System Card+: Responsible AI Framework for Decision Support Systems

Haileleol Tibebu and Ioannis A. Kakadiaris

Computational Biomedicine Lab
Department of Computer Science
University of Houston
Houston, TX, USA
htibebu@illinois.edu,ioannisk@uh.edu

Abstract. Artificial Intelligence's increasing complexity and societal impact demand a robust performance, fairness, inclusivity, and ethical and legal assessment framework. Current approaches often lack the adaptability and context-specific focus to establish responsible AI use. This paper introduces a theoretical framework for AI-based Decision Support Systems, a modular approach designed to address these limitations and bridge the gap between complex AI regulations and their practical implementation. The methodology offers a five-layer benchmarking system. The first layer verifies the performance accountability of the AI system, while the second layer evaluates its fairness. The third layer addresses inclusivity, the fourth examines ethical competency, and the fifth assesses legal compliance. Each layer evaluates AI-based decision support systems throughout the AI life cycle's development, assessment, mitigation, and assurance stages, including data, models, code, and the system. This decomposed, step-by-step approach simplifies compliance efforts, aligning with major international regulations such as the EU AI Act, the NIST AI framework, and the Algorithmic Accountability Act. This work contributes to developing a universal standard for responsible AI practices.

Key words: System-card+, Responsible AI, AI Accountability, Fairness

1 Introduction

Artificial Intelligence (AI) is transforming decision-making across diverse sectors such as healthcare [1, 2], finance [3], and criminal justice [4]. Although these advances hold promise, they also present significant challenges, such as biases, lack of transparency, and concerns over data privacy [5, 6]. Such challenges threaten to erode public trust, amplify societal inequalities, and compromise fundamental human rights [7, 8]. To mitigate these issues, it is essential to embed accountability, fairness, inclusivity, and adherence to ethical and legal standards into AI system development and deployment [9]. Despite ongoing efforts to integrate these principles, there remain considerable gaps in their practical application, which could potentially widen social disparities.

The expansion of AI ethics frameworks signals the importance of this field but also highlights a significant gap between frameworks and their practical implementation [10, 11]. This gap often results in superficial accountability measures, complicating the application of uniform standards across various AI applications, from low-risk scenarios, such as image tagging, to high-stakes contexts [12, 13]. Focusing solely on technical metrics such as model accuracy or data quality can neglect the broader socio-technical contexts, potentially leading AI systems to perpetuate biases, create power imbalances, or infringe on rights [14, 15]. Moreover, the sector-driven nature of AI ethics initiatives has led to inconsistent standards and complicates the creation of a universal ethical foundation [16].

System Card+ is motivated by the original System Card framework [17], which includes 56 criteria and proposes a five-layer structure: accountability for performance, fairness, inclusion, ethical and legal compliance, resulting in a total of 191 evaluation criteria. This separation was motivated by the need to address the complexity and breadth of responsible AI evaluation, covering distinct but interdependent domains such as fairness, inclusivity, ethics, and legal compliance. The limited scope of the original system card framework made it difficult to conduct detailed assessments for each area. By adding layers to the framework, we provide a focused approach that allows stakeholders to conduct domain-specific in-depth evaluations while maintaining the integrity of the overall framework. This approach provides robust guidelines to help industries and researchers develop, deploy, and manage AI technologies responsibly.

The remainder of the paper is structured as follows: Section 2 provides a literature review, reviewing existing AI ethics and governance frameworks. Section 3 describes the System Card+ framework and outlines its components and application. Section 4 discusses our approach's strengths, challenges, and future directions. Section 6 concludes the paper by summarizing the framework's contributions to responsible AI development.

2 Literature Review

Accountable AI systems are essential for public trust and risk mitigation early in the AI life cycle. Kroll *et al.* [18] introduced frameworks for algorithmic transparency and auditing, while Raji *et al.* [19] emphasized the need for ongoing internal algorithmic auditing. Cath *et al.* [20] discussed technical challenges and the necessity for regular audits to uphold public trust. Our prior work [17] developed a performance accountability framework for Decision Support Systems, recognizing risks like bias and opaqueness. It includes 56 detailed criteria covering four dimensions—data, model, code, and system—analyzed through development, assessment, mitigation, and assurance.

Fairness in AI systems is a multidimensional concept involving eliminating bias and ensuring equitable treatment across different demographics. Mehrabi *et al.* [21] reviewed machine learning biases and mitigation strategies. Noble *et al.* [22] and Eubanks *et al.* [23] highlighted how algorithms could reinforce social

inequalities. Other works [24, 25] stress that fairness extends beyond technical solutions, requiring an interdisciplinary approach. Inclusivity demands that AI systems be accessible to all societal sectors, embedding diverse cultures and perspectives. Langdon *et al.* [26] works signify the role of inclusive data frameworks and participatory design. Burnett *et al.* [27] propose a method for diverse representation in data collection. Schlesinger *et al.* [28] emphasized how participatory methods can enhance AI inclusivity. Additionally, Ochmann *et al.* [29] and Zheng *et al.* [30] propose empirical solutions that explore transparency and anthropomorphism interventions as strategies to shape fairness perceptions actively. They also assess the correlation between coverage criteria and the fairness of deep neural networks.

