Revolutionizing AI Governance: Addressing Bias and Ensuring Accountability Through the Holistic AI Governance Framework

Ibrahim Atoum
Department of Artificial Intelligence-Faculty of Science and Information Technology,
Al-Zaytoonah University of Jordan, Amman, 11733, Jordan

***Abstract*—Artificial intelligence (AI) possesses the capacity to transform numerous facets of our existence; however, it concomitantly engenders considerable risks associated with bias and discrimination. This article explores emerging technologies like Explainable AI (XAI), Fairness Metrics (FMs), and Adversarial Learning (AL) for bias mitigation while emphasizing the critical role of transparency, accountability, and continuous monitoring and evaluation in AI governance. The Holistic AI Governance Framework (HAGF) is introduced, featuring a comprehensive, five-layered structure that integrates top-down and bottom-up strategies. HAGF prioritizes foundational principles and resource allocation, outlining five lifecycle-specific phases. Unlike the OECD AI Principles, which offer a general ethical framework lacking holistic perspective and resource allocation guidance, and the Berkman Klein Center's Model, which provides a broad framework but omits resource allocation and detailed implementation, HAGF offers actionable mechanisms. Tailored Key Performance Indicators (KPIs) are proposed for each HAGF layer, enabling ongoing refinement and adaptation to the evolving AI landscape. While acknowledging the need for enhancements in data governance and enforcement, the embedded KPIs ensure accountability and transparency, positioning HAGF as a pivotal framework for navigating the complexities of ethical AI.**

***Keywords*—Artificial intelligence; framework; bias; discrimination; governance; key performance indicators**

I. Introduction

The iterative process of AI model development is crucial for enabling machines to perform human-like tasks such as problem-solving, learning, decision-making, and perception. This process enhances technological systems by developing algorithms that process large datasets, recognize patterns, and make predictions. A critical step is training, which uses labeled data to adjust model parameters and optimize performance. Therefore, the quality and quantity of data are vital for success [1]. Accurate and representative training data directly impacts model performance; biased data can lead to errors, particularly in critical fields like healthcare, finance, and criminal justice [2].

In AI, bias refers to systematic prediction distortions arising from training data and model design, while discrimination involves unfair treatment based on characteristics such as age, race, or gender. Although bias can contribute to discrimination, they are not synonymous; discrimination can occur independently of bias, and reducing bias may not eliminate discrimination [3]. Data labeling and algorithm design biases can perpetuate discriminatory patterns, and a lack of diversity in development teams can hinder identifying and mitigating these biases.

To promote fairness and equity in AI, it is essential to ensure diverse and representative training data, fair feature selection, rigorous evaluation strategies, and effective governance frameworks. These frameworks should include policies and guidelines for safe and ethical AI development, emphasizing auditing, transparency, and stakeholder collaboration [4]. Data augmentation and synthetic data generation are necessary to ensure data accuracy and representativeness, as high-quality data is crucial for avoiding biased AI models.

Four key components support responsible AI development: XAI, FMs, AL, and ongoing M&E [5]. While XAI enhances transparency, it faces challenges such as computational costs. FMs may conflict with accuracy, necessitating a holistic approach. AL can mitigate bias through adversarial examples, and continuous M&E is vital for detecting biases over time. Integrating these elements is essential for developing fair and accountable AI systems.

Governments play a critical role in promoting AI's safe and responsible implementation [6]. They can establish ethical guidelines, engage stakeholders, and enforce principles addressing bias, privacy, safety, and security. Regulation in sensitive areas like healthcare and criminal justice ensures adherence to ethical standards. Governments can also promote industry self-regulation through voluntary codes of conduct and support international cooperation to establish common standards.

This paper compares the Holistic AI Governance Framework (HAGF) with the OECD AI Principles [7] and the Berkman Klein Center's Model for AI Governance [8], focusing on the HAGF's unique contribution to AI governance. The HAGF offers a holistic, cyclical approach to resource allocation, structured around five interconnected components: establishing standards, regulating adherence, promoting AI advancement, fostering transparency and accountability, and encouraging a voluntary code of conduct. Each component includes defined roles and theoretical weights, reflecting a vision for responsible AI development and providing enhanced strategic guidance. The HAGF emphasizes the need for formal

guidelines and policies to ensure ethical AI deployment and effective resource allocation for research and development.

While the OECD AI Principles provide foundational ethical guidelines, they have been criticized for lacking a holistic approach to implementation and sufficient guidance on resource allocation, leading to potentially inconsistent interpretations and inadequate stakeholder engagement [7]. Similarly, though comprehensive, the Berkman Klein Center's Model is criticized for its breadth and lack of specific implementation measures and clear metrics for evaluating effectiveness [8]. The HAGF addresses these shortcomings by focusing on resource allocation and integrating top-down and bottom-up strategies to foster ethical AI practices. Furthermore, this paper examines key technical approaches to mitigating bias and enhancing ethical AI development, proposing KPIs for each phase to monitor effectiveness and facilitate continuous improvement.

Section II presents related work, Section III outlines the importance of gathering diverse data, Section IV explores methods for identifying bias in AI, Section V discusses the management of ethical AI practices, Section VI demonstrates study results, and Section Seven summarizes conclusions.

II. Related Work

Ethical AI frameworks are increasingly recognized for guiding responsible AI development and deployment. Notable frameworks, such as the OECD AI Principles [9, 10] and the Berkman Klein Center's Model for AI Governance [11], provide essential high-level guidance but are often criticized for lacking specific implementation details. Key shortcomings include inadequate resource allocation and insufficient stakeholder engagement, complicating the translation of principles into practice, especially regarding diverse participation [12, 13, 14, 15, 16, 17, 18].

Effectively addressing bias and discrimination in AI systems necessitates implementing concrete strategies. As highlighted by study [19] in the context of surgical care, a comprehensive approach is essential for achieving equity in sensitive areas such as healthcare [1, 5, 6]. This approach should go beyond general data-centric efforts and incorporate targeted interventions at each stage of the AI development lifecycle.

To effectively mitigate bias and discrimination, various strategies can be employed, as illustrated in Fig. 1, including enhancing data quality through data augmentation and bias detection [19, 20, 21], promoting transparency via model interpretability and audit trails [22, 23, 24], and utilizing XAI to build trust [25, 26, 27, 28, 29, 30, 31, 32]. Additionally, continuous monitoring processes and implementing fairness-aware algorithms and metrics are crucial for developing fairer AI systems and reducing discriminatory outcomes [33, 34, 29, 3, 4, 36].

