairness, and user trust are closely linked. Protecting privacy is crucial in building and maintaining user trust, which is essential for the acceptance and success of AI systems. Furthermore, ensuring the proper handling of data is vital for fairness, as data misuse can lead to biased outcomes.

Interconnectedness of Ethical Aspects

The ethical dimensions of AI, including transparency, fairness, and privacy, are not isolated but deeply interconnected (Partnership on AI 2023; Roberts et al. 2021). We can visualize this interconnectedness using a Venn diagram that shows how these aspects overlap and influence each other. For instance, when transparency and fairness intersect, it leads to accountability. Similarly, when fairness and privacy overlap, it underscores the need for nondiscriminatory data practices. The idea of responsible AI use is represented by the central intersection of these aspects in the Venn diagram. This is where all

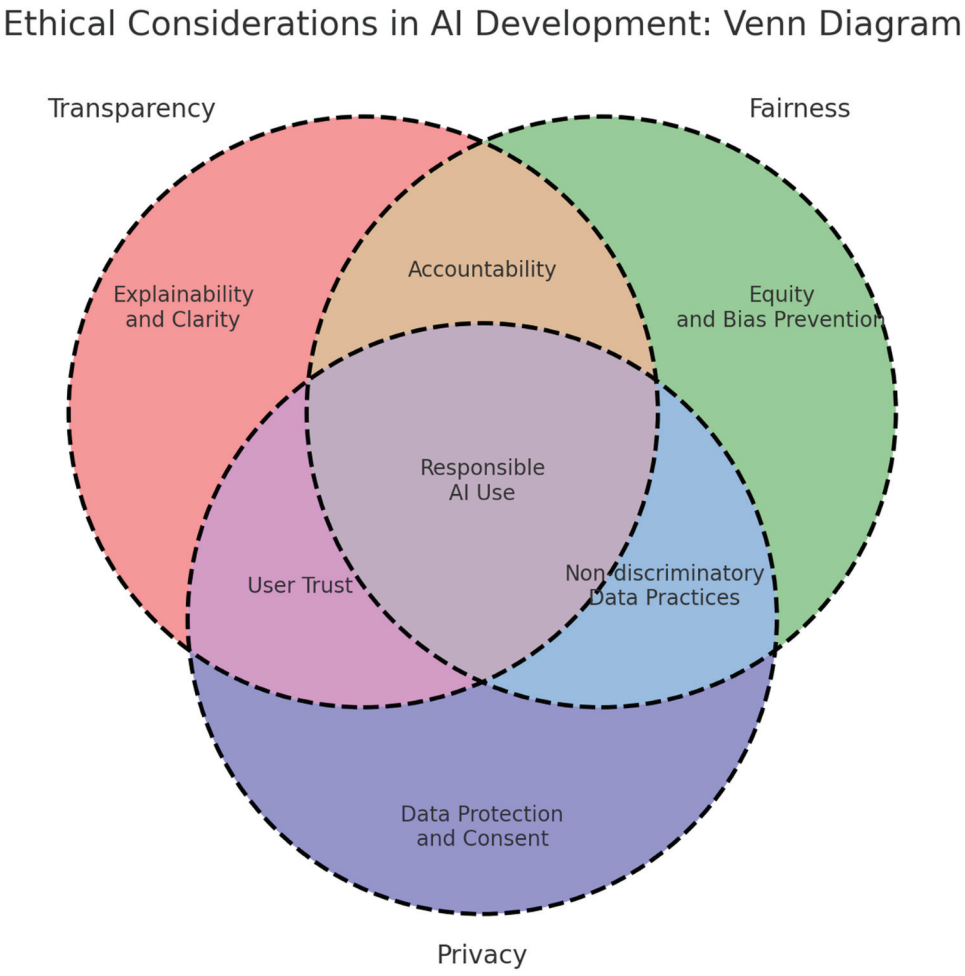

Figure 2. Interconnected concepts of the Framework for developing AI systems that embody Transparency, Explainability and Clarity.

three principles are balanced, leading to AI systems that are ethical, reliable, and trustworthy. We can visualize this in Figure 2.

Implementation Strategies

Incorporating ethical considerations into the development of AI requires a systematic approach. A step-by-step guide for doing this involves first committing to ethical principles. Next, data use and AI training guidelines should be implemented to align with these principles. Regular audits are necessary to detect and correct biases to ensure compliance with ethical standards.

Establishing clear lines of responsibility and accountability in AI-driven decisions is also crucial. Continuous improvement, based on feedback from

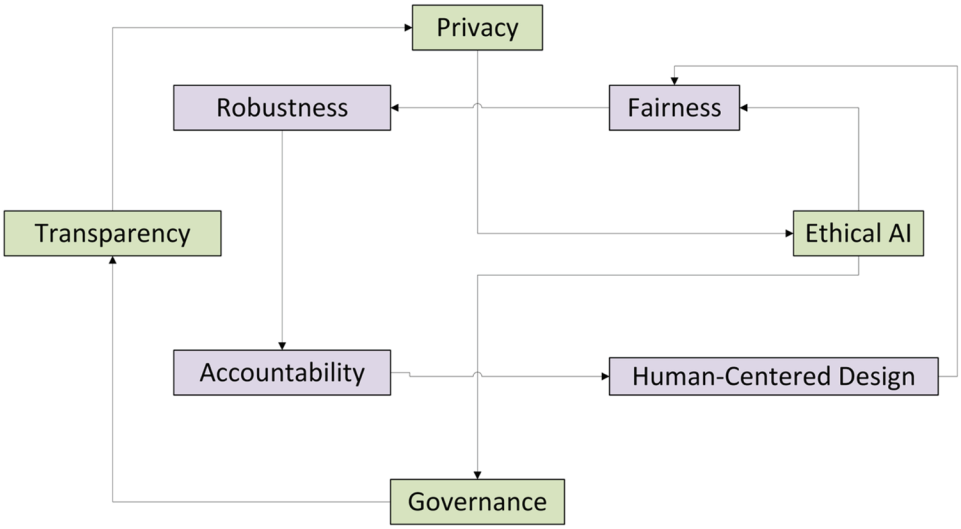

Figure 3. Key elements found in global AI frameworks.

users and stakeholders, should be an integral part of AI development. The ultimate goal is to create AI systems that embody ethical principles, ensuring all users' safety and well-being.

Responsible AI Frameworks

This section will explore AI frameworks from the European Union, the United States, and China. Using a detailed flowchart, we will examine how these frameworks impact each stage of an AI project, from initiation to post-deployment. Additionally, we will use a Venn diagram analysis to compare the EU AI Act (Bommasani et al. 2023), US AI Principles (Tabassi 2023), and China AI Ethics guidelines (Roberts et al. 2021), highlighting their unique features and areas of overlap. This approach will demonstrate how these frameworks share a commitment to ethical standards, privacy protection, and fairness while also providing distinct perspectives on AI development and governance. This analysis is crucial for understanding the multifaceted nature of global AI frameworks and their implications for responsible AI practices.

While this paper primarily focuses on theoretical frameworks and policy analysis, it is important to acknowledge the growing need for empirical research to substantiate the claims made within the scope of AI ethics. In response, we introduce two key case studies that provide concrete examples of how AI ethics frameworks are applied in practice. These case studies were derived from in-depth analyses of recent AI implementations in both the

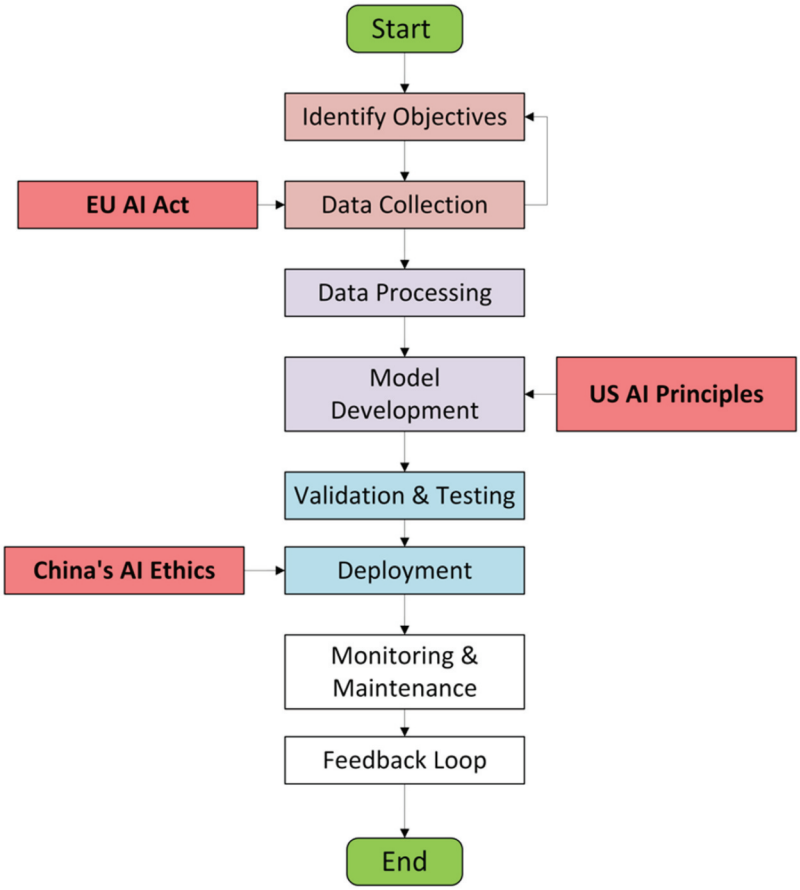

Figure 4. Applied design of responsible and ethical AI practices by integrating global AI frameworks into different stages of AI development and deployment.

healthcare and financial sectors, which serve as critical areas where ethical considerations are paramount.

The first case study examines the deployment of AI diagnostic systems in the European healthcare industry, particularly focusing on IBM Watson Health's use of AI in oncology diagnostics. By conducting structured interviews with healthcare practitioners and reviewing compliance reports, we observed how the stringent transparency and accountability requirements mandated by the European Union's AI Act (European Parliament 2023) influenced the AI system's design and operational transparency. This empirical evidence demonstrates how AI systems were modified to meet regulatory standards, particularly concerning the explainability of diagnostic recommendations provided to medical professionals and patients.