Ethical AI must respect human dignity and societal values. Floridi and Cowls [31] presented a unified ethics framework, while Mittelstadt *et al.* [32] discussed the ethical challenges of transparency. Jobin *et al.* [33] and Binns [34] surveyed global ethics guidelines and the philosophical underpinnings of machine learning biases, respectively. Cath *et al.* [20] and Moor [35] also contributed to discussions on AI governance and ethical reflections. The legal regulation of AI systems is rapidly evolving, with various jurisdictions attempting to address the unique challenges posed by Voigt *et al.* [36] provided a detailed analysis of the EU General Data Protection Regulation (GDPR) and its implications. The author emphasizes data protection and user consent. Madiega *et al.* [37] discussed the EU AI Act by the European Parliament, aiming to establish comprehensive legal standards for AI development and deployment. Mittelstadt *et al.* [38] highlighted principles' limitations in ensuring ethical AI and for more concrete legal and regulatory measures.

Despite notable advancements in AI assessment methodologies, significant gaps persist, including the absence of a universal and adaptable framework. Stahl *et al.* [39] and Floridi [40] emphasized the need for AI frameworks that maintain flexibility across diverse AI applications and evolving technologies. Sector-specific challenges, such as those identified by Hagendorff [11] in healthcare and Zhang *et al.* [41] in finance, highlight the limitations of generic guidelines. Jobin *et al.* [9] and Veale and Brass [42] critiqued the static nature of many frameworks, underscoring the necessity for dynamic, flexible ethical guidelines. This fragmented approach complicates implementation for companies and evaluation by policymakers, revealing a critical need for more integrated and adaptable AI governance frameworks; our theoretical framework aims to address these gaps.

We propose a five-layered accountability methodology for AI-based decision support systems to address this significant gap. Our framework starts by assessing the general requirements of an AI system, followed by an independent evaluation of fairness, inclusivity, ethical considerations, and legal compliance across four key dimensions: data, model, code, and system. Each dimension is further examined through stages of development, assessment, mitigation, and assurance. The combination of layers verifies a dynamic and context-sensitive implementation of a responsible AI-based decision system.

3 Methodology

The System Card+ framework is a structured five-layer methodology comprising 191 criteria for comprehensively evaluating the AI system. Although this study provides an overview of the framework's structure, description, and objectives, evaluation methods for each criterion will be developed in subsequent phases of the project.

3.1 Data Sources

We use a combination of primary and secondary data sources to establish a knowledge base for framework development. We conducted semi-structured discussions with focus groups and key stakeholders, including AI developers, legal experts, public policy experts, and representatives from marginalized communities. These primary sources provided insights into ethical AI development's practical challenges and requirements. We also reviewed the literature, systematically analyzing existing frameworks, guidelines, and case studies across diverse industries and geographical regions. Our sources included peer-reviewed academic journals, industry white papers, and policy documents. Our previous single-layer system card is presented in Table 1.

Table 1. The Original Single Layer System Cards for AI-Based Decision Support Systems

Category	Development		Assessment		Mitigation		Assurance	
Data	Data Dictionary	(C111)	Privacy, Data	(C211)	Anonymization	(C311)	Data Protection	(C411)
	Datasheet, Collection Process	(C112)	Fairness, Data	(C212)	Security	(C312)	Datasheet, Maintenance	(C412)
	Datasheet, Composition	(C113)	Quality, Labels	(C213)	Datasheet, Uses	(C413)		
	Datasheet, Motivation	(C114)	Inspectability	(C214)				
	Datasheet, Preprocessing	(C115)						
Model	Reproducibility, Model	(C121)	Interpretability	(C221)	Adversarial Training	(C321)	Privacy, Model	(C421)
	Design Transparency, Model	(C122)	Fairness, Model	(C222)	Explanations, Mitigation	(C322)	Uses, Model	(C422)
	Documentation, Model	(C123)	Testing, Adversarial	(C223)	Fairness, Mitigation	(C323)	Documentation, Capabilities	(C423)
	Selection, Model	(C124)	Privacy, Training	(C324)	Explainability	(C424)		
Code	Reproducibility, Code	(C131)	Privacy, Code	(C231)	Review, Code	(C331)	Certification, Developer	(C431)
	Design Transparency, Code	(C132)	Security, Code	(C232)	Diversity, Team	(C332)	Due Diligence	(C432)
	Documentation, Code	(C133)	Testing Cards	(C233)				
System	Documentation, Development	(C141)	Awareness, Public	(C241)	Monitoring, Fairness	(C341)	Record Keeping, Operational	(C441)
	Plans, Maintenance	(C142)	Risk, Humans	(C242)	Monitoring, Performance	(C342)	Uses, System	(C442)
			Training, Operator	(C243)	Oversight, Human	(C343)	Documentation, Acceptability	(C443)
			Accuracy, System	(C244)	Harms, Remedies	(C344)	Insurance	(C444)
					Mechanism, Feedback	(C345)	Rating, Risk	(C445)
					Security	(C346)		

3.2 Criteria for Item Inclusion

The items included in our framework were selected based on the following criteria:

Relevance: Each item was evaluated for its pertinence to accountability, fairness, inclusively, ethics, and legal compliance principles. We verify that the items address AI development's most pressing ethical and legal challenges. We also provide supporting academic literature for the framework.
Evidence-Based: We prioritized items supported by empirical evidence or validated through case studies. An empirical foundation was prioritized by integrating findings from peer-reviewed studies, industry reports, and validated case studies. We used quantitative and qualitative data from real-world applications to validate each item's efficacy and maintain an ongoing review of the latest research to update and refine the framework based on new empirical findings.