Implementing these technical solutions requires integration into a comprehensive governance framework that articulates ethical principles and offers actionable steps, including resource considerations and stakeholder involvement. The growing use of AI across various sectors—such as healthcare, finance, education, and the judiciary—highlights the necessity for robust governance frameworks to tackle unique ethical challenges [1, 5, 6, 37, 38, 39, 40, 41, 42].

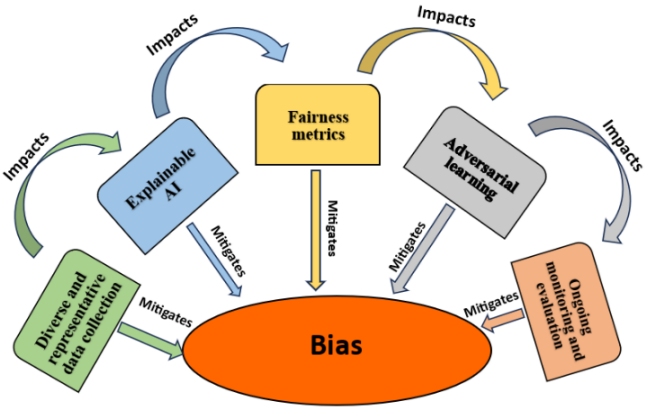

Fig. 1. Strategies for mitigating bias in AI systems.

The HAGF is proposed, which builds on existing efforts by offering a cyclical and interconnected approach to AI governance. HAGF emphasizes resource allocation and combines top-down policy development with bottom-up self-regulation, ensuring continuous refinement. Its focus on resource allocation at each phase, with specific weights assigned to sub-phases, effectively addresses implementation challenges. Critical areas include risk identification, data protection, transparency, and oversight mechanisms.

HAGF's cyclical nature facilitates continuous adaptation to the evolving AI landscape, while KPIs measure effectiveness and progress in ethical AI development. Unlike broader frameworks, HAGF offers specific guidance on data protection, providing organizations with concrete compliance mechanisms. While HAGF has notable strengths, it also recognizes limitations, including the need for deeper guidance on specific ethical issues and implementation challenges; these will be addressed through ongoing refinement and the inclusion of KPIs to track progress. Furthermore, the growing emphasis on auditing AI systems for fairness and bias aligns with HAGF's focus on continuous monitoring, addressing implicit bias, and promoting inclusivity as essential components of the framework.

III. Inclusive Data Collection

Diverse data is essential for developing fair and inclusive AI. Collecting information from various sources and involving multiple stakeholders helps identify potential biases during design and training, enhancing AI accuracy and fairness. In contrast, limited data diversity can lead to overfitting, reduced performance, and the perpetuation of biases against marginalized groups, as illustrated in Fig. 2. The figure underscores these interconnected challenges, including ethical concerns. Therefore, addressing these issues requires careful attention to data collection, model development, and evaluation.

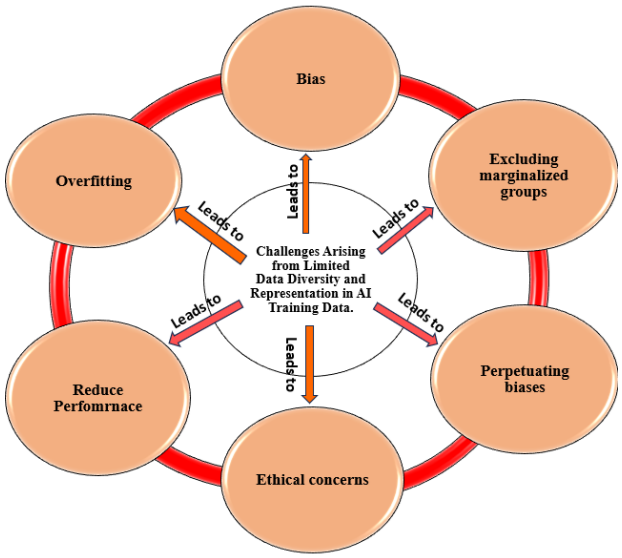

Fig. 2. Challenges arising from limited data diversity and representation.

To address these problems, carefully curating the data to ensure accuracy and representativeness and using techniques such as data augmentation and synthetic data generation to increase diversity and representation. Data curation, the practice of meticulously selecting, organizing, and maintaining data for long-term quality and usability [40], encompasses collection, validation, transformation, integration, and archiving activities. This ensures a trustworthy and well-documented resource for effective analysis and reuse. Building on this foundation, data augmentation transforms existing data to create new, diverse examples, ultimately enhancing model accuracy and robustness [20]. These transformations, tailored to the data type and the problem at hand, can help reduce overfitting and address imbalanced datasets by generating new examples for underrepresented classes, thereby improving overall model performance.

High-quality data is essential for training AI models (as discussed earlier), but challenges arise in acquiring diverse data [40]. Synthetic data generation can address limited or biased training data by creating variations from existing information to improve model fairness and robustness [20]. However, collecting representative data poses difficulties [21, 24, 26, 43, 44]. Biased training data leads to biased AI models, a significant concern for critical applications [45]. Data governance efforts, like audits and fairness benchmarks, aim to ensure the use of representative data, promoting fair and inclusive AI decision-making [46].

IV. Unveiling Bias in AI: XAI, FMs, AL, and M&E

As AI permeates our lives, ensuring fairness and trust is paramount. Three guardians stand watch over responsible development: XAI acts as a transparency lens, revealing how AI makes decisions. FMs, watchful guardians, identify potential biases for equal treatment by AI. Finally, AL strengthens models by mimicking real-world challenges. M&E continuously safeguards against bias creep, ensuring ongoing fairness through regular checks. This ensures AI remains fair and accountable. Despite the oversight role of these guardians, they come with challenges, as outlined in Table I.

Opaque AI systems can lead to biased decisions. XAI cuts through this, revealing AI's reasoning and becoming a powerful tool for fighting bias in critical healthcare fields [27]. Fig. 3 illustrates a typical AI model's limitation: it does not explain its predictions. XAI addresses this by adding an explanation layer and building trust through transparency.

TABLE I. Summary of the Challenges of AI Bias Mitigation Technology

Diverse and representative data collection	***XAI***	***FMs***	***AL***	***On-Going Monitoring***
Data may be incomplete or biased	Computationally expensive	contextual challenges and considerations.	Resource intensive	Resource-intensive.
difficult and computationally expensive	Explanation-challenged	Trade-offs between fairness and performance	limited efficacy	Data bias and incompleteness.
The potential for privacy concerns arises from biases and power imbalances in data collection	Explanations bias	vulnerability to manipulation	over-specialization	Limitations in detecting emerging issues
Potential biases and power imbalances in data collection	Privacy implications	Fairness checklist pitfall	Ethical concerns arise from the possibility of malicious use of AI	The opacity of AI systems limits transparency and interpretability, raising privacy concerns.