The second case study involves an empirical assessment of AI fraud detection systems in the financial services sector within the United States. Through direct engagement with industry experts and an analysis of

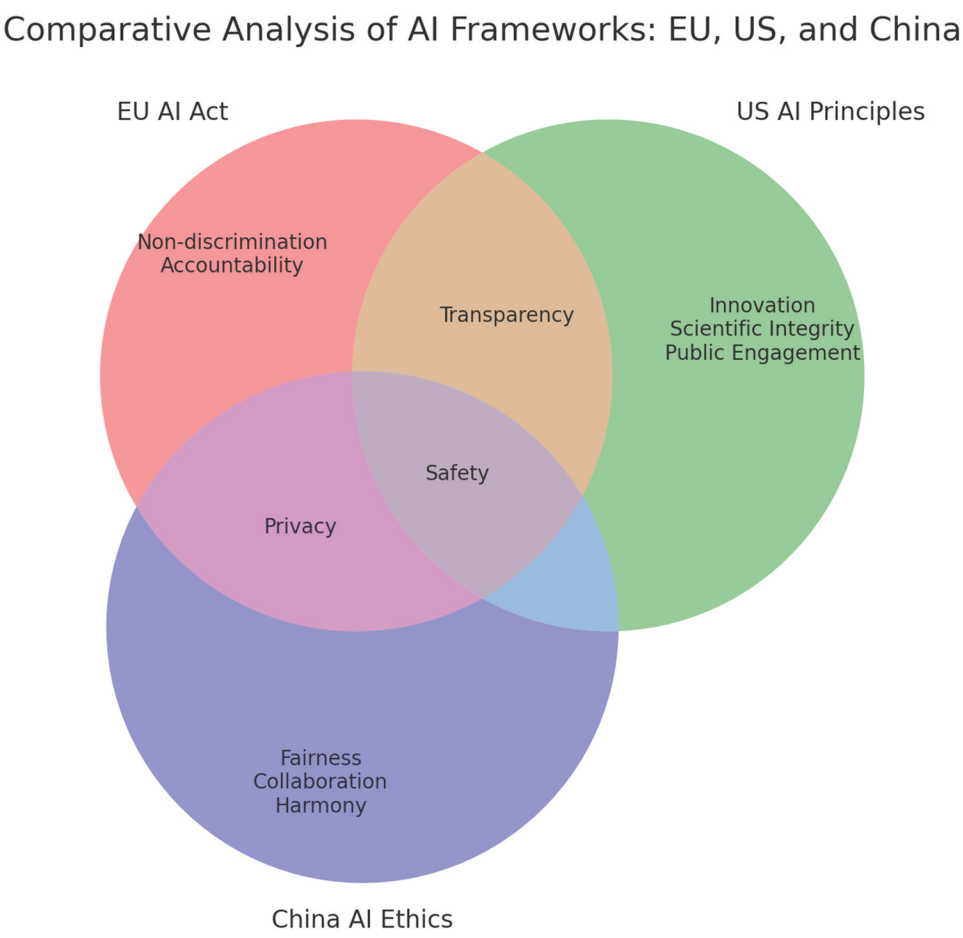

Figure 5. Different global AI frameworks overlap in some areas but maintain unique characteristics in others.

internal auditing processes, we explored how JP Morgan's AI-powered fraud detection system aligns with the flexible, innovation-driven guidelines of the NIST AI Risk Management Framework (NIST 2023a; Tabassi 2023). The study reveals how these frameworks permit adaptive risk management strategies, providing companies with the autonomy to tailor their ethical standards while maintaining a balance between innovation and accountability.

These case studies provide empirical evidence to support the theoretical and policy analysis discussed in this paper. They illustrate how global AI ethics frameworks are not just abstract concepts but operational guidelines that have tangible impacts on AI design, deployment, and compliance. By incorporating these real-world applications, we aim to bridge the gap between theory and practice, offering a more robust foundation for understanding the dynamics of AI governance across various industries.

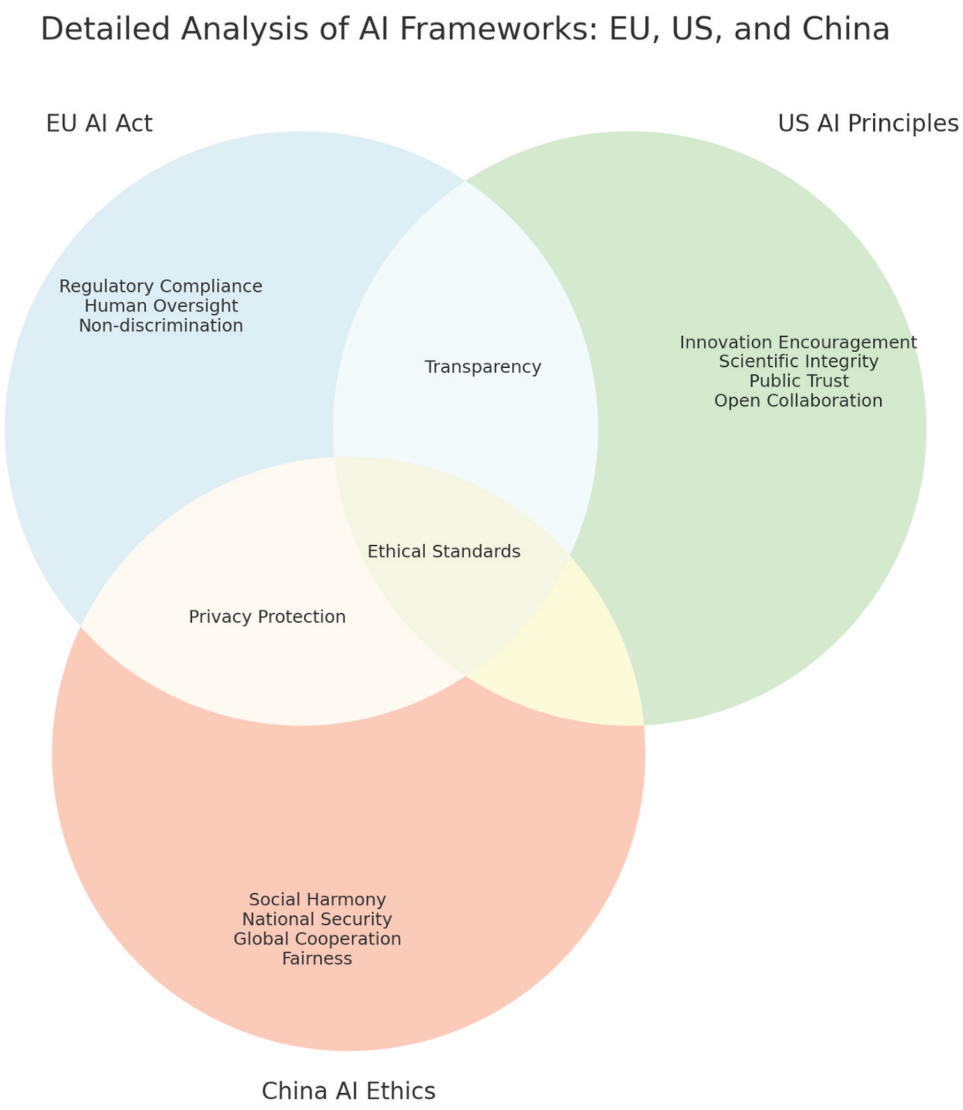

Figure 6. The complexity of AI frameworks across different regions and the potential for collaboration and divergence highlights the need for global collaboration and harmonisation of AI ethics and regulations.

Development and Deployment Flowchart

Creating and implementing AI systems is a complex process that requires compliance with international frameworks to ensure ethical and responsible outcomes. This section provides a comprehensive flowchart that integrates AI frameworks from the European Union, the United States, and China, mapping their application throughout the AI project lifecycle.

The flowchart is based on the EU AI Act (European Parliament 2023), US AI Principles (NIST 2024c), and China AI Ethics guidelines (Provisions on the

Administration of Deep Synthesis Internet Information Services, Personal Information Protection Law of the People's Republic of China PRC 2022). It begins with initiating an AI project, where objectives are defined and relevant data is collected. The EU AI Act's guidelines on ethical data use and transparency are essential at this stage. The US AI Guidelines come into play as the project progresses to data processing and AI model development, emphasizing innovation, fairness, and accountability.

During the validation and testing phase, the model must align with the set objectives and comply with ethical standards per these frameworks. The deployment phase sees the integration of China's AI Ethics guidelines, which prioritize social harmony, national security, and global cooperation. After deployment, continuous monitoring and maintenance are essential to ensure the AI system functions as intended and adheres to ethical standards. This phase is critical for incorporating feedback and insights, allowing for iterative improvements based on real-world performance and impact.

Figure 3 illustrates how global AI frameworks are necessary and applicable at different AI development and deployment stages. This integration ensures a holistic approach to responsible AI practices that align with global standards and ethical considerations.

Figure 3 shows the steps involved in developing and deploying AI systems. It incorporates the latest AI frameworks worldwide, including those from the EU, the US (NAIAC 2024, NIST 2024a), and China (Interim Measures for the Management of Generative Artificial Intelligence Services, Personal Information Protection Law of the People's Republic of China PRC 2023; Li 2017; Provisions on the Administration of Deep Synthesis Internet Information Services, Personal Information Protection Law of the People's Republic of China PRC 2022; The State Council People Republic of China 2017), and highlights critical points where these frameworks intersect. The process flow begins with the start of the AI project and moves on to identifying objectives and collecting relevant data. The EU AI Act's guidelines may be considered at this stage. The collected data is then processed, and the AI model is developed according to the principles outlined in the US AI Guidelines. The model is then validated and tested to meet the set objectives. Deployment of the AI system in a real-world environment is the next step, and considerations from China's AI Ethics guidelines come into play here. Continuously monitoring and maintaining the AI system post-deployment is essential. The flowchart also includes a feedback loop that involves revisiting objectives and processes based on feedback and new insights. Once all the steps are completed, the AI project cycle ends. The flowchart in Figure 4 ensures responsible and ethical AI practices by integrating global AI frameworks into different stages of AI development and deployment.

Figure 4 visualizes how various AI frameworks integrate with the AI development and deployment process. Each stage of the AI process is marked, from

the beginning to the end. The EU, the US, and China AI frameworks are highlighted with individual annotations. The arrows linking these frameworks to specific stages in the process are designed to avoid text overlap, ensuring clarity and readability.

The flowchart effectively illustrates the integration of global AI frameworks into the AI development lifecycle, emphasizing the importance of considering these guidelines at different stages for responsible AI practices. The annotations for the AI frameworks from the EU, US, and China are strategically positioned to avoid obstructing any other flowchart elements. The connecting arrows are designed with a specific arc to neatly link the frameworks to their respective stages in the AI process without crossing over any text.