3.3 Framework Development

To establish the framework for AI systems, we identified five critical dimensions: performance accountability, fairness, inclusivity, ethics, and legal compliance. These dimensions were chosen based on their relevance to the entire life cycle of AI systems. Each dimension focuses on specific aspects of automated decision support systems. These dimensions are selected based on their ability to address the holistic accountability challenges inherent in AI systems. Each dimension is further structured into four operational phases: Development, Assessment, Mitigation, and Assurance. These phases are applied across four key components: Data, Model, Code, and System. Core ethics and operational requirements were identified through extensive literature reviews and multi-discipline stakeholder consultations. We mapped vital challenges and requirements to the appropriate dimensions using this process.

The performance accountability layer provides a baseline detection mechanism that checks the minimum requirements of the AI system. This is the foundational layer to check that all subsequent fairness, inclusivity, ethics, and legal compliance are built upon a solid accountability ground. This layer is essential for the early identification of potential issues and establishing a standard for monitoring and reporting. It spans Data, Model, Code, and System components across the Development, Assessment, Mitigation, and Assurance phases. The fairness layer addresses the need to detect and mitigate biases in AI systems to promote equitable outcomes. It is essential to prevent discrimination and ensure fair treatment across all user groups. Each phase includes strategies for analyzing data for biases, correcting biases in AI models, evaluating code for fairness, and ensuring the system performs equitably under training and testing conditions. The inclusivity layer verifies that AI systems are accessible and beneficial to diverse user groups. This dimension addresses the need for comprehensive data representation and user-centric design practices. Each layer includes actions to confirm datasets are inclusive, validate models for multicultural considerations and accessible design, and verify system-level inclusivity. The ethical integrity layer embeds ethical principles into AI operations and evaluates AI systems aligned with societal values and ethical norms. Each layer includes measures to secure ethical sourcing and consent management in data, ethical design, impact analysis in models, adherence to ethical coding standards, and regular ethical

audits at the system level. The legal compliance layer checks systems' compliance with laws and regulations. This dimension is critical for avoiding legal pitfalls and following regulatory compliance. Each layer includes actions to verify data licenses and regulatory adherence, review legal compliance in models, audit code licenses and export compliance, and verify system-level regulatory submissions and privacy compliance. The following section presents the proposed five-layer AI accountability framework for an AI-based decision support system: performance, fairness, inclusively, ethics, and legal compliance.

3.4 Performance Accountability Layer for AI-Based Decision Support Systems

The **Performance Accountability layer**, as presented in Table 2, provides a structured approach to secure responsible and transparent deployment of AI systems. It delineates clear protocols and checks across the entire AI life cycle, encompassing the development, assessment, mitigation, and assurance phases, thus ensuring that each phase rigorously evaluates and meets high accountability standards.

Development: The *Data Dictionary* (A111) evaluates the clarity and accuracy of data terminology, ensuring all participants have a unified understanding of the data elements used. The *Datasheet for Collection Process* (A112) measures the transparency and appropriateness of the methods and environments from which data is collected. The *Datasheet for Composition* (A113) evaluates the diversity and representativeness of data sets, ensuring that they are free from bias and inclusion. The *Datasheet for Motivation* (A114) assesses the alignment of data collection objectives with the project goals, verifying the rationale behind the data needs. The *Datasheet for Preprocessing* (A115) measures the adequacy and correctness of data preparation techniques, including cleaning and normalization processes. The reproducibility of the *Model* (A121) evaluates consistency in model outputs under identical scenarios, which is crucial for validation. The *Design Transparency* (A122) assesses the openness of model architecture, facilitating stakeholder understanding and trust. **Model Documentation** (A123) evaluates the comprehensiveness of the information provided about the model, including its purpose, assumptions, and limitations. The *Model Selection* (A124) process measures strategic alignment with ethical considerations and project objectives. *Code Reproducibility* (A131) evaluates the consistency of the software code in producing the same results under the same conditions. *Design Transparency of Code* (A132) measures the clarity of the code's documentation, ensuring it is understandable and maintainable. *Code Documentation* (A133) assesses the detail and clarity in describing the code's functionalities and limitations. *Documentation for System* (A141) evaluates the thoroughness of documenting the system's design and development process. *Maintenance Plans* (A142) assess the robustness of strategies for the system's long-term health and functionality.

Table 2. Performance Accountability Layer for AI-Based Decision Support Systems

Category	Development	Assessment	Mitigation	Assurance
Data	Data Dictionary	(A111) Inspectability	(A211) Anonymization	(A311) Data Protection (A411)
	Datasheet, Collection Process	(A112)		Datasheet, Maintenance (A412)
	Datasheet, Composition	(A113)		Datasheet, Uses (A413)
	Datasheet, Motivation	(A114)		
	Datasheet, Preprocessing	(A115)		
Model	Reproducibility, Model	(A121) Interpretability	(A221) Adversarial Training	(A321) Uses, Model (A421)
	Design Transparency, Model	(A122) Testing, Adversarial	(A223) Explanations, Mitigation	(A322) Documentation, Capabilities (A423)
	Documentation, Model	(A123)		Explainability (A424)
	Selection, Model	(A124)		
Code	Reproducibility, Code	(A131) Testing Cards	(A231) Review, Code	(A331) Certification, Developer (A431)
	Design Transparency, Code	(A132) Compliance Review	(A232) Diversity, Team	(A332) Due Diligence (A432)
	Documentation, Code	(A133)		
System	Documentation, System	(A141) Training, Operator	(A241) Oversight, Human	(A341) Record Keeping, Operational (A441)
	Plans, Maintenance	(A142)	Mechanism, Feedback	(A345) Uses, System (A442)
		Security	(A346)	Documentation, Acceptability (A443)
				Insurance (A444)
				Rating, Risk (A445)