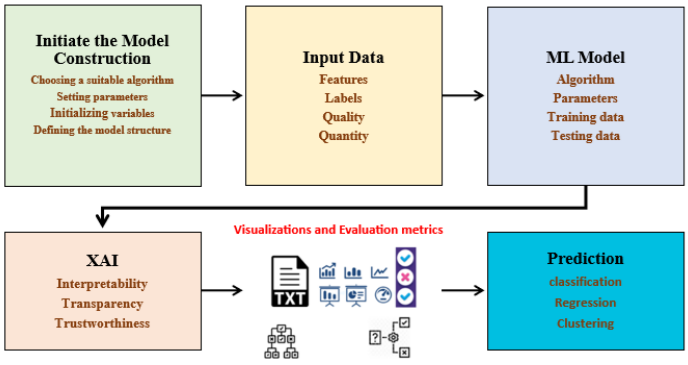

Fig. 3. The placement of XAI in AI model construction.

The National Institute of Standards (NIST) defines transparency, interpretability, explicability, and fairness as key principles for trustworthy AI [28]. These principles are essential for fostering user trust and ensuring AI systems operate ethically and accountable. They align with XAI, which employs techniques like decision trees and other methods to make AI reasoning more transparent for users [29]. By providing clear insights into how decisions are made, XAI helps demystify complex algorithms, enabling users to understand better and challenge AI outputs. This transparency is crucial for individual users, regulatory compliance, and societal acceptance of AI technologies, as illustrated in Fig. 4.

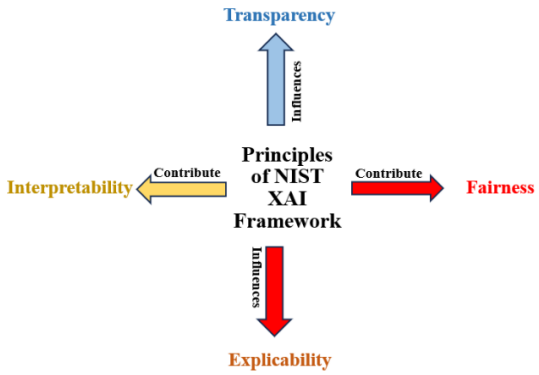

Fig. 4. NIST Key components of AI governance.

However, XAI faces challenges like computational cost, limited explanations, privacy concerns, and unclear explanations that hinder adoption [30]. A multi-disciplinary approach combining machine learning, human-computer interaction, psychology, and ethics is crucial. Balancing technical progress with societal needs and ethics is key to successful XAI [30]. This means developing transparent, interpretable, and ethical XAI solutions prioritizing fairness, accountability, and privacy. A multi-stakeholder approach involving users, experts, and marginalized groups is necessary. Only then can the potential of XAI for trusted and inclusive AI that benefits society be unlocked.

FMs like Equal Opportunity (EO) ensure fair and unbiased AI [31, 32]. These quantitative measures assess bias by examining how AI models treat different groups. For example, EO focuses on consistent True Positive Rates (correct identification of true positives) across groups. Other metrics like Demographic Parity (DP) ensure similar positive outcomes across groups, while Sufficiency (Calibration/Predictive Parity) checks for accurate positive probability predictions [47]. However, FMs face challenges. Defining and measuring them can be difficult; they might conflict with accuracy, and focusing solely on them can lead to a rigid fairness approach. Despite these limitations, FMs are essential for promoting fairness, transparency, and accountability in AI development. It must be recognized that FMs are not a standalone solution [47]. Over-reliance on them risks oversimplification, and manipulation is a potential concern. A balanced approach that considers FMs within a broader fairness definition and measurement framework is necessary. This collaborative effort, acknowledging trade-offs and integrating fairness throughout the AI lifecycle, is key to building more equitable and trustworthy AI systems.

AL tackles bias in AI by introducing cleverly crafted training data designed to fool the system and expose its weaknesses [48, 49]. These examples, like images or text, are false positives during training. By encountering these errors, AI models learn to become more robust against mistakes caused by biased data. AL offers benefits like improved accuracy and reduced cyberattack vulnerability [50]. However, it's not a silver bullet. While promising in improving AI robustness and reducing bias, AL's effectiveness depends on context. It can be computationally expensive and may not address all biases, particularly those rooted in deeper societal issues [51].

Moreover, concerns about overfitting and potential misuse exist. Therefore, AL is most effective as part of a broader strategy for fair AI, alongside continued research focused on responsible innovation and ethics [51]. This combined approach unlocks AL's potential while mitigating risks, allowing AI to harness adversarial training responsibly to create fairer and more trustworthy technologies.

AI systems require continuous M&E to prevent bias from creeping in and ensure ongoing fairness. Regular audits, testing, and user engagement are crucial to identifying and addressing potential biases [36]. FMs, like demographic parity and equal opportunity, can track how outcomes are distributed across different groups and highlight any disparities [36]. Human oversight, through approaches involving humans in the decision-making loop, can help review and intervene in potentially biased decisions [36].

However, challenges exist, including resource-intensive monitoring, requiring expertise and time to analyze data [42]. Biased or incomplete surveillance data can lead to inaccurate conclusions [37]. New biases that emerge over time or issues not present in the training data may also be missed [38]. The lack of transparency in some AI systems can make it challenging to identify bias [52].

Finally, the lack of transparency in some AI systems can make it challenging to identify bias [52], and privacy concerns arise when collecting and analyzing sensitive data [52].

V. Ethical and Responsible AI Management

Governments can promote the safe and responsible implementation of AI by establishing ethical guidelines and standards [54] [55] [56], consulting experts, engaging stakeholders, and formulating clear ethical principles, all of which monitor compliance and implement sanctions to build public trust and ensure the usefulness of AI. Ethics, in general, is a set of principles that help us distinguish between right and wrong. AI ethics is the process of studying how to improve the beneficial impact of AI while minimizing negative consequences. Rapid changes in AI systems raise profound ethical concerns by embedding biases, threatening human rights, and more.

Different international organizations [57] have issued some basic requirements for an AI system to be considered trustworthy. These organizations initially define values necessary for an ethical AI model and move beyond this towards an actionable policy. The most common requirements are explainability, reliability, robustness, security, accountability, privacy and data governance, human agency,

and oversight, legality, fairness, and safety, as shown in Fig. 5; all of these requirements should be evaluated throughout the life cycle of designing an AI model.