The Venn diagram represents the AI frameworks from the EU, the US, and China. Each circle in the diagram represents a different region's AI framework: EU AI Act, US AI Principles, and China AI Ethics. The overlaps between the circles indicate areas of common focus or principles shared between these frameworks. Individual sections highlight unique aspects of each framework.

This Venn diagram in Figure 5 visually demonstrates how different global AI frameworks overlap in some areas while maintaining unique characteristics in others. It reflects the diverse approaches to AI governance and ethics across these regions.

The Venn diagram in Figure 5 comprehensively analyses the AI frameworks from the European Union, the United States, and China. It highlights the areas of collaboration and divergence between the three regions and demonstrates the complexity of AI frameworks across different regions.

The European Union's AI Act focuses on human oversight, nondiscrimination, and regulatory compliance. The US AI Principles emphasize innovation encouragement, public trust, and open collaboration. China's AI Ethics prioritizes social harmony, national security, and global cooperation. These unique elements reflect the individual priorities and cultural perspectives of each region.

The EU and the US share values in transparency and ethical standards. The EU and China both emphasize privacy protection and ethical standards. The US and China find common ground in ethical standards and a focus on fairness. These common areas between the two frameworks indicate the potential for global collaboration and harmonization of AI ethics and regulations.

The central overlap in the Venn diagram highlights the areas where the EU, US, and China share common principles such as ethical standards and privacy protection. These areas offer opportunities for global collaboration and harmonization of AI ethics and regulations.

Each region has unique focus areas that reflect its priorities and cultural perspectives. These divergent approaches could lead to different approaches in

AI development and governance. The Venn diagram demonstrates the potential for collaboration and divergence in the global AI landscape.

The Venn diagram analysis in Figure 6 shows the complexity of AI frameworks across different regions and the potential for collaboration and divergence. It highlights the need for global collaboration and harmonization of AI ethics and regulations to ensure that AI development and governance align with ethical and societal values.

The Venn diagram in Figure 6 shows the EU AI Act i in light blue, the US AI Principles are in light green, and the China AI Ethics framework in coral.

Real-World Applications of Global AI Frameworks

In order to provide a clearer understanding of how AI frameworks function in practice, it is essential to examine real-world applications of these frameworks across different regions. A relevant example can be found in the application of the European Union's AI Act within the healthcare sector. The use of AI-powered diagnostic tools, such as IBM's Watson Health, was subject to scrutiny under the EU's stringent regulations on transparency and explainability. Under the AI Act, companies deploying such AI systems in high-risk sectors are required to provide detailed documentation of the algorithms used, as well as explainability mechanisms that allow medical professionals and patients to understand AI-driven decisions. This regulatory requirement has led to the modification of AI models to ensure compliance, particularly by providing more transparent decision-making processes that can be audited by healthcare regulators.

In contrast, the United States' more innovation-centric approach, as embodied by the NIST AI Risk Management Framework, can be observed in the deployment of AI in the financial services sector. For instance, JP Morgan's AI-powered fraud detection system operates within a framework that emphasizes risk mitigation through best practices and industry standards rather than rigid regulatory oversight. The NIST framework encourages companies to develop internal policies tailored to their operational risks, allowing for greater flexibility in AI implementation. As a result, JP Morgan has developed proprietary methods for continuous monitoring and auditing of AI models to ensure they remain effective while balancing the need for innovation with ethical considerations.

China's AI governance, which prioritizes state control and societal harmony, can be seen in the government's use of facial recognition systems for public security. The deployment of such systems, governed by China's Provisions on the Administration of Deep Synthesis Internet Information Services, Personal Information Protection Law of the People's Republic of China (PRC) (2022), illustrates how the state leverages AI under a framework that prioritizes national security. In this

case, AI technologies are used to monitor public spaces, but their ethical implications, particularly in terms of privacy, are handled within a governance model that differs significantly from those in Western democracies. The Chinese government's emphasis on AI as a tool for societal stability underscores the unique application of their framework in practice.

These examples highlight the varied approaches of AI frameworks across different regions and sectors, demonstrating how global AI policies are shaped by contextual factors and applied in practical, high-impact scenarios. By examining such real-world cases, we can better understand the strengths and limitations of these frameworks and the challenges in harmonizing AI ethics on a global scale.

Specific Recommendations for Real-World Applications of Global AI Frameworks Include

Technical Solutions for Embedding Ethical Principles in AI Systems. Embedding ethical principles such as fairness, transparency, and accountability into AI systems requires sophisticated algorithmic approaches that ensure these objectives are met without compromising the system's performance. **Algorithmic fairness** can be addressed using methods such as *differential fairness* and *fair representation learning*. For instance, algorithms like the **Fair Representation Learning (FRL)** model aim to mitigate bias by transforming raw data into a latent representation that is invariant to sensitive attributes, such as race or gender, without losing important predictive power. The FRL method applies adversarial learning to ensure the model cannot easily infer sensitive attributes, thus reducing bias while maintaining accuracy. This can be particularly useful in sectors like finance, where historical biases in credit scoring datasets often lead to unfair outcomes. Incorporating these fairness constraints during the model training phase ensures that discriminatory patterns in the data are not propagated by the AI system.

Transparency is enhanced through the use of **explainable AI (XAI)** techniques. One common approach is the implementation of **Local Interpretable Model-agnostic Explanations (LIME)**, which provides users with interpretable approximations of complex models, enabling end-users and auditors to understand and evaluate individual predictions. LIME works by perturbing input data and observing how changes impact predictions, thus constructing simpler, interpretable models locally around specific instances. This method is particularly valuable in high-stakes fields like healthcare, where understanding the rationale behind AI-driven diagnoses is critical for building trust and accountability. For example, LIME has been effectively applied in medical imaging to explain how AI systems identify tumor regions, offering transparency to both clinicians and patients.

Obtaining Real-World Probabilistic Data for Legislation. To create more robust AI legislation that addresses real-world challenges, **probabilistic data collection** is necessary. A critical solution lies in **data-driven simulations** that use real-world probabilistic distributions of AI outcomes across various domains. These simulations can leverage **Bayesian inference models** to analyze the probability of ethical failures, such as biased decisions or transparency breaches, under different regulatory scenarios. For example, Bayesian models can assess the likelihood of biased outputs in loan approval systems based on varying regulatory constraints, allowing policymakers to quantitatively evaluate the trade-offs between stringent regulation and innovation. By incorporating such probabilistic assessments, policymakers can develop legislation grounded in empirical evidence, ensuring that ethical guidelines are both practical and enforceable in diverse sectors.

Moreover, **quantitative data collection** from real-world AI deployments could utilize techniques such as **differential privacy** to protect sensitive information while still gathering meaningful insights. For instance, in healthcare, collecting large-scale patient data from AI diagnostic tools while maintaining patient privacy can be achieved through differential privacy algorithms that introduce noise into datasets, ensuring that individual records cannot be re-identified. This allows regulators to gather accurate statistics on AI system performance, such as prediction accuracy and error rates, without violating privacy laws. These real-world data points can then be used to fine-tune legislative frameworks to ensure they are reflective of practical AI use and compliant with privacy standards.

Algorithmic Solutions to Ensure Fair and Ethical AI for End-Users. To ensure AI systems are perceived as fair and ethical by end-users, several algorithmic approaches can be integrated into the development lifecycle. One promising method is the use of **fairness constraints** in model optimization, such as **Equalised Odds** and **Demographic Parity**. The Equalised Odds algorithm ensures that an AI system has equal true positive and false positive rates across different demographic groups, ensuring that no group disproportionately benefits or suffers from the system's decisions. This technique has been successfully implemented in judicial systems where AI models are used for bail and sentencing recommendations, reducing the racial disparities commonly observed in earlier models.

Fairness-aware learning algorithms can also be embedded into machine learning pipelines to monitor and adjust for bias during the training process. For example, the **Fairness through Awareness (FTA)** framework adjusts decision boundaries within models to ensure that similar individuals are treated similarly, thereby reducing unfair bias. This algorithm calculates distances in a fairness-sensitive space and ensures that individuals who are close in this space receive similar predictions. This has been applied in hiring

algorithms to ensure that applicants with similar qualifications, regardless of demographic attributes, are treated equitably.

Furthermore, end-user engagement with AI systems can be improved through **interactive transparency mechanisms**. For instance, **counterfactual explanations** can be used to provide users with actionable insights into how decisions could change if certain inputs were modified. In credit scoring systems, for example, a counterfactual explanation might inform a user that their loan was denied due to a low credit score and suggest specific steps, such as reducing credit card debt, that would lead to approval. By providing users with clear, actionable insights, these systems not only increase trust but also empower users to engage more meaningfully with AI-driven decisions.

Algorithmic Accountability and Continuous Monitoring. To maintain ongoing fairness and ethical standards, continuous monitoring of AI systems is essential. This can be achieved through **algorithmic auditing frameworks** that regularly assess AI systems for adherence to ethical principles post-deployment. **Post-hoc fairness auditing tools**, such as **AI Fairness 360 (AIF360)**, provide an open-source toolkit that measures and mitigates bias in deployed models. These tools can be integrated into AI governance processes, ensuring that models remain fair and unbiased as they encounter new data in real-world environments. AIF360 evaluates fairness through multiple metrics, such as disparate impact and statistical parity, and enables continuous recalibration of models to maintain ethical performance.

Incorporating **algorithmic accountability systems** with real-time feedback loops ensures that biases introduced by shifts in data distributions (data drift) are swiftly detected and mitigated. Techniques such as **drift detection algorithms**, including **ADWIN** (Adaptive Windowing), continuously monitor the performance of AI models and trigger retraining when significant deviations from expected behavior are detected. By automating the detection of ethical breaches and recalibrating models in response, these systems ensure that AI remains both effective and ethically compliant over time.

Comparative Venn Diagram Analysis

This section uses a Venn diagram analysis to compare the AI frameworks of the European Union (EU), the United States (US), and China. The diagram represents each framework, highlighting their unique features and areas of overlap, revealing both collaborative potentials and divergent approaches.

The EU's AI Act prioritizes human oversight, nondiscrimination, and strict regulatory compliance. It reflects the EU's emphasis on protecting citizens' rights in the digital age. The US AI Principles prioritize fostering innovation, ensuring public trust, and promoting open collaboration. This mirrors the US's emphasis on market-driven and innovation-led AI development. On the

other hand, China's AI Ethics framework emphasizes the importance of social harmony, national security, and global cooperation. It reflects China's approach to balancing technological advancement with social stability and state security.

The intersections of these frameworks in the Venn diagram highlight shared principles and potential areas for international cooperation. For example, the EU and the US emphasize ethical standards and transparency. The EU and China share a common focus on privacy protection and the ethical use of AI. The US and China converge on encouraging ethical standards and fairness in AI.