Assessment: *Inspectability* (A211) measures the system's capability to examine data and algorithms, ensuring transparency and accountability effectively. *Interpretability of Model* (A221) evaluates the ability of stakeholders to understand and rationalize the model's decision-making processes. *Adversarial Testing of Model* (A223) assesses the model's resilience against inputs designed to challenge or deceive it, measuring its robustness and security. *Testing Cards for Code* (A231) evaluate the code's performance under various conditions to identify potential issues. *Compliance Review of Code* (A232) measures adherence to regulatory and ethical standards, ensuring legal compliance. *Operator Training* (A241) evaluates the effectiveness of training provided to operators, ensuring they can manage and interact with the system correctly. *System Security* (A346) measures the strength of the security measures to protect the system from unauthorized access and cyber threats.

Mitigation: *Data Anonymization* (A311) evaluates the effectiveness of techniques to protect personal information, ensuring privacy and compliance with data protection laws. *Adversarial Training* (A321) measures how well the model can handle unexpected or extreme inputs, enhancing its performance and reliability. *Mitigation Explanations* (A322) assess the clarity and adequacy of the explanations provided for actions taken to mitigate identified risks. *Code Review* (A331) evaluates the code for security vulnerabilities and operational efficiency, ensuring robust and secure software. *Team Diversity* (A332) is assessed to measure its impact on reducing biases in the development process. *Human Oversight* (A341) evaluates the effectiveness of human monitoring and control mechanisms over the AI system. *Feedback Mechanisms* (A345) are measured for their effectiveness in capturing user feedback and enabling continuous improvement.

Assurance: *Data Protection* (A411) measures the effectiveness of implemented security protocols and access controls in safeguarding data. *Maintenance of Datasheets* (A412) and *Uses of Datasheets* (A413) verify continuous accuracy and relevance of data documentation. *Model Uses* (A421) assesses to affirm the model's applications are as intended and comply with design specifications. *Documentation of Model Capabilities* (A423) evaluates the accuracy and completeness of information regarding what the model can do. *Explainability of Model* (A424) measures how well users can understand the model's reasoning. *Devel-*

oper Certification (A431) assesses the qualification and expertise of developers against established standards. *Due Diligence for Code* (A432) measures the thoroughness of evaluations conducted to assure the code's quality and security standards. *Operational Record Keeping* (A441), *Documentation of System Uses* (A442), and *Documentation of System Acceptability* (A443) are evaluated for their comprehensiveness and adherence to operational and ethical standards. *Insurance* (A444) and *Risk Rating* (A445) measure the management and mitigation of financial and operational risks associated with the system.

3.5 Fairness layer for AI-Based Decision Support Systems

The Fairness layer for AI-Based Decision Support Systems establishes fairness throughout the AI system's life cycle. This layer helps identify and mitigate biases within AI systems to uphold equitable outcomes across all user groups. Table 3 summarizes the fairness layer.

Development: *Equity in Data* (B111) evaluates the equality of data collection practices to confirm that no group is unfairly favored or disadvantaged. *Bias Detection in Data* (B112) measures potential biases within the dataset, allowing for early identification and correction. *Sampling Integrity* (B113) inspect the methods used to gather data, ensuring they represent the target population accurately and fairly. *Fairness Metrics for Data* (B114) evaluate the fairness levels within the data to ensure they meet predefined fairness standards. *Fairness-by-Design for Models* (B121) inspect the fairness principles used in the model design phase to address potential issues proactively. *Bias Detection in Models* (B122) measures the mechanism that detects potential biases during model training and operation, ensuring they are identified and mitigated. *Fairness Benchmarking for Models* (B123) inspect the standards to continuously evaluate the model's fairness, promoting adherence to fairness criteria.
Assessment: *Fairness Assessment of Data* (B211) evaluates the data for fairness, ensuring no discriminatory biases are present. *Impact Assessment for Data* (B212) analyzes the potential effects of data handling and processing decisions on different demographic groups, checking for unintended discriminatory impact. *Fairness Metrics for Models* (B221) measures how models perform against established fairness benchmarks throughout their life cycle. *Sensitivity Analysis for Models* (B222) tests models against various inputs to identify conditions under which the model's outputs may be unfairly biased, ensuring robustness and fairness.

Table 3. Fairness Layer for AI-Based Decision Support Systems

Category	Development	Assessment	Mitigation	Assurance
Data	Equity, Data	(B111) Fairness, Data	(B211) Bias Mitigation, Data	(B311) Third-Party Assessment, Data (B411)
	Bias Detection, Data	(B112) Impact Assessment, Data	(B212) Rebalancing Techniques	(B312) Transparency Reports, Data (B412)
	Sampling Integrity, Data	(B113)		
	Fairness Metrics, Data	(B114)		
Model	Fairness-by-Design, Model	(B121) Fairness Metrics, Model	(B221) Bias Mitigation, Model	(B321) Third-Party Assessment, Model (B421)
	Bias Detection, Model	(B122) Sensitivity Analysis, Model	(B222) Fair Optimization, Model	(B322) Transparency Reports, Model (B422)
	Fairness Benchmarking, Model	(B123)		
Code	Automated Testing, Code	(B131) Compliance Review, Code	(B231) Fairness Refactoring, Code	(B331) Compliance Certificates, Code (B431)
	Audit Trails, Code	(B132) Fairness Test, Code	(B232) Equity Enhancements, Code	(B332) Transparency Reports, Code (B432)
System	Alert, System	(B141) User Feedback, System	(B241) Operation Protocols, System	(B341) Assurance Policies, System (B441)
	Stress Tests, System	(B142) Demographic Performance, System	(B242) Calibration, System	(B342) Fairness Logs, System (B442)
	Impact Assessments, System	(B143)		