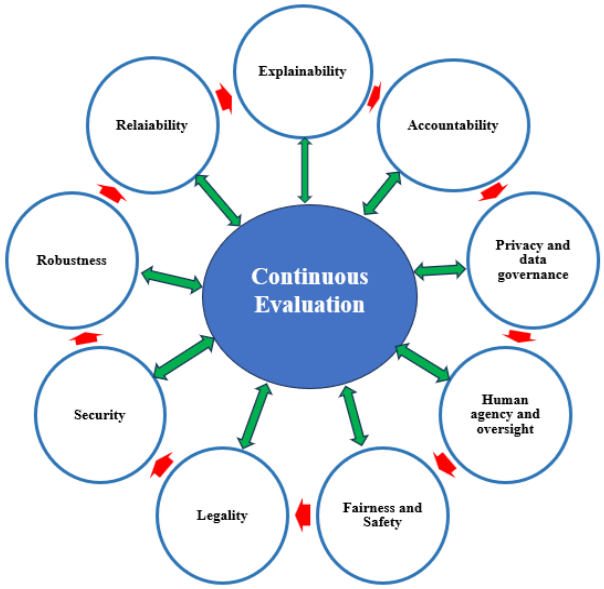

Fig. 5. Basic requirements for an AI model to be considered trustworthy.

Governments play a crucial role in shaping AI's responsible development and use. This includes establishing regulations to address data protection, privacy, and the use of AI in sensitive areas like healthcare and criminal justice. These regulations would focus on data protection, transparency, accountability, and ethical AI use, mitigating potential risks.

These regulations would focus on data protection, transparency, accountability, and ethical AI use, mitigating potential risks. Furthermore, governments can encourage industry self-regulation through voluntary codes of conduct and oversight mechanisms, potentially overseen by regulatory bodies that conduct audits. International collaboration can further solidify these efforts by establishing common standards.

However, responsible AI development shouldn't stifle innovation. Governments can promote AI advancement by investing in research and development, funding academic research, and fostering public-private partnerships. Additionally, supporting educational and training programs equips the workforce with the skills to navigate this evolving technological landscape [41]. This multifaceted approach ensures responsible AI development while fostering innovation and economic growth.

VI. Results

HAGF, presented in Fig. 6, offers a cyclical and interconnected approach to AI governance, emphasizing resource allocation and a multi-faceted strategy combining top-down policy development, bottom-up self-regulation, and continuous refinement. This comprehensive framework establishes a unified approach to AI governance, incorporating a coherent set of definitions and principles for responsible AI utilization. HAGF aligns with both model governance and rights-based models, focusing on ethical AI development, stakeholder engagement, and accountability. It provides a systematic approach to overseeing AI systems throughout their lifecycle.

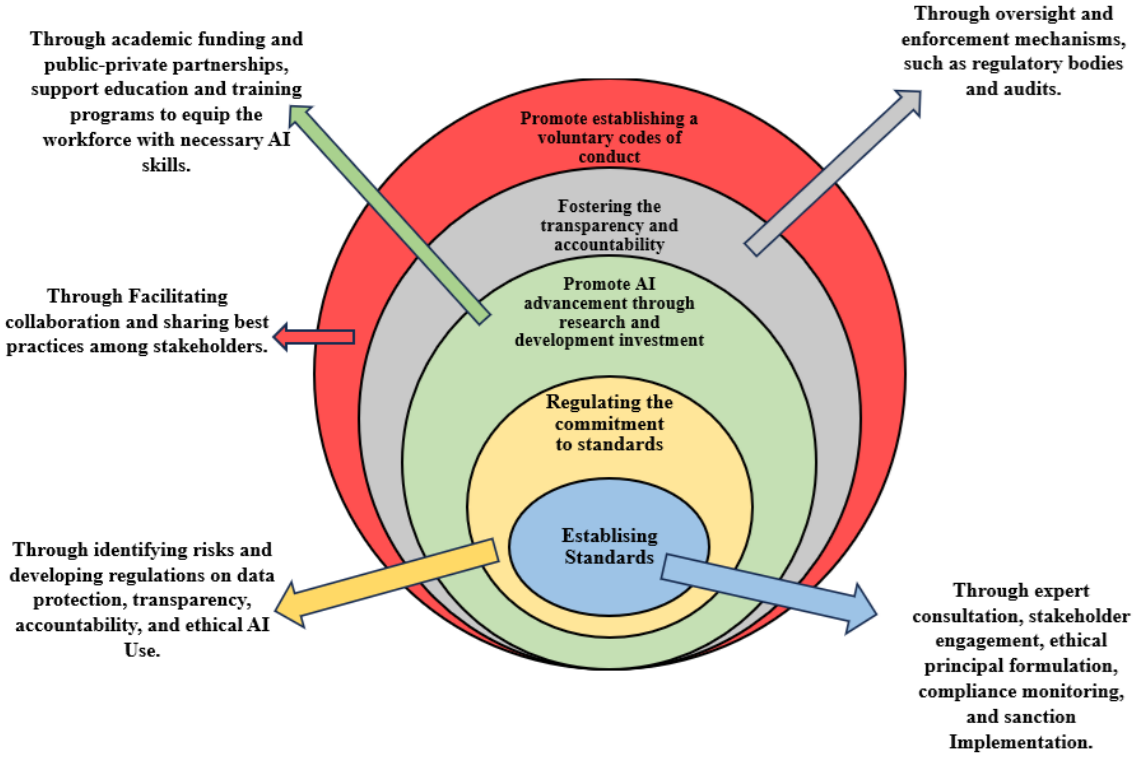

Fig. 6. HAGF – AI Governance framework.

The HAGF framework consists of five interconnected and cyclical components: establishing standards, regulating commitment to those standards, promoting AI advancement, fostering transparency and accountability, and encouraging a voluntary code of conduct. Each of these components plays a specific role in advancing ethical AI governance.

The phases are assigned theoretical weights, as illustrated in Fig. 7, to clearly convey the authors' prioritized vision for responsible AI development and deployment. These weights act as a powerful tool for clarifying the conceptual relationships between the components, providing strategic guidance for implementation, and stimulating discussion and debate regarding prioritization.

These assigned weights enhance the rigor and transparency of the HAGF framework, moving beyond abstract principles to deliver a concrete, actionable, and transparent representation of the authors' vision for ethical AI governance. By emphasizing these weights, the framework illustrates the importance of each component and fosters a deeper understanding of how they interconnect to promote a responsible AI ecosystem.