At the central intersection of the Venn diagram, where all three frameworks overlap, lies a shared commitment to ethical standards, privacy protection, and ensuring fairness. This common ground suggests opportunities for global collaboration and the harmonization of AI ethics and regulations.

However, each framework's divergent aspects reflect each region's varying priorities and cultural perspectives. These differences could lead to distinct approaches in AI development and governance globally. Therefore, the Venn diagram highlights the potential for collaboration and underscores the need to understand and respect diverse perspectives in the global AI landscape.

Policy Frameworks

This section expands into a thorough analysis of AI Ethics Policy Frameworks worldwide. It covers significant regions such as the European Union, the United States, China, Canada, Japan, India, and Australia. Figure 7 presents the data in Cartesian graphs, comparing these frameworks across four key ethical dimensions: Transparency, Accountability, Fairness, and Privacy. Each framework is evaluated in detail and rated, providing an insightful understanding of how countries prioritize these dimensions in their AI policies. This graph highlights each framework's unique priorities and focus areas and emphasizes the diversity and commonalities in global approaches to AI ethics. Policymakers can leverage these insights to identify areas for improvement and develop comprehensive, ethically aligned AI policies. Moreover, the section expands upon this analysis to include additional critical dimensions such as human oversight and national security, broadening the scope to encompass broader sociopolitical implications of AI technology. Figure 8 analyses global trends and presents a roadmap for harmonization in AI ethics, advocating for ongoing international dialogue and cooperation to foster a responsible and ethical global AI ecosystem.

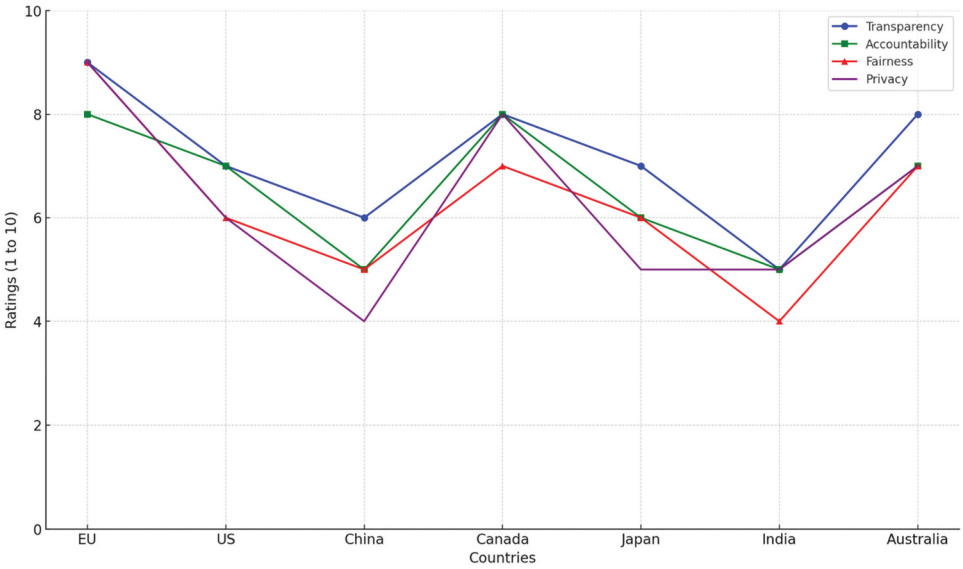

Figure 7. Cartesian graph - Comparison of AI Ethics Policy Frameworks Across Countries.

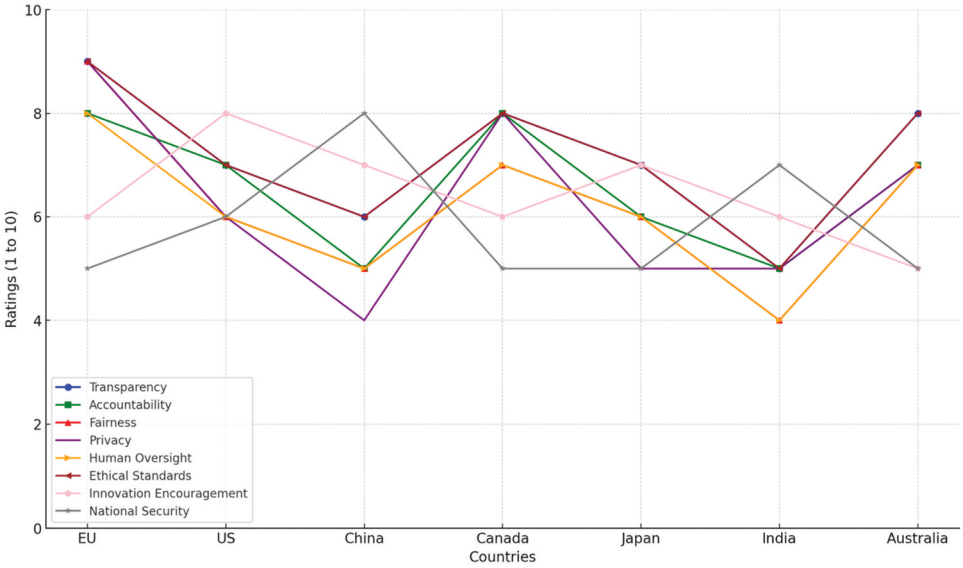

Figure 8. Cartesian graph of Global AI Ethics Policy Frameworks: a tool for policymakers to understand the priorities of different countries regarding AI ethics.

Criteria for Rating AI Ethics Policy Frameworks Across Countries

The ratings for each country's AI ethics policy framework were based on a detailed analysis of public documents, policy white papers, regulatory guidelines, and academic literature. The scores for each dimension, transparency, accountability, fairness, privacy, human oversight, ethical standards,

innovation encouragement, and national security, were evaluated according to the following specific criteria:

Transparency (1–10 Scale)

Transparency was assessed by the extent to which each country's framework mandates openness regarding the design, implementation, and decision-making processes of AI systems.

- **9-10**: Countries with explicit, legally enforced transparency requirements for AI systems, including mandates for explainability and public accountability (e.g., the European Union).
- **6-8**: Countries that encourage transparency but do not mandate it as a legal requirement across all AI applications (e.g., the United States).
- **4-5**: Countries where transparency is mentioned in policies but with few practical enforcement mechanisms (e.g., China).
- **1-3**: Minimal or no formal focus on transparency in the AI framework.

Accountability (1–10 Scale)

Accountability measures the robustness of legal and regulatory mechanisms that hold developers, companies, and governments responsible for the outcomes of AI systems.

- **9-10**: Countries with well-defined liability frameworks that assign clear responsibility to AI developers or operators (e.g., the EU AI Act).
- **6-8**: Countries where accountability is encouraged through voluntary compliance frameworks but lacks mandatory enforcement (e.g., the US with the NIST AI Risk Management Framework).
- **4-5**: Countries with vague accountability measures, often handled at the discretion of private entities or lacking centralized regulation (e.g., Japan).
- **1-3**: Countries where accountability frameworks are non-existent or still in early development phases.

Fairness (1–10 Scale)

Fairness was evaluated based on how well a country's policy framework addresses bias in AI algorithms and ensures equitable outcomes across demographic groups.

- **9-10**: Countries with explicit fairness requirements for AI systems, mandating fairness audits and bias mitigation techniques (e.g., Canada, EU).
- **6-8**: Countries that encourage fairness but with less stringent or optional auditing practices (e.g., the US).
- **4-5**: Countries where fairness is an aspirational goal with limited practical implementation (e.g., India).

- **1-3**: Minimal or no focus on fairness in AI regulation.

Privacy (1–10 Scale)

Privacy was assessed by how each country's framework protects user data in the context of AI and how it aligns with global standards like the GDPR.

- **9-10**: Countries with robust privacy regulations, including explicit rules on AI data use (e.g., EU, Canada).
- **6-8**: Countries with general data privacy laws but limited AI-specific guidelines (e.g., the US, Japan).
- **4-5**: Countries with privacy regulations that are inconsistently applied or underdeveloped in relation to AI (e.g., India, Australia).
- **1-3**: Countries with minimal focus on privacy in AI contexts, or where data protection laws are not enforced effectively (e.g., China).

Human Oversight (1–10 Scale)

This criterion assessed the role of human oversight in AI decision-making, particularly in high-risk sectors like healthcare or autonomous vehicles.

- **9-10**: Countries mandating human oversight in high-risk AI decisions, ensuring human intervention in critical areas (e.g., EU AI Act).
- **6-8**: Countries that recommend but do not legally enforce human oversight (e.g., the US).
- **4-5**: Countries where human oversight is mentioned, but enforcement mechanisms are vague or absent (e.g., Japan, India).
- **1-3**: Little to no emphasis on human oversight in AI policy frameworks.

Ethical Standards (1–10 Scale)

Ethical standards were scored based on how well a country's AI policies adhere to global ethical frameworks (such as UNESCO's AI ethics recommendations) and promote ethical AI development.

- **9-10**: Countries with a clearly defined, internationally aligned ethical framework for AI (e.g., the EU, Canada).
- **6-8**: Countries with ethical guidelines for AI but limited in scope or enforcement (e.g., Japan, the US).
- **4-5**: Countries that mention ethical AI but lack a coherent, enforceable framework (e.g., India).
- **1-3**: Minimal or no formal focus on AI ethics in public policy.

Innovation Encouragement (1–10 Scale)

This criterion measured the balance between promoting AI innovation and enforcing ethical guidelines. Countries that foster innovation while maintaining a robust ethical framework scored higher.

- **9-10**: Countries with innovation-centric policies that support AI research and development while integrating ethical guidelines (e.g., US, Canada).
- **6-8**: Countries with strong innovation policies but less rigorous ethical enforcement (e.g., Japan).
- **4-5**: Countries where innovation is promoted but at the cost of ethical standards (e.g., China).
- **1-3**: Countries where innovation in AI is stifled due to excessive regulation or a lack of resources (e.g., minimal focus).

National Security (1–10 Scale)

National security was evaluated based on how countries incorporate AI within their national security strategies, including defense, cybersecurity, and surveillance.

- **9-10**: Countries where AI plays a significant role in national security frameworks, with clear policies on military AI, surveillance, and cyber defense (e.g., China, the US).
- **6-8**: Countries that include AI in national security policies but with fewer explicit regulations on its use in defense (e.g., Australia, Japan).
- **4-5**: Countries with some mention of AI in national security contexts but lacking concrete policies (e.g., India).
- **1-3**: Minimal focus on AI for national security, or policies that are still in early development (e.g., the EU).