Mitigation: *Bias Mitigation for Data* (B311) inspect using bias mitigation techniques such as data resampling or reweighing to mitigate bias in the data t. *Rebalancing Techniques* (B312) evaluates the weight adjustment or presence of certain data points to substantiate a balanced and representative dataset. *Bias Mitigation for Models* (B321) inspect the corrective measures to adjust model behaviors that lead to unfair outcomes, ensuring equitable model performance. *Fair Optimization for Models* (B322) evaluates the fine-tuning of the model parameters to optimize fairness metrics and to adjust the model to perform equitably across all user groups. *Fairness Refactoring of Code* (B331) reviewing existing code to enhance fairness, removing or modifying parts of the code base that could lead to biased outcomes. *Equity Enhancements for Code* (B332) asses the mechanism taken to introduce new code elements designed to enhance the fairness of operations.
Assurance: *Third Party Assessments for Data and Models* (B411, B421) involves evaluating the external audits to verify the fairness of data handling and model outputs, providing independent verification of fairness practices. *Transparency Reports for Data, Models, and Code* (B412, B422, B432) evaluates the mechanism for disclosing fairness practices and outcomes to the public, ensuring accountability and transparency. *Compliance Certificates for Code* (B431) Check certification that the code meets all regulatory and ethical standards related to fairness. *Assurance Policies for Systems* (B441) inspect the set forth guidelines and standards to maintain fairness throughout the system's operational life to provide a layer for ongoing fairness checks. *Fairness Logs for Systems* (B442) evaluates the records of fairness-related decisions and actions, offering a traceable history that supports audits and continuous improvement.

3.6 Inclusivity layer for AI-Based Decision Support Systems

The Inclusivity layer affirms that AI systems are designed, developed, and deployed to be accessible and beneficial to diverse user groups. This definition addresses potential barriers that could exclude individuals or communities from fully participating in or benefiting from AI technologies. The summary of the inclusivity layer is presented in Table 4. The layer is operationalized across four interconnected phases:

Development: *Inclusion Criteria for Data* (C111) evaluates the standards to verify that data collection practices include diverse populations. *Language Variety in Data* (C112) evaluates the range of languages included in the dataset, ensuring broad linguistic representation. *Diverse Demography in Data* (C113) verifies that data sources reflect a wide demographic spectrum, enhancing the system's applicability to diverse user groups. *Selection Criteria for Models* (C121) enforce that models are chosen based on their ability to operate inclusively across varied demographics. *Interpretability of Models* (C122) measures how easily people from diverse backgrounds can understand model decisions. *Inclusive Design of Code* (C131) assesses code structures to verify that they are accessible and understandable to developers of varying abilities. *Language Support in Code* (C132) checks the extent to which the code accommodates multiple languages, facilitating wider usage. *Accessibility Standards for Systems* (C141) sets benchmarks to confirm system interfaces and functionalities are accessible to users with different abilities.
Assessment: *Coverage Evaluation of Data* (C211) refers to assessing data representativeness to cover a broad range of demographic groups as feasible. It measures how well the data covers the needs of diverse user groups, ensuring no demographic is overlooked. *Translation Assessment of Data* (C212) evaluates the accuracy and effectiveness of data translation, verifying linguistic fidelity and inclusiveness. *Inclusion Impact of Models* (C221) assesses the impact of model operations on various demographic groups, ensuring equitable outcomes. *Cultural Appropriateness of Models* (C222) evaluates models for cultural sensitivity, ensuring that they respect and accurately represent cultural nuances. *Accessibility Audit of Code* (C231) conducts thorough reviews of code accessibility, ensuring that all users can interact with the system effectively. *Cultural Testing of Systems* (C241) tests systems to ensure they perform appropriately across different cultural contexts, verifying cultural adaptability and sensitivity.
Mitigation: *Translation Mitigation in Data* (C311) evaluates the corrections to improve translation quality in datasets and to validate linguistic accuracy and inclusivity. *Language Expansion in Data* (C312) inspect the additional languages added into the dataset to broaden the system's linguistic reach. *Bias Mitigation in Models* (C321) evaluates the techniques to reduce or eliminate biased outcomes against any demographic gro p. *Representation Repair in Models* (C322) inspect the mechanisms provided to correct under-representation issues within model training sets to confirm diverse perspectives are adequately represented. **Accessibility Enhancements in Models** (C323) inspect the model's usability for users with disabilities. *Cultural Sensitivity in Models* (C324) evaluate model responses to reflect cultural diversity accurately and respectfully. *Accessibility Enhancements in Code* (C331) measures the code usability, ensuring it is accessible to developers with diverse needs. *Cultural Feedback in Systems* (C341) inspect the collections and integration of feedback from diverse cultural backgrounds to refine system operations. *Inclusive Updates in Systems* (C342) evaluates the system updates that enhance inclusivity, addressing identified gaps in system performance and accessibility.