Phase 1 allocated 20%, focusing on developing a strong foundation for ethical guidelines for AI. This moderately resource-intensive phase involves forming a steering committee, conducting comprehensive research, formulating clear standards, and gathering feedback from diverse stakeholders. Notably, the sub-phase weights prioritize core values, with the highest allocation to formulating ethical guidelines (6%). Furthermore, stakeholder engagement (5%) was emphasized, recognizing the importance of engagement from all stakeholders. In addition, expert consultation (4%) informs the guidelines on best practices, while compliance monitoring (3%) and sanctions enforcement (2%) focus on developing the framework. It is important to note that implementation and enforcement will be addressed in subsequent phases.

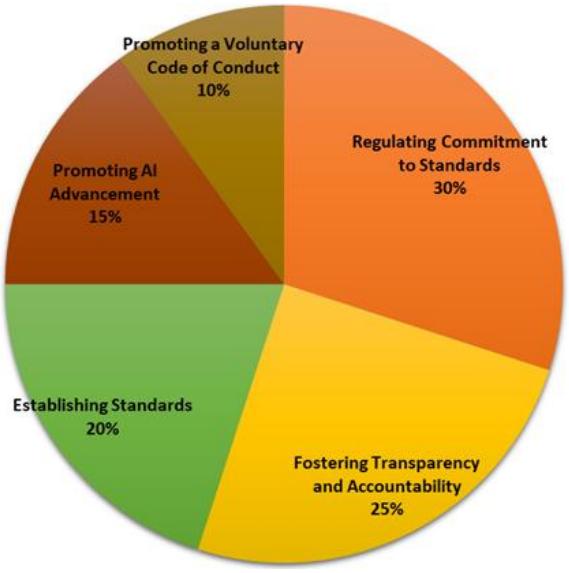

Fig. 7. HAGF phase weights.

The second phase has been assigned the highest weight, with 30% dedicated to translating ethical guidelines into action through robust compliance mechanisms. This resource-intensive phase emphasizes risk identification and the development of regulations for data protection, transparency, accountability, and ethical AI use. Due to the complexity of compliance, ongoing monitoring, regular audits, and potential legal action are necessary. Key focus areas in this phase include risk identification and assessment (9%), data protection regulations (8%), transparency and explainability regulations (7%), accountability mechanisms and sanctions (4%), and collaboration with authorities and dispute resolution (2%). This distribution effectively reflects the relative effort and significance of ensuring compliance with ethical AI standards.

15% has been allocated to the third phase, which aims to foster AI innovation through targeted investments. Notably, Funding Allocation (6%) is the primary focus, receiving the largest share, highlighting the importance of strategically distributing financial resources. In addition, Public-Private Partnerships (4%) leverage both resources and expertise, while Talent Development (3%) works to build a skilled workforce. Ethical Research Practices (1%) ensure that ethical standards are maintained throughout the research process, and Evaluation and Impact Assessment (1%) facilitate continuous learning and improvement. As a result, this distribution emphasizes funding while also recognizing the significance of partnerships, talent, ethics, and evaluation.

In the fourth phase, 25% of the weight is allocated, shifting the focus to building trust through responsible AI practices. Oversight and Enforcement Mechanisms (10%) are essential for ensuring compliance with established guidelines. Furthermore, Regular Audits (8%) assess adherence to these guidelines and pinpoint areas for improvement, while Public Disclosure (7%) enhances transparency by making information about AI systems accessible. This distribution prioritizes oversight and audits, highlighting their importance for accountability, and acknowledges the role of public disclosure in fostering trust.

Finally, 10% of the weighting is allocated to the fifth phase, which emphasizes encouraging self-regulation and best practices in AI. Facilitating Collaboration Among Stakeholders (5%) is vital for uniting diverse perspectives and fostering consensus. Additionally, Sharing Best Practices (3%) allows organizations to learn from each other and implement effective strategies. Promoting Inclusivity (2%) ensures that the codes encompass a broad range of viewpoints. Consequently, this distribution prioritizes collaboration and sharing best practices as essential drivers for developing and adopting meaningful voluntary codes of conduct.

A summary of the weighting assigned to each sub-phase within HAGF is provided in Fig. 8. This figure illustrates a project prioritizing compliance, risk management, and transparency, with Oversight & Enforcement (10%) and Risk Identification (9%) receiving the heaviest weight. Data Protection and Regular Audits (8% each) are also strongly emphasized. Transparency & Explainability (7%) and Ethical Principles and Funding (6% each) are moderately weighted. Stakeholder Engagement (5%), Accountability & Sanctions, Public-Private Partnerships, and Expert Consultation (4% each) are important but less emphasized. Supporting activities

include Talent Development (3%), Public Disclosure (3%), Collaboration (2%), Ethical Research (1%), Evaluation (1%), and Inclusivity (2%). The focus is on robust mechanisms and ongoing monitoring of ethical AI practices.

HAGF aligns with several emerging best practices in AI governance. Its emphasis on ethical development, stakeholder engagement, and accountability resonates with the principle of human-centered AI. The framework prioritizes transparency and accountability, addressing the growing demand for explainable AI (XAI). Furthermore, it implicitly considers risks at various stages of the AI lifecycle and acknowledges the importance of diverse stakeholder involvement, promoting a multi-stakeholder approach. Its cyclical nature facilitates continuous refinement, ensuring the framework remains relevant and effective.

Model Governance Frameworks, including HAGF, the Berkman Klein Center's framework, and the OECD AI Principles, provide blueprints for organizations seeking to implement responsible AI practices. These frameworks offer structured, principled approaches to governance, applicable across various contexts, and serve as models for navigating the complex challenges of ethical AI. While not legally binding or technically specific, they share a commitment to fostering ethical AI practices, each with unique attributes, as highlighted in Table II.

HAGF, in particular, offers significant ethical guidelines for AI development, emphasizing interconnectedness and resource allocation. It promotes stakeholder collaboration and underscores the importance of transparency and accountability, laying a solid foundation for responsible AI practices and positioning HAGF as a valuable contributor to the field.

HAGF framework offers significant advantages through its actionable implementation, breaking down ethical AI governance into specifically weighted sub-phases that prioritize critical areas and inform resource allocation. This explicit weighting compels strategic thinking, clarifying the relative importance of each component. Unlike other frameworks that provide broad guidance, HAGF delivers detailed direction, particularly in data protection, equipping organizations with concrete mechanisms for compliance. Its cyclical and adaptable nature fosters continuous improvement, ensuring relevance in a rapidly evolving landscape.