Justification for the Selection of Criteria and Scores

The chosen criteria for evaluating AI ethics policy frameworks, transparency, accountability, fairness, privacy, human oversight, ethical standards, innovation encouragement, and national security, were carefully selected to reflect the core dimensions that are essential for ethically sound, socially beneficial, and technologically responsible AI systems. These dimensions are well-established in policy discourse and academic literature as the pillars of AI ethics and governance, ensuring that AI development aligns with societal values and mitigates potential harms.

Transparency

Transparency is a cornerstone of AI ethics, as highlighted in both academic and regulatory discussions (Floridi et al. 2018). Transparent AI systems allow

stakeholders to understand how decisions are made and ensure accountability. The choice of transparency as a criterion is supported by regulatory frameworks such as the European Union's **GDPR** and **AI Act**, which place explicit demands on AI systems to be explainable and open to public scrutiny. Studies have shown that lack of transparency is one of the main causes of public distrust in AI systems (Wachter, Mittelstadt, and Russell 2023). Therefore, countries with clear legal mandates for transparency received higher scores, while those with voluntary or vague transparency guidelines scored lower.

Accountability

Accountability ensures that AI developers, operators, and users are held responsible for the outcomes produced by AI systems. This criterion is justified by the recognition in the literature that without clear accountability structures, it becomes difficult to address failures or harms caused by AI systems (Mittelstadt 2019). The **EU AI Act** introduces comprehensive provisions that assign legal responsibility, providing a strong model for accountability. Countries like the United States, with voluntary compliance through frameworks like the **NIST AI Risk Management Framework**, received intermediate scores due to the lack of enforceability. The necessity of accountability is also a major theme in academic literature, particularly in the context of complex AI systems where multiple stakeholders are involved in the design and deployment (de Bruin and Floridi 2017; Floridi et al. 2018; Turilli and Floridi 2009).

Fairness

Fairness in AI systems addresses concerns about bias and discrimination, which are well-documented issues in AI applications (Binns 2018). Countries with explicit fairness requirements in their AI policies, such as the European Union and Canada, received higher scores because their frameworks mandate fairness audits and bias mitigation practices. Literature on fairness in AI often points to the limitations of algorithmic systems to ensure equitable outcomes across demographic groups without specific regulatory intervention (Du 2023; IBM 2018). Nations with minimal or non-enforceable fairness provisions, such as India and China, scored lower due to the absence of robust bias-mitigation mechanisms.

Privacy

Privacy is a critical concern in AI, especially in systems that rely on vast amounts of personal data. The **GDPR** in the EU sets a high global benchmark for data protection and privacy, justifying the high score for the EU in this dimension. In contrast, countries like the United States, where privacy regulations such as **HIPAA** are domain-specific and not universally applicable to AI systems, scored lower (HIPAA 1996). Privacy as a criterion is grounded in the

principle that ethical AI must protect individuals' rights to control their data, a concern that is pervasive in academic and policy literature (Mittelstadt 2019).

Human Oversight

Human oversight in AI decision-making is essential to prevent over-reliance on automated systems, particularly in critical sectors like healthcare and law enforcement (European Parliament 2023). The **EU AI Act** again leads the way by mandating human oversight in high-risk AI applications. The literature emphasizes the importance of preserving human judgment in AI-assisted decision-making, especially in cases involving moral or legal consequences (Jobin, Ienca, and Vayena 2019). Countries that recommend but do not mandate human oversight scored lower, as voluntary oversight often fails in real-world applications, particularly where operational efficiency is prioritized over human intervention.

Ethical Standards

Ethical standards are increasingly seen as vital for aligning AI development with human values. UNESCO's **Recommendation on the Ethics of AI** and similar initiatives by the **OECD** have provided blueprints for ethical AI, focusing on principles such as beneficence, non-maleficence, autonomy, and justice (UNESCO 2023). Countries like Canada and the EU, which have adopted comprehensive ethical guidelines, scored highly. These standards are crucial for ensuring that AI operates within moral and legal boundaries. In contrast, countries that lack specific ethical frameworks for AI, such as India and China, received lower scores, reflecting the underdevelopment of ethical considerations in their AI policies.

Innovation Encouragement

The balance between encouraging AI innovation and enforcing ethical standards is a key concern for policymakers (Brynjolfsson and Mcafee 2014; Evans 2015). Countries like the United States scored high in this dimension due to their innovation-centric policies, such as the **NIST AI Risk Management Framework**, which promotes industry-led solutions and fosters a favorable environment for AI research and development. The literature supports the notion that innovation thrives when there is flexibility and minimal regulatory overhead, but with the caveat that ethical guardrails must not be neglected (Bostrom and Yudkowsky 2014). Countries that focus excessively on regulation, potentially stifling innovation, or that lack sufficient incentives for AI research, scored lower.

National Security

National security considerations, particularly regarding the development of **autonomous weapons systems (AWS)** and **AI-enhanced cybersecurity**, are

becoming a critical component of AI policy (Singer 2009). Countries like the US and China, where AI plays a substantial role in national defense strategies, scored highly. The literature on **autonomous systems** highlights the importance of regulating AI to prevent unintended consequences in military applications (Singer 2009). Countries that have not yet integrated AI into their national security strategies or have underdeveloped AI governance in this area, such as the EU, scored lower.

Justification for Scoring. The scores for each dimension were derived from a combination of the following sources:

- **Policy Documents and Regulations**: Key regulatory frameworks such as the **EU AI Act** (Bommasani et al. 2023; European Parliament 2023, FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety | The White House, 2023; Mozumder et al. 2022), **GDPR** (GDPR 2018; ICO 2018), **NIST AI Risk Management Framework** (NIST 2024a, 2024c; Tabassi 2023), and national AI strategies were directly analyzed to assess the strength and comprehensiveness of each country's AI governance mechanisms.
- **Academic Literature**: Foundational texts on AI ethics, fairness, transparency, and accountability were referenced to establish baseline expectations for what constitutes best practices in each dimension (Binns 2018; Floridi et al. 2018; Jobin, Ienca, and Vayena 2019).
- **Real-World Case Studies**: Examples of AI implementation in various sectors, including healthcare, finance, and national security, were examined to contextualize the practical impacts of each policy framework.

The criteria were selected to ensure a comprehensive assessment of each country's approach to AI ethics, focusing on both regulatory stringency and the practical application of ethical principles. Each score reflects the extent to which the country's framework addresses key challenges associated with AI governance, ensuring a balanced evaluation that is informed by both policy analysis and academic insights.

Comparative Analysis Using Cartesian Graphs

This section provides a detailed comparative analysis of AI Ethics Policy Frameworks from a global perspective. The analysis covers major regions such as the European Union, the United States, China, Canada, Japan, India, and Australia. The study examines these frameworks against key ethical dimensions, including Transparency, Accountability, Fairness, and Privacy, using a series of Cartesian graphs.

Each framework is rated on a scale of 1 to 10 across these dimensions. The Cartesian graph format helps to visualize how each country's framework measures up in these critical areas. For instance, the EU's framework prioritizes privacy and accountability, reflected in its high scores in these areas. On the other hand, the US framework might score higher on transparency because of its focus on open data and innovation. China's framework, emphasizing social harmony and national security, might have different strengths and weaknesses.

This comparative analysis provides insights into the priorities and focus areas of different countries and highlights the diversity and commonalities in approaches to AI ethics globally. The graphical representation aids in understanding the complexities of each framework, offering a clear view of how nations are navigating the ethical landscape of AI development.

The Cartesian graph in Figure 7 compares AI Ethics Policy Frameworks across countries, such as the EU, the US, China, Canada, Japan, India, and Australia. It evaluates them on Transparency, Accountability, Fairness, and Privacy.

Figure 7 presents an overview of how countries prioritize various elements of AI ethics in their policy frameworks. The four aspects used for rating are transparency, accountability, fairness, and privacy. Each aspect is rated on a scale from 1 to 10.

The Blue Line represents transparency, which reflects how policies are openly communicated and implemented. The Green Line indicates accountability, measuring the extent to which AI developers and users are held accountable for their systems as per the frameworks. The Red Line represents fairness, measuring the extent to which the policies ensure fair and unbiased AI systems. Lastly, the Purple Line shows the importance of user privacy and data protection in the policies.

The graph offers valuable insights into the global landscape of AI ethics. It highlights similarities and differences in national approaches to regulating AI that can help inform future policy development. The ratings of each aspect for each country can help policymakers identify areas for improvement in their policies.

This graph provides a tool for policymakers to understand the priorities of different countries regarding AI ethics. By focusing on transparency, accountability, fairness, and privacy, policymakers can create policies that address the concerns of stakeholders and help establish ethical guidelines for AI development and use.

Extended Analysis and Global Implications

In this section, we expand into an evaluation matrix that incorporates additional dimensions that are increasingly relevant to AI ethics. These dimensions

include the roles of human oversight and national security, reflecting AI technology's broader socio-political implications.

Including human oversight highlights the necessity for human intervention and judgment in AI systems, a principle strongly supported by the EU framework. Conversely, national security is crucial in the frameworks of countries like China and the US, where AI's involvement in defense and intelligence is a pivotal consideration.

This comprehensive analysis emphasizes the global trends in AI policy frameworks and the potential for harmonizing AI ethics. The section explores the possibility of converging principles and standards despite each region's diverse cultural, political, and social contexts. While complete uniformity may be unattainable, the potential for international collaboration and consensus-building on core principles is significant. Such harmonization could facilitate the establishment of universally accepted norms and standards, ensuring that AI development aligns with ethical and societal values.

The analysis of similarities and differences in Figure 8 concludes that we need continued dialogue and cooperation among nations to cultivate a responsible and ethical global AI ecosystem.

The Cartesian graph in Figure 8 includes lines representing concepts inspired by earlier Venn diagrams that explored the issues surrounding AI governance.

The lines on the graph represent different aspects of AI ethics that have been identified as necessary. The blue line represents transparency, which refers to the openness and clarity in AI policy communication. The green line represents accountability, which signifies the extent of responsibility in AI development and use. The red line represents fairness, which denotes the importance of ensuring unbiased AI systems. The purple line represents privacy, which highlights the importance of user privacy and data protection. The orange line represents human oversight, which signifies humans' involvement and oversight in AI processes. The brown line represents ethical standards, which means adherence to ethical guidelines in AI. The pink line represents innovation encouragement, which reflects support for innovative AI development. The gray line represents national security, emphasizing AI's role in national security.