Assurance: *Third Party Assessment of Data* (C411) checks if there is an external audit to verify that data handling practices meet inclusivity standards. *Third Party Assessment of Models* (C421) inspect that models operate fairly across all demographics through independent evaluations. *Third Party Assessment of Code* (C431) inspects that code development practices adhere to inclusivity standards using the external body. *Third Party Assessment of Systems* (C442) evaluates the overall system operations to confirm they comply with established inclusivity benchmarks using the external body. *Transparency Report for Data* (C412) inspects the disclosures about the inclusivity practices and outcomes for data, promoting transparency. *Accessibility Reports for Models* (C422) evaluates if the accessibility features and performance of models are disclosed to verify transparency and accountability. *Compliance Audits of Code* (C432) inspect the code practices meet all regulatory and ethical standards for inclusivity. *Inclusive Logs for Systems* (C441) evaluates the records of all actions taken to enhance system inclusivity to enforce traceable history that supports audits and continuous improvement.

Table 4. Inclusion Layer for AI-Based Decision Support Systems

Category	Development	Assessment	Mitigation	Assurance
Data	Inclusion Criteria, Data (C111)	Coverage Evaluation, Data (C211)	Translation Mitigation (C311)	Third-Party Assessment, Data (C411)
	Language Variety, Data (C112)	Translation Assessment, Data (C212)	Language Expansion, Data (C312)	Transparency Report, Data (C412)
	Diverse Demography, Data (C113)			
Model	Selection Criteria, Model (C121)	Inclusion Impact, Model (C221)	Bias Mitigation, Model (C321)	Third-Party Assessment, Model (C421)
	Interpretability, Model (C122)	Cultural Appropriateness, Model (C222)	Representation Repair, Model (C322)	Accessibility Reports, Model (C422)
			Accessibility Enhancements, Model (C323)	Inclusion Monitoring, Model (C423)
			Cultural Sensitivity, Model (C324)	
Code	Inclusive Design, Code (C131)	Accessibility Audit, Code (C231)	Accessibility Enhancements, Code (C331)	Third-Party Assessment, Code (C431)
	Language Support, Code (C132)			Compliance Audits, Code (C432)
System	Accessibility Standards, System (C141)	Cultural Testing, System (C241)	Cultural Feedback, System (C341)	Inclusive Logs, System (C441)
			Inclusive Updates, System (C342)	Third-Party Assessment, System (C442)

3.7 Ethical layer for AI-Based Decision Support Systems

The **Ethical Layer for AI-Based Decision Support Systems** delineates a layer designed to embed ethical considerations into the entire life cycle of AI systems. This layer emphasizes development, assessment, mitigation, and assurance that ethical standards are consistently applied and maintained. A summary of the ethical layer is presented in Table 5.

Development: *Consent Protocols for Data* (D111) evaluate the guidelines for obtaining informed consent, ensuring that data collection is transparent and ethically sound. *Ethical Source of Data* (D112) examines that all data used is sourced following strict ethical guidelines, promoting responsibility in data gathering. *Ethical Guidelines for Models* (D121) evaluates the ethical development of models to align with core values and principles. *Value Alignment in Models* (D122) examines that the models' functions align with the ethical standards and values of the organization. *Ethical Guidelines for Code* (D131) analyzes the integration of ethical standards in the coding process, ensuring ethical compliance from the ground up. *Stakeholder Engagement in Systems* (D141) evaluates the

effectiveness of involving stakeholders in the development process, ensuring their voices are considered in decision-making. *Transparency in Systems* (D142) measures the clarity and openness with which system processes are communicated to users and stakeholders.

Assessment: *Consent Assessment for Data* (D211) evaluates the adherence to consent protocols throughout the data life cycle, ensuring ongoing compliance with ethical standards. *Impact Assessment for Data* (D212) measures the potential social and ethical impacts of how data is used within the system. **Ethical Compliance for Models** (D221) assesses models to confirm that they operate within ethical guidelines and standards. *Environmental Impact of Models* (D222) evaluates the environmental footprint of models, promoting sustainability in model deployment. *Ethical Compliance for Code* (D231) checks code against ethical standards to substantiate its operations are ethically sound. *Transparency Assessment for Code* (D232) measures how transparent code operations are conducted and documented. *Transparency Assessment for Systems* (D241) verifies the transparency of system operations, enhancing accountability.

Mitigation: *Consent Verification for Data* (D311) confirms that data usage continues to comply with initial consent terms throughout its use. *Diversity Mitigation for Data* (D312) introduces strategies to ensure diverse data representation and prevent bias. *Value Adjustments for Models* (D321) modifies models to enhance alignment with ethical values. *Impact Mitigation for Models* (D322) applies corrective actions to minimize negative impacts identified during model assessments. *Mitigation for Code* (D331) involves rectifying any identified ethical issues in the code. *Transparency Mitigation for Code* (D332) increases the clarity and openness of code operations. *Stakeholder Feedback for Systems* (D341) incorporates mechanisms for collecting and addressing feedback from system users. *Ethical Mitigation for Systems* (D342) confirms ethical concerns identified are adequately addressed.

Assurance: *Third-party Review for Data* (D411) involves evaluating external audits to validate data's ethical handling and usage. *Third-party Review for Models* (D421) asses independent evaluation of models to verify their ethical integrity. *Third-party Review for Code* (D431) evaluates that code practices meet ethical standards through external verification. *Third-party Review for Systems* (D441) evaluates the overall system operations to verify they comply with ethical standards. extitEthics Report for Data (D412) examines disclosure on the ethical practices and outcomes for data management. *Ethics Report for Models* (D423) and *Ethics Report for Code* (D433) review the robustness of the ethical operations and modifications in models and code. *Value Audits for Models* (D422) specifically review how well models reflect the ethical values they represent. *Developer Ethical Training for Code* (D432) evaluate code developers' understanding and implementation of ethical practices. *Transparency Log for Systems* (D442) reviews the record of all transparency-related actions and decisions within the system. *Stakeholder Reviews for Systems* (D443) analyzes stakeholder feedback to uphold system operations that remain ethically aligned and transparent.