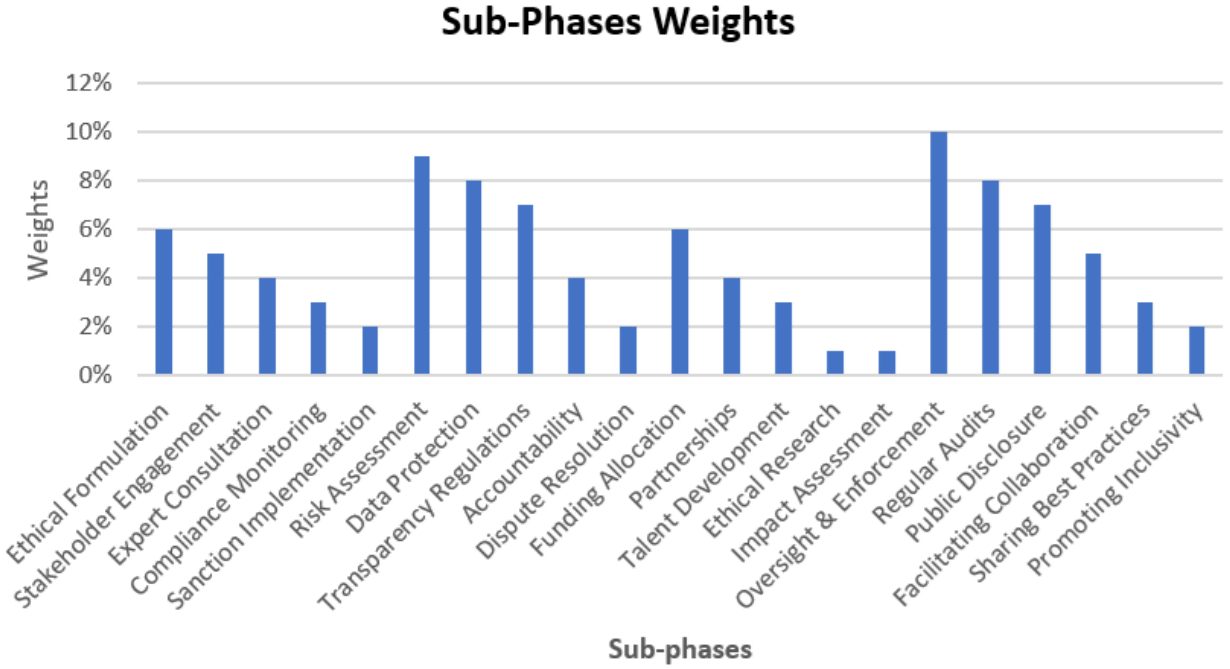

Fig. 8. Percentage weighting of sub-phases in HAGF.

TABLE II. COMPARING AI GOVERNANCE FRAMEWORKS: HAGF, OECD, AND BERMAN KLEIN

Framework	*Scope*	*Focus*	*Holistic Approach*	*Resource Allocation*	*Practical Focus*
HAGF Framework	Provides ethical guidelines for AI development and use, may offer more specific guidance	Interconnectedness and resource allocation.	Yes, it considers various interconnected components	Emphasizes the importance of funding and support	Provides a more concrete and actionable framework
OECD AI Principles	Provides general ethical guidelines for AI development and use	A general approach to ethical considerations	No	It does not explicitly address resource allocation	It gives more general principles
Berkman Klein Center's Model	Provides a comprehensive framework for AI governance	A holistic approach, considering various aspects of AI governance	Yes	It does not explicitly address resource allocation	Provides a general framework with guidance on implementation

However, HAGF has areas for potential improvement. It would strengthen its applicability by providing more detailed guidance on specific ethical issues, such as enhancing its focus on data governance with clear directives for responsible data practices. While its current complexity may pose challenges for some organizations, these issues are manageable and can be addressed through further refinement. HAGF's reliance on consistent funding and expertise highlights its potential for growth and adaptation. Measuring the effectiveness of its ethical guidelines and fostering public trust will enable it to evolve with rapid technological change.

Improvements in its enforcement mechanisms, particularly concerning voluntary codes of conduct, are also essential. As illustrated in Table III, a comprehensive set of KPIs has been developed to facilitate ongoing improvement by measuring the effectiveness of each HAGF phase and tracking progress toward ethical AI development. These KPIs cover critical aspects such as standards development and adoption, regulatory compliance, AI research investment, public trust, stakeholder engagement, and workforce impact. This clear mechanism for monitoring HAGF's performance helps identify areas for enhancement and ensures continuous progress in ethical AI practices, fostering accountability and transparency.

This holistic, cyclical approach, coupled with its emphasis on resource allocation and integration of top-down and bottom-up governance strategies, positions HAGF as an effective tool for promoting responsible AI development and use. Addressing the identified weaknesses and aligning more closely with best practices will further enhance HAGF's effectiveness in fostering ethical AI practices. Recognizing these limitations, as reflected in the insights from Table III, offers a balanced perspective on HAGF's potential and challenges.

TABLE III. KPIS FOR EVALUATING AN AI GOVERNANCE FRAMEWORK

Phase	***Key Performance Indicator (KPI)***
Establishing standards	- Number of AI-related standards and guidelines developed and published. - Level of stakeholder engagement in the standard-setting process. - Clarity and Comprehensibility of the established standards. - Adoption rate of the established standards by AI developers and users. - Number of revisions made to the standards. - Time taken to develop and publish the standards
Regulating the commitment to standards	- Compliance rate with established AI standards. - Number and severity of non-compliance incidents. - Effectiveness of enforcement mechanisms (e.g., audits, sanctions). - Level of public trust in the AI governance framework. - Number of AI-related regulations enacted. - Time taken to resolve non-compliance incidents. - Level of industry engagement in the regulatory process.
Promoting AI advancement through research and development investment	- Amount of funding allocated for AI research and development. - Number of successful AI research projects funded. - Number of new AI-based products or services developed. - Growth of the AI industry in revenue, employment, and innovation. - Collaboration rate between academia and industry on AI projects. - Number of AI-related patents filed. - Return on investment (ROI) for AI research funding.
Fostering transparency and accountability	- Number of AI systems with publicly available information on their algorithms and data sources. - Level of public trust and confidence in AI systems. - Number of successful cases of AI-related accountability measures (e.g., investigations, legal actions). - Frequency and quality of AI impact assessments. - Number of AI explainability tools or frameworks developed. - Number of AI bias detection and mitigation strategies implemented. - Level of public awareness and understanding of AI systems.
Promoting establishing a voluntary code of conduct.	- Number of organizations that have adopted voluntary codes of conduct for AI. - Level of adherence to the codes of conduct by organizations. - Positive impact of codes of conduct on ethical and responsible AI development. - Public perception of the effectiveness of voluntary codes of conduct. - Number of industry-led initiatives promoting responsible AI. - Level of diversity and inclusion in the development and governance of AI systems. - Number of self-assessments or audits conducted by organizations to ensure compliance with codes of conduct.