These lines provide a comprehensive view of how different countries address multiple facets of AI ethics in their policies. They reflect a broader range of considerations in the global discourse on AI governance. By considering the different aspects of AI ethics, policymakers can create fair, transparent, and accountable policies while encouraging innovation and protecting user privacy and data. This is essential to building trust in AI and ensuring that it is used to benefit society.

Strategies to Mitigate Bias

This section expands into the issue of bias in Artificial Intelligence (AI) systems and its significant impact on fairness and effectiveness. We introduce Figure 8, a network diagram visually representing various interconnected strategies crucial for mitigating bias. These include ensuring data diversity to avoid biased AI models, conducting regular audits to detect and correct biases, employing bias detection tools, and emphasizing the importance of algorithmic transparency. In addition, we emphasize the importance of integrating diverse development teams and providing ethical AI training to reduce unconscious bias in AI design and development. The section then transitions to Figure 10, which provides a more detailed network representation, highlighting the synergies and dependencies among these strategies. This section demonstrates how a collective, multifaceted approach, comprising both technical and organizational measures, is vital for developing AI systems that are equitable, fair, and aligned with ethical standards. It is essential to note that mitigating bias in AI is an ongoing process that requires continuous vigilance and adaptation to evolving AI technologies and societal norms.

Comprehensive Mitigation Strategies

Bias in AI systems is of utmost importance, as it can significantly impact the fairness and effectiveness of these technologies. This section reviews various strategies aimed at mitigating bias, presented as a network diagram showcasing the interconnectedness and collective importance of these strategies.

The network diagram encompasses a range of approaches, each linked to demonstrate how they complement and reinforce one another. Key strategies include ensuring data diversity and using datasets representing all relevant demographics to prevent biased AI models. Regular audits are crucial to identifying and addressing biases that may develop over time. Additionally, the network emphasizes the importance of using bias detection tools, which employ specialized algorithms to uncover and address biases in AI systems.

The diagram in Figure 9 emphasizes the significance of collectively implementing these strategies, indicating that the most effective approach to mitigating bias involves a multifaceted effort. This includes technical solutions and organizational and procedural measures to ensure that AI systems are developed and operated in a manner that minimizes bias and promotes fairness.

The diagram in Figure 9, shows the connections and assigned weights between different strategies to mitigate bias in AI. The thickness of the lines corresponds to the strength of the connection, and the weights are labeled on the diagram. The strategies are arranged circularly to emphasize their interconnectedness and importance. By implementing them collectively, the risk of

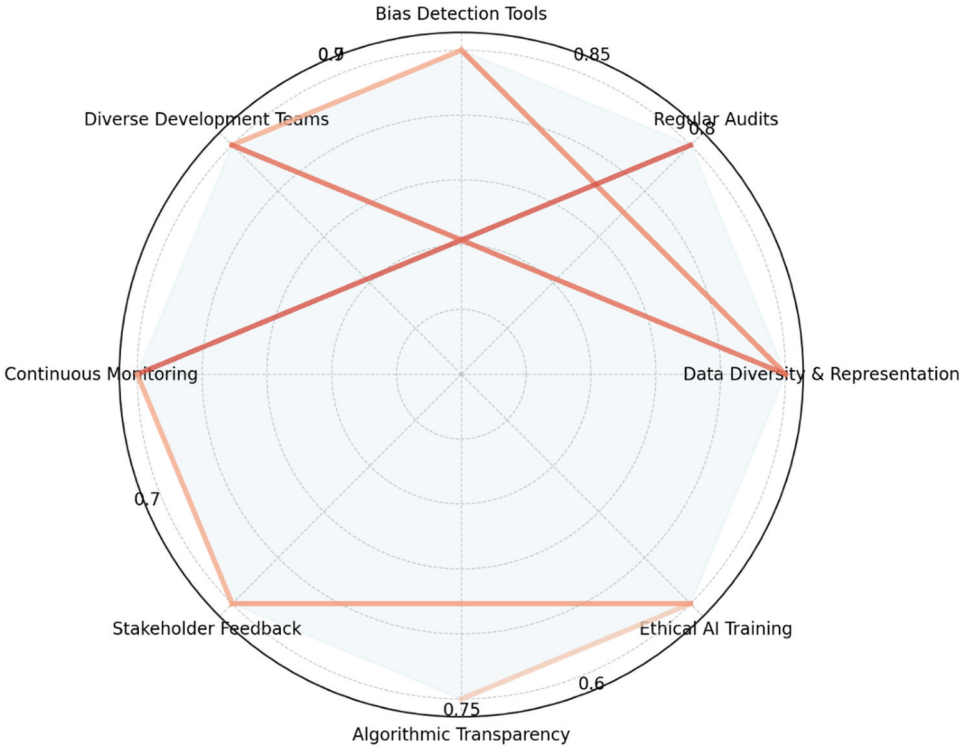

Figure 9. The significance of a collective implementation of AI Strategies: connections and Weights in Strategies to Mitigate Bias in AI (Colour-coded by Strength).

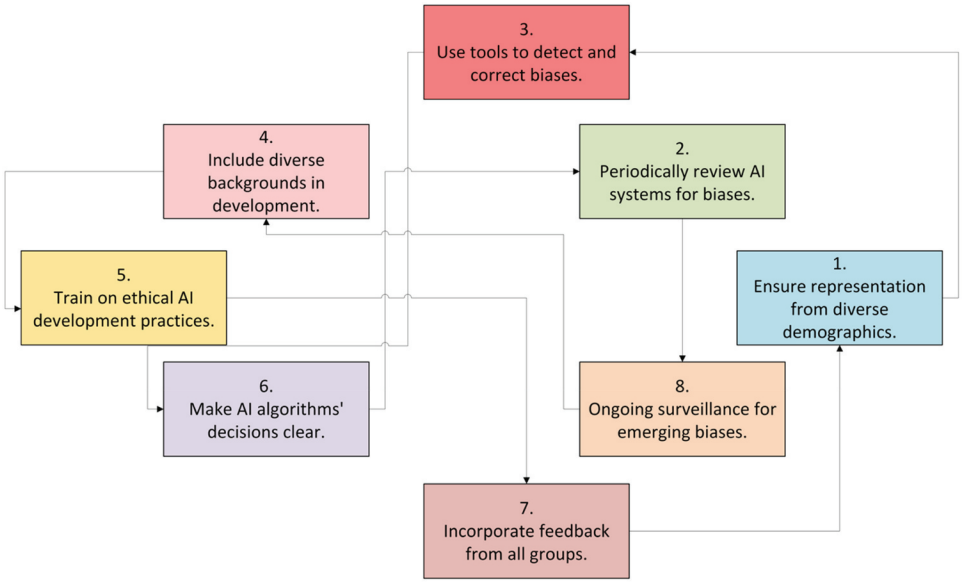

Figure 10. Strategies to Mitigate AI Bias.

bias in AI systems can be significantly reduced, leading to more equitable and trustworthy AI solutions.

One key strategy is to ensure data diversity and representation. This involves using diverse data representing all relevant demographics to avoid biased models in AI systems. Regular audits are also essential to identify and rectify any biases that may have crept in over time. Bias detection tools are another essential strategy. These tools utilize specialized software to detect biases in AI algorithms, which can then be corrected. Diverse development teams can also help minimize unconscious biases in designing and developing AI systems.

Providing ethical AI training to AI professionals is another way to mitigate bias in AI. This training educates AI professionals on ethical considerations and avoiding bias. Algorithmic transparency is also crucial to reducing bias in AI. By making the workings of AI algorithms transparent, biases can be identified and corrected more quickly. Involving various stakeholders, including those from underrepresented groups, to provide feedback on AI systems and their outputs can also significantly reduce bias. Finally, continuously monitoring AI systems is essential to quickly identify and address any biases that may emerge over time. Collectively implementing these strategies can significantly reduce the risk of bias in AI systems, leading to more equitable and trustworthy AI solutions.

The flowchart in Figure 10 outlines a network representation of strategies for mitigating bias in AI. The enhanced diagram provides more context and displays the connections between these strategies, providing a comprehensive approach to addressing this issue.

As shown in Figure 9, data diversity is one of the primary strategies to mitigate bias in AI. It is essential to ensure that data is collected from diverse demographics. This strategy is linked to bias detection tools, emphasizing the importance of having a wide range of data to identify and correct biases. Regular audits are another crucial component of mitigating bias in AI systems. Periodic reviews of AI systems for biases are required and are connected to continuous monitoring. This highlights the need for ongoing assessments to ensure the AI system remains unbiased.

Using bias detection tools is also essential in mitigating bias in AI systems. This strategy links to algorithm transparency, underscoring detection tools' role in making AI decisions more straightforward. Ensuring that the AI algorithm is transparent makes it easier to identify and correct any potential biases.

Diverse teams are also vital to mitigating bias in AI systems. Having diverse backgrounds in development teams can significantly reduce unconscious biases. This strategy is connected to ethical training, showing the importance of diverse perspectives in ethical AI development. This ensures that the AI system is developed ethically and unbiasedly.

Ethical training is another crucial strategy in mitigating bias in AI systems. Training on ethical AI development practices is vital, and this connects to stakeholder feedback. This illustrates the role of ethical considerations in incorporating diverse viewpoints, ensuring that the AI system is developed with the interests of all stakeholders in mind.

Making AI algorithms' decisions clear is another essential strategy for identifying biases. This is connected to regular audits, highlighting the need for transparency in ongoing assessments. Ensuring that the AI algorithm's decisions are transparent makes it easier to identify and correct any potential biases.

Incorporating feedback from all groups, including underrepresented ones, is essential in developing an unbiased AI system. This relates back to data diversity, emphasizing the role of inclusive feedback in ensuring diverse data representation. Diverse feedback and ongoing surveillance for emerging biases is necessary to mitigate bias in AI systems. This relates to diverse teams, underscoring the need for continuous oversight by teams with varied backgrounds and perspectives. Diverse teams make it easier to identify and correct potential biases.

The network diagram in Figure 9 illustrates the interconnected nature of these strategies, showing how each contributes to a comprehensive approach to mitigating bias in AI systems. By implementing these strategies, AI systems can be developed more ethically and unbiasedly, with the interests of all stakeholders in mind.

Ethical Training and Diverse Teams

Providing ethical training to AI professionals is crucial to making them aware of potential biases and fostering an ethical culture in AI development. This training should cover the ethical implications of AI, the significance of diversity in datasets, and ways to detect and mitigate bias.

Forming diverse teams is also a vital strategy. Teams composed of people from diverse backgrounds bring unique perspectives to the AI development process, which can help identify biases that a more homogeneous group may overlook. The diversity here refers to demographic factors and variations in expertise, experience, and viewpoints.