Table 5. Ethical Layer for AI-Based Decision Support Systems

Category	Development		Assessment		Mitigation		Assurance	
Data	Consent Protocols, Data	(D111)	Consent Assessment, Data	(D211)	Consent Verification, Data	(D311)	Third-Party Review, Data	(D411)
	Ethical Source, Data	(D112)	Impact Assessment, Data	(D212)	Diversity Mitigation, Data	(D312)	Ethics Report, Data	(D412)
Model	Ethical Guidelines, Model	(D121)	Ethical Compliance, Model	(D221)	Value Adjustments, Model	(D321)	Third-Party Review, Model	(D421)
	Value Alignment, Model	(D122)	Environmental Impact, Model	(D222)	Impact Mitigation, Model	(D322)	Value Audits, Model	(D422)
							Ethics Report, Model	(D423)
Code	Ethical Guidelines, Code	(D131)	Ethical Compliance, Code	(D231)	Mitigation, Code	(D331)	Third-Party Review, Code	(D431)
			Transparency Assessment, Code	(D232)	Transparency Mitigation, Code	(D332)	Developer Ethical Training, Code	(D432)
							Ethical Report, Code	(D433)
System	Stakeholder Engagement, System	(D141)	Transparency Assessment, System	(D241)	Stakeholder Feedback, System	(D341)	Third-Party Review, System	(D441)
	Transparency, System	(D142)			Ethical Mitigation, System	(D342)	Transparency Log, System	(D442)
							Stakeholder Reviews, System	(D443)

3.8 Legal layer for AI-Based Decision Support Systems

The Legal layer verifies that AI systems comply with applicable laws, regulations, and legal obligations throughout their life cycles. This dimension is essential to mitigate legal risks, protect stakeholders, and foster trust in AI technologies. The summary of the legal layer is presented in Table 6. The layer is operationalized across four interconnected phases:

Development: *License Verification for Data* (E111) affirms that all data used complies with applicable licensing agreements. *Regulation Tracking for Data* (E112) measures adherence to changing legal regulations, ensuring data handling remains compliant. *Consent Logs for Data* (E113) maintains records of user consents, facilitating audits and compliance checks. *Contract Compliance for Models* (E121) verifies that model development adheres to contractual obligations and legal standards. *Patent Verification for Models* (E122) safeguards model designs and functionalities do not infringe on existing patents. *Compliance Documentation for Models* (E123) maintains thorough records of compliance efforts, supporting legal audits. *License Audit for Code* (E131) checks code components against license requirements to verify compliance. *Secrets Management in Code* (E132) Check if a mechanism is in place to secure sensitive data and proprietary algorithms within the code. *Export Compliance for Code* (E133) confirms code exports adhere to international trade laws. *Regulatory Submissions for Systems* (E141) evaluate the submission system specifications and operations for regulatory approval. *Privacy Compliance for Systems* (E142) evaluates system designs to confirm they meet privacy laws and standards. *Accessibility Audit for Systems* (E143) verifies that system interfaces comply with accessibility regulations.

Assessment: *Legal Review of Data* (E211) evaluates the legal implications of data usage, ensuring compliance with privacy and data protection laws. *Privacy Assessment for Data* (E212) measures how well personal data is protected against unauthorized access and breaches. *Consent Audit for Data* (E213) checks the effectiveness of consent management processes. *Legal Testing for Models* (E221) assesses models against legal risks and compliance requirements. *Liability Review for Models* (E222) evaluates potential liabilities arising from model applications. *Regulatory Review for Models* (E223) verifies that models comply with specific industry regulations. *Security Check for Code* (E231) evaluates code for vulnerabilities that could lead to legal breaches. *Compliance Monitoring for*

Code (E232) continuously checks code compliance with legal standards. *Documentation Check for Code* (E233) verifies that all code documentation meets legal documentation standards. *Compliance Verification for Systems* (E241) Verify that system operations comply with all relevant legal requirements. *Safety Check for Systems* (E242) evaluates system components and operations for safety compliance. *Legal Monitoring for Systems* (E243) Check if there is a continuously observed system operation to ensure ongoing legal compliance.

Mitigation: *Breach Protocols for Data* (E311) evaluates the procedures to manage and mitigate data breaches effectively. *Consent Updates for Data* (E312) evaluates the consent mechanisms to remain compliant with evolving legal standards. *Rectification Processes for Data* (E313) will assess the measure's robustness to correct any identified legal non-compliance in data handling g. *Compliance Adjustments for Models* (E321) evaluate the process if the models are modified to correct legal and compliance issues identified. *Liability Mitigation for Models* (E322) evaluates strategies to reduce potential legal liabilities associated with model applications. *Legal Remediation for Models* (E323) assesses the processes for correcting legal deficiencies identified in models. *Code Corrections for Code* (E331) evaluates the measures taken to rectify legal non-compliance found in code audits. *Compliance Patching for Code* (E332) assesses the updates made to code to meet new legal standards. *Security Updates for Code* (E333) evaluates code security enhancements to prevent legal breaches. *Compliance Upgrades for Systems* (E341) evaluates system updates to enhance legal compliance. *Regulatory Adjustments for Systems* (E342) assesses system modifications to align with updated regulations *Legal Reassessments for Systems* (E343) evaluates the processes for reevaluating systems to make sure that they meet current legal standards.