VII. CONCLUSION

Bias and discrimination in AI systems present significant challenges; however, emerging technologies such as Explainable AI (XAI), Fairness Metrics (FMs), and Adversarial Learning (AL), combined with robust governance frameworks, offer promising solutions. This article introduced the Holistic AI Governance Framework (HAGF), a comprehensive and cyclical approach to ethical AI governance. HAGF comprises five interconnected components: establishing standards, regulating adherence to those standards, promoting AI advancement, fostering transparency and accountability, and encouraging voluntary codes of conduct. By prioritizing resource allocation and integrating top-down policies with bottom-up self-regulation, HAGF addresses the complexities of

AI governance. Its emphasis on expert consultation, stakeholder engagement, and ethical principle formulation lays a strong foundation for effective standards. Additionally, focusing on risk identification, data protection, and accountability through regulation and oversight cultivates a culture of compliance. Investment in AI innovation and the promotion of voluntary codes of conduct supports both responsible practices and technological advancement. To ensure HAGF's effectiveness and adaptability in an evolving AI landscape, a robust suite of Key Performance Indicators (KPIs) is proposed, measuring standards, regulatory effectiveness, public trust, societal impact, and research advancement. Ongoing research, stakeholder engagement, and continuous evaluation and refinement of HAGF—particularly regarding data rights and algorithmic accountability—will be crucial for realizing a just and equitable AI future.

REFERENCES

[1] I. Castiglioni, L. Rundo, M. Codari, G. Di Leo, C. Salvatore, M. Interlenghi, et al., "AI applications to medical images: From machine learning to deep learning," Phys. Med., vol. 83, pp. 9–24, Mar. 2021. DOI: 10.1016/j.ejmp.2021.02.006

[2] C. Rudin, "Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, May 2019. DOI: 10.1038/s42256-019-0048-x

[3] X. Ferrer, T. Nuenen, J. M. Such, M. Cote, and N. Criado, "Bias and discrimination in AI: A cross-disciplinary perspective," IEEE Technol. Soc. Mag., vol. 40, no. 2, pp. 72–80, 2021. DOI: 10.1109/MTS.2021.3056293

[4] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," Humanit. Soc. Sci. Commun., vol. 10, no. 1, pp. 1–12, 2023. DOI: 10.1057/s41599-023-02079-x

[5] S. Reddy, S. Allan, S. Coghlan, and P. Cooper, "A governance model for the application of AI in health care," J. Am. Med. Inform. Assoc., vol. 27, no. 3, pp. 491–497, Mar. 1 2020. DOI: 10.1093/jamia/ocz192.

[6] Dankwa-Mullan, I., Ndoh, K., Akogo, D., Rocha, H. A. L., & Juaçaba, S. F. (2025). Artificial Intelligence and Cancer Health Equity: Bridging the Divide or Widening the Gap. *Current Oncology Reports*, 1-17.

[7] Prabhakar, P., Pati, P. B., & Parida, S. (2025). Navigating Legal and Ethical Dimensions in AI, IoT, and Cloud Solutions for Women's Safety. In *Developing AI, IoT and Cloud Computing-based Tools and Applications for Women's Safety* (pp. 155-191). Chapman and Hall/CRC.

[8] D. F. Engstrom and A. Haim, "Regulating Government AI and the Challenge of Sociotechnical Design," Annu. Rev. Law Soc. Sci., vol. 19, no. 1, p. 19, 2023. DOI: 10.1146/annurev-lawsocsci-120522-091626

[9] Da Mota, M. (2024). Toward an AI Policy Framework for Research Institutions. *Artificial Intelligence*.

[10] Fu, Y., & Weng, Z. (2024). Navigating the ethical terrain of AI in education: A systematic review on framing responsible human-centered AI practices. *Computers and Education: Artificial Intelligence*, 100306.

[11] Papagiannidis, E. (2024). Responsible AI governance in practice: The strategic impact of responsible AI governance on business value and competitiveness.

[12] Khan, M. A. (2024). *Responsible AI Principles in Product Management Practices* (Doctoral dissertation, PhD thesis, Aalborg University).

[13] Challoumis, C. (2024, October). Building a sustainable economy-how ai can optimize resource allocation. In *XVI International Scientific Conference* (pp. 190-224).

[14] Constantinides, M., Bogucka, E., Quercia, D., Kallio, S., & Tahaei, M. (2024). RAI guidelines: Method for generating responsible AI guidelines grounded in regulations and usable by (non-) technical roles. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1-28.

[15] Leslie, D., & Perini, A. M. (2024). Future Shock: Generative AI and the international AI policy and governance crisis.

[16] Kaminski, M. E., & Malgieri, G. (2024). Impacted Stakeholder Participation in AI and Data Governance. *Forthcoming (2024-25)*.

[17] Lee, S. U., Perera, H., Liu, Y., Xia, B., Lu, Q., Zhu, L., ... & Whittle, J. (2024). Responsible AI Question Bank: A Comprehensive Tool for AI Risk Assessment. *arXiv preprint arXiv:2408.11820*.

[18] Porter, Z., Habli, I., McDermid, J., & Kaas, M. (2024). A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI and Ethics*, *4*(2), 593-616.

[19] A. Ahmad, et al., "Equity and Artificial Intelligence in Surgical Care: A Comprehensive Review of Current Challenges and Promising Solutions," BULLET: Jurnal Multidisiplin Ilmu, vol. 2, no. 2, pp. 443–455, 2023.

[20] Paproki, A., Salvado, O., & Fookes, C. (2024). Synthetic Data for Deep Learning in Computer Vision & Medical Imaging: A Means to Reduce Data Bias. *ACM Computing Surveys*.

[21] J. S. Park, et al., "Understanding the representation and representativeness of age in AI data sets." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021. DOI: 10.1145/3461702.3462590

[22] Zhang, Kuan, et al. "Mitigating bias in radiology machine learning: 2. Model development." Radiology: Artificial Intelligence 4.5 (2022): e220010. DOI: 10.1148/ryai.220010

[23] Gunning, David, et al. "XAI—Explainable artificial intelligence." Science robotics 4.37 (2019): eaay7120. DOI: 10.1126/scirobotics.aay7120.