By combining ethical training, diversity in teams, and technical strategies such as data diversity and regular audits, AI systems can be developed in a way that is more equitable, fair, and aligned with ethical standards. Countering bias is a continuous process that necessitates ongoing attention and adaptation as AI technologies progress.

Emerging AI Technologies and Their Ethical Challenges

Large language models (LLMs), such as GPT-3 and GPT-4, exemplify the growing power of generative AI. These models are trained on vast datasets, often scraping data from the internet, which raises significant concerns over **data provenance**, **copyright infringement**, and **privacy violations**. The opaque nature of these models complicates efforts to ensure that they are free from biases present in the training data, such as discriminatory language, misinformation, or unintentional perpetuation of harmful stereotypes. Despite employing fine-tuning and debiasing techniques, these models are still prone to producing biased outputs due to the inherent limitations of the training data and the probabilistic nature of their generation processes. For instance, techniques such as **reinforcement learning from human feedback (RLHF)** have been deployed to mitigate harmful outputs, but they remain insufficient in addressing the deeper systemic biases embedded within the underlying datasets. This calls for more sophisticated techniques, such as **adversarial training**, where adversarial examples are used to iteratively refine models and expose hidden biases. Additionally, **federated learning** presents a promising approach for enhancing the ethical training of LLMs by allowing models to learn from decentralized, anonymized data, thus reducing the ethical risks associated with data centralization and privacy violations.

Another critical issue with LLMs lies in their ability to produce convincing but factually incorrect or **hallucinatory outputs**. This problem, often referred to as the "hallucination problem," presents ethical challenges in high-stakes domains such as healthcare or law, where accurate information is paramount. Current mitigation strategies include **truth-verification models** that cross-check generated content against verified databases and **automated fact-checking systems**integrated into the model's inference pipeline. However, these methods are still evolving and are far from fully resolving the issue. Ethical frameworks for LLM deployment must, therefore, include rigorous post-deployment monitoring and real-time validation mechanisms to ensure the integrity of the outputs, particularly in applications where misinformation could have profound societal impacts.

Autonomous systems, including autonomous vehicles, drones, and robotic systems, pose additional ethical challenges related to **safety**, **accountability**, and **decision-making autonomy**. A key ethical dilemma arises in the context of **autonomous decision-making** in unpredictable environments. For instance, in the case of autonomous vehicles, ethical frameworks must account for the so-called **trolley problem** scenarios, where the system must make life-and-death decisions in the event of an unavoidable accident. Traditional rule-based ethical systems, such as **deontological** or **utilitarian** approaches, often fail to provide clear solutions in these nuanced scenarios. Consequently, emerging solutions

involve the use of **ethical AI algorithms** like **multi-objective optimization**, which allows systems to balance competing ethical principles – such as minimizing harm and respecting human autonomy – by assigning dynamic weights to different ethical outcomes based on real-time environmental factors.

Furthermore, **accountability in autonomous systems** presents a unique challenge, especially in cases where systems operate with minimal human oversight. **Explainable AI (XAI)** plays a critical role here, enabling transparency in decision-making processes by providing interpretable insights into how the system reached a specific decision. Techniques such as **attention mechanisms** and **saliency maps** can be employed to highlight the features that most influenced an autonomous system's decision, making it easier for regulators and auditors to understand and assess the fairness and safety of these decisions. However, the effectiveness of XAI in highly complex, real-time autonomous systems remains limited, necessitating the development of **causal inference** models that can provide a more comprehensive understanding of decision-making pathways and their underlying ethical implications.

The issue of **algorithmic accountability** in autonomous weapons systems (AWS) presents perhaps the most acute ethical challenge. The development and deployment of AWS raise profound concerns over **autonomous lethality**—the ability of a system to make life-or-death decisions without human intervention. Current discussions on **international AI governance** focus on the need to restrict the deployment of AWS through legally binding treaties, but enforcement mechanisms remain elusive. From a technical standpoint, one proposed solution involves embedding **human-in-the-loop (HITL)** mechanisms that ensure critical decisions, particularly those involving the use of lethal force, require human validation before execution. This integration of human oversight into decision-making processes is critical to preventing unintended harm and ensuring compliance with international humanitarian law. Additionally, ongoing research into **ethical-by-design architectures** aims to build ethical constraints directly into the system's operational framework, limiting the scope of actions that an autonomous system can take based on predefined ethical guidelines.

Finally, the deployment of **swarm intelligence** in autonomous drones and robots introduces challenges related to **collective decision-making** and **distributed accountability**. In swarm systems, decisions are often made collectively by a distributed group of agents, with no single agent being responsible for the final outcome. This creates significant ethical ambiguity in determining accountability when swarm systems malfunction or cause harm. Solutions such as **distributed ledger technologies (DLT)**, including **blockchain**, have been proposed to ensure that every decision made within the swarm is recorded in a transparent and immutable way, providing a traceable log of actions that can be audited for accountability purposes.

Discussion

The introduction of fairness, transparency, and accountability into AI systems, while crucial for ensuring ethical standards, introduces a significant financial burden and operational complexity, especially in sectors where fast innovation is a competitive necessity.

One of the primary economic costs arises from the increased complexity in developing AI systems that adhere to ethical guidelines. Implementing fairness-aware learning algorithms, such as demographic parity or equalized odds, requires additional computational resources and extensive testing during the training phase. These fairness constraints are not simply add-ons but require a fundamental rethinking of the algorithmic design, particularly in cases where performance optimization conflicts with fairness. For instance, in financial services, ensuring that loan approval algorithms do not exhibit bias may necessitate retraining models with diverse datasets and applying fairness constraints throughout the development cycle. This extended development process incurs higher labor costs, requires greater infrastructure investment, and often results in longer timeframes to achieve regulatory compliance. Additionally, privacy-preserving techniques, such as differential privacy and federated learning, add further complexity. Federated learning, which enables model training across distributed datasets without centralizing sensitive data, requires more sophisticated system architectures and secure communication channels, increasing both the cost and technical difficulty of implementation.

Operationally, the impact of strict ethical guidelines is felt through the need for ongoing compliance and continuous monitoring of AI systems. Ethical frameworks such as the European Union's AI Act mandate that high-risk AI applications, particularly in fields like healthcare and criminal justice, undergo continuous auditing to ensure ethical standards are maintained post-deployment. These operational costs are amplified by the need to integrate real-time fairness monitoring tools, such as AI Fairness 360, which check for bias drift or decision-making anomalies as AI systems encounter new data. These tools require continuous computational resources, infrastructure support, and personnel dedicated to auditing and model recalibration. For industries such as financial services, where AI systems are deployed in real-time environments like high-frequency trading, maintaining fairness and compliance adds layers of complexity to the operational workflow. This constant need for recalibration can also result in downtime, during which systems must be reevaluated and updated, leading to delays in decision-making processes and potential disruptions to business continuity.

The financial impact of these ethical requirements also affects innovation cycles and speed to market. In highly competitive sectors like autonomous driving or AI-driven diagnostics, time-to-market is often crucial for gaining a first-mover advantage. Companies that invest heavily in ethical compliance –

such as model transparency, fairness audits, and explainability – may experience delays in bringing products to market. For instance, the requirement to integrate explainability mechanisms, such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations), into AI models often necessitates additional development and testing phases. This extends the overall project timeline and may place companies at a competitive disadvantage against those who prioritize rapid deployment over ethical oversight. The delay not only impacts short-term revenue but also affects long-term strategic positioning, particularly in industries where technological leadership is key to maintaining market share.

Beyond development and operational costs, legal compliance and regulatory risk are significant financial considerations for companies implementing strict ethical guidelines. Regulatory frameworks like the GDPR and the upcoming EU AI Act impose severe penalties for noncompliance, with fines that can reach up to 4% of a company's global revenue for violations of data privacy and transparency requirements. To mitigate these risks, companies often need to invest heavily in legal teams, external audits, and compliance infrastructures. This introduces an additional cost layer as companies must allocate resources not just for initial development but also for ongoing compliance management. The cyclical nature of compliance – where systems must be continuously updated, audited, and re-certified to meet evolving standards – creates long-term financial commitments that extend well beyond the initial implementation of AI systems.

Despite these costs, emerging technologies offer potential solutions that could mitigate some of the financial and operational burdens associated with ethical AI. Automated machine learning (AutoML) systems are increasingly capable of incorporating fairness and transparency checks into their development pipelines, reducing the need for manual intervention and thus lowering labor costs. Additionally, distributed ledger technologies (DLT), such as blockchain, can help track AI decisions in a transparent and immutable way, thereby simplifying post-deployment audits and reducing the cost of maintaining ethical standards. Nevertheless, while these technologies offer some relief, they come with their own set of technical challenges and infrastructural costs, which require additional investment and expertise to implement effectively.

The implementation of strict ethical guidelines in AI development significantly impacts economic and operational aspects of AI projects. While these guidelines are crucial for ensuring fairness, transparency, and accountability, they introduce substantial costs at every stage of the AI lifecycle, from development through to post-deployment monitoring and compliance. Balancing these ethical obligations with the need for innovation and market competitiveness remains a challenge, particularly for companies operating in highly

dynamic and competitive sectors. The evolving landscape of AI governance, coupled with emerging cost-saving technologies, will be critical in determining how companies navigate the financial and operational implications of ethical AI development.

Conclusion

This study has undertaken an examination of the ethical imperatives surrounding AI, particularly the principles of transparency, fairness, and privacy, in the context of its prevalent influence across sectors such as healthcare, finance, and communication. The deployment of AI technologies in these domains brings with it profound ethical challenges that necessitate a strong and inclusive framework to safeguard individual rights and societal interests. Through a comparative analysis of international AI policy frameworks from the European Union, the United States, and China, this research has clarified the conflicting ethical priorities that shape AI governance globally.

This work clarifies the ethical principles of privacy, transparency, and fairness, addressing regional challenges and interdependencies. By distinguishing how these principles operate independently yet interactively across frameworks, the paper offers a refined conceptual foundation necessary for global governance. A primary contribution is the proposed set of integration criteria (Interoperability, Normative Cohesion, Cultural Adaptability, and Transparency of Process). These criteria provide a structured foundation for aligning ethical principles across diverse international frameworks, supporting cross-border AI compatibility while respecting region-specific values and regulatory approaches. The graphical representations represent the individual and corelated interdependencies and conflicts among frameworks, avoiding oversimplification and enhancing analytical clarity. This provides a visual tool for understanding ethical elationships in global AI governance.

The analysis reveals marked variations in how different regions balance the demands of innovation against the ethical principles of privacy, fairness, and accountability. While certain jurisdictions, such as the European Union, emphasize stringent regulatory oversight and data protection, others, including the United States, adopt a more flexible, innovation-centric approach. These divergences underscore the complexities involved in striving for a harmonized global standard for ethical AI governance. Nevertheless, this study has articulated several strategic interventions to mitigate algorithmic bias, including the deployment of fairness-aware algorithms, regular audits, and the incorporation of diverse development teams. These interventions are essential in fostering equitable and trustworthy AI systems.

Furthermore, this research has highlighted the critical need for sustained international collaboration and dialogue to bridge the gaps in global AI ethics frameworks. It is increasingly evident that no single jurisdiction can fully

address the multi-faceted ethical challenges posed by AI in isolation. Instead, the path forward demands a concerted, cooperative effort that leverages shared principles while respecting regional variations in regulatory and cultural priorities.

This study advances the argument that ethical considerations must be embedded at every stage of the AI development lifecycle, from inception through to deployment and beyond. The recommendations herein aim to inform policymakers, regulators, and AI developers, encouraging the pursuit of AI systems that are not only innovative and technologically advanced but also aligned with the highest ethical standards. As AI continues to evolve and exert its transformative potential, the need for vigilance, adaptability, and cross-border cooperation remains paramount in ensuring that these technologies serve the common good, promoting fairness, accountability, and trust in their application.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Engineering and Physical Sciences Research Council [EP/S035362/1].

ORCID

Petar Radanliev http://orcid.org/0000-0001-5629-6857

Data Availability Statement

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Due to the sensitive nature of the data related to AI ethics and privacy considerations, access to the data may be restricted. Specific details regarding the data sources, including international AI policy frameworks from the EU, US, China, Canada, Japan, India, and Australia, are documented within the study. All data shared will be compliant with ethical guidelines and privacy standards as outlined in the General Data Protection Regulation (GDPR) and other relevant data protection laws.

References

Aldoseri, A., K. N. Al-Khalifa, and A. M. Hamouda. 2023. Re-Thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges. *Applied Sciences* 13 (12):7082. doi: 10.3390/APP13127082.

Bécue, A., I. Praça, and J. Gama. 2021. Artificial intelligence, cyber-threats and industry 4.0: Challenges and opportunities. *Artificial Intelligence Review* 54 (5):3849–86. doi: 10.1007/s10462-020-09942-2.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. doi: 10.1145/3442188.3445922.

Binns, R. 2018. Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, vol. 81, 149–59, PMLR. https://proceedings.mlr.press/v81/binns18a.html.

Bommasani, R., K. Klyman, D. Zhang, and P. Liang. 2023. *Do foundation model providers comply with the draft EU AI act?* Center for Research on Foundation Models (CRFM): Stanford Center for Research on Foundation Models.

Bostrom, N., and E. Yudkowsky. 2014. The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence* 316–34. doi: 10.1017/CBO9781139046855.020.

Brynjolfsson, E., and A. Mcafee. 2014. The second machine age: Work, progress, and prosperity in a time of brilliant technologies. In The second machine age: Work, progress, and prosperity in a time of brilliant technologies. Worldwide: W.W. Norton & Company 978-0-393-35064-7 https://wwnorton.com/books/the-second-machine-age/ .

de Bruin, B., and L. Floridi. 2017. The ethics of cloud computing. *Science and Engineering Ethics* 23 (1):21–39. doi: 10.1007/s11948-016-9759-0.

de Fine Licht, K., and J. de Fine Licht. 2020. Artificial intelligence, transparency, and public decision-making. *AI & Society* 35 (4):917–26. doi: 10.1007/s00146-020-00960-w.

Du, M. 2023. *Awesome-Fairness-in-AI*. GitHub Repository. https://github.com/datamllab/awesome-fairness-in-ai.

European Parliament. 2023. *AI act: A step closer to the first rules on artificial intelligence | news | European Parliament*. https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence.

Evans, K. 2015. The second machine age: Work, progress, and prosperity in a time of brilliant technologies by eric Brynjolfsson and Andrew McAfee. *Journal of Business & Finance Librarianship* 20 (3):244–46. doi: 10.1080/08963568.2015.1044355.

FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety | The White House. 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/.

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al. 2018. AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines* 28 (4):689–707. doi: 10.1007/s11023-018-9482-5.

GDPR. 2018. *What is GDPR, the EU's new data protection law? - Gdpr.Eu*. https://gdpr.eu/what-is-gdpr/.

Helbing, D., B. S. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, J. Van Den Hoven, R. V. Zicari, and A. Zwitter. 2018. Will democracy survive big data and artificial intelligence? In Towards digital enlightenment: Essays on the dark and light sides of the digital revolution, 73–98. Springer International Publishing. doi: 10.1007/978-3-319-90869-4_7.

HIPAA. 1996. *Health insurance portability and accountability act of 1996 (HIPAA) | CDC*. https://www.cdc.gov/phlp/publications/topic/hipaa.html.

Hosny, A., C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts. 2018. Artificial intelligence in radiology. *Nature Rev Cancer* 18 (8):500. doi: 10.1038/S41568-018-0016-5.

IBM. 2018. *AI fairness 360 – open source*. Open Project. https://www.ibm.com/opensource/open/projects/ai-fairness-360/.

ICO. 2018. *Information commissioner's office (ICO): The UK GDPR*. UK GDPR Guidance and Resources. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/lawful-basis-for-processing/consent/.

Interim Measures for the Management of Generative Artificial Intelligence Services, Personal Information Protection Law of the People's Republic of China (PRC). 2023.

ISO. 2023. *ISO/IEC DIS 42001 - information technology — artificial intelligence — management system*. https://www.iso.org/standard/81230.html.

Jobin, A., M. Ienca, and E. Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence 2019* 1 (9):389–99. doi: 10.1038/s42256-019-0088-2.

Li, L. 2017. China's manufacturing locus in 2025: With a comparison of "Made-in-China 2025" and "Industry 4.0". *Technological Forecasting & Social Change* 135:66–74. doi: 10.1016/J.TECHFORE.2017.05.028.

Malhotra, Y. 2018. Cognitive computing for anticipatory risk analytics in intelligence, surveillance, & reconnaissance (ISR): Model risk management in artificial intelligence & machine learning (presentation slides). *SSRN Electronic Journal*. doi: 10.2139/ssrn.3111837.

McCorduck, P., and C. Cfe. 2004. *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press. https://books.google.com/books?hl=en&lr=&id=r2C1DwAAQBAJ&oi=fnd&pg=PP1&dq=Pamela+McCorduck+%22Machines+Who+Think&ots=UnmXIiuRtM&sig=JAh90Eu07MGvjS5OgFq1CMy6-gc.

Meissner, G. 2020. Artificial intelligence: Consciousness and conscience. *AI & Society* 35 (1):225–35. doi: 10.1007/s00146-019-00880-4.

MeitY. 2023. Artificial intelligence committees reports | Ministry of electronics and information technology, Government of India. *Artificial Intelligence Committees Report*. https://www.meity.gov.in/artificial-intelligence-committees-reports.

Mijwil, M. M., M. Aljanabi, and ChatGPT. 2023. Towards artificial intelligence-based cybersecurity: The practices and ChatGPT generated ways to combat cybercrime. *Iraqi Journal for Computer Science and Mathematics* 4 (1):65–70. doi: 10.52866/IJCSM.2023.01.01.0019.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1 (11):501–07. doi: 10.1038/s42256-019-0114-4.

Mozumder, M. A. I., M. M. Sheeraz, A. Athar, S. Aich, and H.-C. Kim. 2022. Overview: Technology roadmap of the future trend of metaverse based on IoT, blockchain, AI technique, and medical domain metaverse activity. *International Conference on Advanced Communication Technology (ICACT)* 256–61. doi: 10.23919/ICACT53585.2022.9728808.

NAIAC. 2024. *AI safety: National AI advisory committee*. https://ai.gov/wp-content/uploads/2024/06/FINDINGS-RECOMMENDATIONS_AI-Safety.pdf.

NIST. 2023a. *AI risk management framework | NIST*. National Institute of Standards and Technology. https://www.nist.gov/itl/ai-risk-management-framework.

NIST. 2023b. *Artificial intelligence | NIST*. https://www.nist.gov/artificial-intelligence.

NIST. 2024a. *AI risk management framework | NIST*.

NIST. 2024b. *AI standards | NIST*. https://www.nist.gov/artificial-intelligence/ai-standards.

NIST. 2024c. *Department of commerce announces new guidance, tools 270 days following president Biden's executive order on AI | NIST*. https://www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following.

Office for Artificial Intelligence and Department for Science, Innovation & Technology. 2023. *A pro-innovation approach to AI regulation* 978-1-5286-4009-1 (London: Crown copyright).

Partnership on AI. 2023. *Partnership on AI and the ethical AI framework for social good*. https://partnershiponai.org/.

Provisions on the Administration of Deep Synthesis Internet Information Services, Personal Information Protection Law of the People's Republic of China (PRC). 2022.

Roberts, H., J. Cowls, J. Morley, M. Taddeo, V. Wang, and L. Floridi. 2021. The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & Society* 36 (1):59–77. doi: 10.1007/s00146-020-00992-2.

Shu, Y., J. Zhang, and H. Yu. 2021. Fairness in design: A tool for guidance in ethical artificial intelligence design. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12774:500–10. doi: 10.1007/978-3-030-77626-8_34.

Singer, P. W. 2009. *Wired for war: The robotics revolution and conflict in the twenty-first century*, 499. https://books.google.com/books/about/Wired_for_War.html?id=AJuowQmtbU4C.

The State Council People Republic of China. 2017. *Made in China 2025; the state council people Republic of China*. http://english.gov.cn/2016special/madeinchina2025/ .

Tabassi, E. 2023. *AI risk management framework | NIST*. doi: 10.6028/NIST.AI.100-1.

Turilli, M., and L. Floridi. 2009. The ethics of information transparency. *Ethics and Information Technology* 11 (2):105–12. doi: 10.1007/s10676-009-9187-9.

UNESCO. 2023. *Recommendation on the ethics of artificial intelligence | UNESCO*. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.

Wachter, S., B. Mittelstadt, and C. Russell. 2023. *Health care bias is dangerous. But so are fairness' algorithms | WIRED*. Wired. https://www.wired.com/story/bias-statistics-artificial-intelligence-healthcare/.

Yu, K. H., A. L. Beam, and I. S. Kohane. 2018. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2 (10):719–31. doi: 10.1038/S41551-018-0305-Z.