[24] Landers, Richard N., and Tara S. Behrend. "Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models." American Psychologist 78.1 (2023): 36. DOI: 10.1037/amp0000972

[25] Karim, Md Rezaul, et al. "Explainable AI for Bioinformatics: Methods, Tools and Applications." Briefings in Bioinformatics 24.5 (2023): bbad236. DOI: 10.1093/bib/bbad236

[26] S. Akhai, "From Black Boxes to Transparent Machines: The Quest for Explainable AI." Available at SSRN 4390887 (2023). DOI: 10.2139/ssrn.4390887

[27] Eke, C. I., & Shuib, L. (2024). The role of explainability and transparency in fostering trust in AI healthcare systems: a systematic literature review, open issues and potential solutions. *Neural Computing and Applications*, 1-36.

[28] Radanliev, P. (2025). AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. *Applied Artificial Intelligence*, *39*(1), 2463722.

[29] Kostopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. *Electronics*, *13*(14), 2842.

[30] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., ... & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301.

[31] Schoenherr, Jordan Richard, et al. "Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness." *IEEE Transactions on Technology and Society* 4.1 (2023): 9-23.

[32] P. Phillips, C. Hahn, P. Fontana, A. Yates, K. Greene, D. Broniatowski, et al., Four Principles of Explainable Artificial Intelligence, NIST Interagency/Internal Report (NISTIR). Gaithersburg, MD: National Institute of Standards and Technology, 2021 [online], https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399. Accessed Jul. 17, 2023., DOI: 10.6028/NIST.IR.8312.

[33] Bellamy, Rachel KE, et al. "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias." arXiv preprint arXiv: (2018).

[34] C. Dwork, et al., "Fairness through awareness." Proceedings of the 3rd innovations in theoretical computer science conference. 2012. DOI: 10.1145/2090236.2090255

[35] Almajali, M. H., Nasrawin, L., Alqudah, F. T., Althunibat, A. A., & Albalawee, N. (2023). Technical Service Error as a Pillar of

Administrative Responsibility for Artificial Intelligence (AI) Operations. *International Journal of Advances in Soft Computing & Its Applications, 15*(3).

[36] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. PMLR, 2018.

[37] Cascella, Marco, et al. "Crossing the AI Chasm in Neurocritical Care." Computers 12.4 (2023): 83. DOI: 10.3390/computers12040083

[38] J. S. Brownstein, B. Rader, C. M. Astley, and H. Tian, "Advances in Artificial Intelligence for infectious-disease surveillance," N. Engl. J. Med., vol. 388, no. 17, pp. 1597–1607, Apr. 27 2023. DOI: 10.1056/NEJMra2119215

[39] Qatawneh, A. M., Lutfi, A., & Al Barrak, T. (2024). Effect of Artificial Intelligence (AI) on Financial Decision-Making: Mediating Role of Financial Technologies (Fin-Tech). HighTech and Innovation Journal, 5(3), 759-773.

[40] Abu Afifa, M. M., Nguyen, T. H., Le, M. T. T., Nguyen, L., & Tran, T. T. H. (2024). Accounting going digital: a Vietnamese experimental study on artificial intelligence in accounting. *VINE Journal of Information and Knowledge Management Systems*.

[41] D. Schiff, "Education for AI, not AI for Education: The Role of Education and Ethics in National AI Policy Strategies," Int. J. Artif. Intell. Educ., vol. 32, no. 3, pp. 527–563, 2022., doi:DOI: 10.1007/s40593-021-00270-2

[42] G. Said, et al., "Adapting Legal Systems to the Development of Artificial Intelligence: Solving the Global Problem of AI in Judicial Processes," International Journal of Cyber Law, vol. 1, p. 4, 2023.

[43] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, et al., "Scientific discovery in the age of artificial intelligence," Nature, vol. 620, no. 7972, pp. 47–60, Aug. 2023. DOI: 10.1038/s41586-023-06221-2

[44] Y. Chen, E. W. Clayton, L. L. Novak, S. Anders, and B. Malin, "Human-centered design to address biases in artificial intelligence," J. Med. Internet Res., vol. 25, p. e43251, Mar. 24 2023. DOI: 10.2196/43251

[45] D. Hartmann et al., "Addressing the Regulatory Gap: Moving Towards an EU AI Audit Ecosystem Beyond the AIA by Including Civil Society," arXiv preprint arXiv:2403.07904, 2024.

[46] M. Roshanaei, "Towards best practices for mitigating artificial intelligence implicit bias in shaping diversity, inclusion and equity in higher education," Educ. Inf. Technol., pp. 1–26, 2024.

[47] N. Zhou, Z. Zhang, V. N. Nair, H. Singhal, and J. Chen, "Bias, fairness and accountability with artificial intelligence and machine learning algorithms," Int. Stat. Rev., vol. 90, no. 3, pp. 468–480, 2022. DOI: 10.1111/insr.12492

[48] Pagano, Tiago P., et al. "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods." Big data and cognitive computing 7.1 (2023): 15. DOI: 10.3390/bdcc7010015.

[49] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples." arXiv preprint arXiv: (2014).

[50] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018. DOI: 10.1145/3278721.3278779

[51] D. Madras, et al., "Learning adversarially fair and transferable representations." International Conference on Machine Learning. PMLR, 2018.

[52] Nwakanma, Cosmas Ifeanyi, et al. "Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review." Applied Sciences 13.3 (2023): 1252. DOI: 10.3390/app13031252

[53] Dhirani, Lubna Luxmi, et al. "Ethical dilemmas and privacy issues in emerging technologies: a review." Sensors 23.3 (2023): 1151. DOI: 10.3390/s23031151

[54] J. Cahill, V. Howard, Y. Huang, J. Ye, S. Ralph, and A. Dillon, "Intelligent Work: Person Centered Operations, Worker Wellness and the Triple Bottom Line," Commun. Comput. Inf. Sci., vol. 1421, pp. 307–314, 2021., doi:DOI: 10.1007/978-3-030-78645-8_38

[55] M. Cole, C. Cant, F. Ustek Spilda, and M. Graham, "Politics by Automatic Means? A Critique of Artificial Intelligence Ethics at Work," Front. Artif. Intell., vol. 5, p. 869114, Jul. 15 2022., doi:DOI: 10.3389/frai.2022.869114

[56] "Ethics of Artificial Intelligence." UNESCO, 26 Sept. 2023, www.unesco.org/en/artificial-intelligence/recommendation-ethics.

[57] "Ethics Guidelines for Trustworthy AI." Shaping Europe's Digital Future, 8 Apr. 2019, digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.