Assurance: *Compliance Certification for Data* (E411) checks if there is a compliance certification for the data and also assesses the certification process, ensuring data handling practices meet all legal requirements. *Privacy Guarantees for Data* (E412) evaluates the robustness of privacy protections to confirm compliance with la s. *Legal Reporting for Data* (E413) assesses the provision of detailed reports on data compliance status. *Legal Clearance for Models* (E421) evaluates processes, ensuring that models meet all legal requirements before deployment. *Regulatory Compliance for Models* (E422) assesses compliance of models with specific industry regulations. *Audit Trails for Models* (E423) evaluates the logs of all compliance checks and measures taken for adjustments. *License Audits for Code* (E431) assesses the auditing of code components for legal compliance. *Ethical Coding for Code* (E432) evaluates adherence to the highest ethical and legal standards in coding practices. *Document Compliance for Code* (E433) assesses verifying all code-related documents for legal compliance. *Legal Conformance for Systems* (E441) evaluates comprehensive reviews, ensuring that entire systems meet legal standards. *Safety Compliance for Systems* (E442) assesses adherence to legal requirements for safety protocol. *Accessibility Verification for Systems* (E443) evaluates whether system accessibility meets or exceeds legal requirements.

Table 6. Legal Layer for AI-Based Decision Support Systems

Category	Development	Assessment	Mitigation	Assurance	
Data	License Verification	(E111) Legal Review	(E211) Breach Protocols	(E311) Compliance Certification	(E411)
	Regulation Tracking	(E112) Privacy Assessment	(E212) Consent Updates	(E312) Privacy Guarantees	(E412)
	Consent Logs	(E113) Consent Audit	(E213) Rectification Processes	(E313) Legal Reporting	(E413)
Model	Contract Compliance	(E121) Legal Testing	(E221) Compliance Adjustments	(E321) Legal Clearance	(E421)
	Patent Verification	(E122) Liability Review	(E222) Liability Mitigation	(E322) Regulatory Compliance	(E422)
	Compliance Documentation	(E123) Regulatory Review	(E223) Legal Remediation	(E323) Audit Trails	(E423)
Code	License Audit	(E131) Security Check	(E231) Code Corrections	(E331) License Audits	(E431)
	Secrets Management	(E132) Compliance Monitoring	(E232) Compliance Patching	(E332) Ethical Coding	(E432)
	Export Compliance	(E133) Documentation Check	(E233) Security Updates	(E333) Document Compliance	(E433)
System	Regulatory Submissions	(E141) Compliance Verification	(E241) Compliance Upgrades	(E341) Legal Conformance	(E441)
	Privacy Compliance	(E142) Safety Check	(E242) Regulatory Adjustments	(E342) Safety Compliance	(E442)
	Accessibility Audit	(E143) Legal Monitoring	(E243) Legal Reassessments	(E343) Accessibility Verification	(E443)

4 Discussion

System Card+ builds on existing frameworks by providing life cycle-specific criteria across transparency, reproducibility, ethics, and regulatory alignment. Unlike high-level guidelines, it embeds actionable measures within each stage of the AI life cycle. T's structured approach sets it apart by ensuring robust compliance with ethical and legal standards while promoting fair and inclusive AI practices. The framework's modular, five-layer design enables it to be integrated into existing AI workflows with minimal disruption.

The framework's coverage of the entire AI life cycle from data collection and multidimensional to deployment and post-production—offers substantial benefits in ensuring responsible AI deployment. At the data collection stage, the framework emphasizes transparency and accuracy through documentation and reproducibility. This establishes a solid foundation where data is collected, ethically processed, and stored to safeguard its integrity and reliability. During model development, the focus on reproducibility and design transparency ensures that AI models are built on unbiased, well-documented data. This enhances their trustworthiness and robustness. Deployment phases benefit from continuous monitoring and assessment protocols. These protocols align the AI system's performance with ethical and regulatory standards. In the post-production stage, the framework's maintenance plans and regular audits verify that AI systems operate responsibly. They adapt to new data and evolving regulatory requirements. This life cycle approach facilitates early detection and rectification of biases and promotes continuous improvement.

The framework's alignment with international AI regulations, particularly the EU AI Act, is a significant strength. The EU AI Act categorizes AI systems based on risk levels and mandates specific requirements for high-risk AI applications. The Performance Accountability Layer, emphasizing detailed documentation and transparency, directly supports the Act's requirements for traceability and accountability. The framework meets the stringent documentation and oversight demanded by the Act by ensuring comprehensive records of data sources, model decisions, and system operations. The Fairness Layer aligns with the Act's non-discrimination and fairness mandates by incorporating fairness-by-design principles and a continuous bias detection mechanism. T's proactive approach assures that AI systems comply with the Act's standards for equitable

treatment across all demographic groups. The Ethical Layer addresses the Act's emphasis on ethical AI by embedding consent protocols, value alignment, and stakeholder engagement throughout the AI life cycle. The Legal Layer affirms adherence to all relevant legal standards, including those outlined in the EU AI Act, through rigorous legal reviews and compliance monitoring. By integrating these comprehensive measures, the framework not only aligns with but also reinforces the principles and requirements of the EU AI Act, ensuring robust legal and ethical compliance in AI deployments.

Each layer of the framework addresses specific aspects of responsible AI. The framework's comprehensive approach upholds that potential loopholes are systematically addressed through its detailed layers. The ch layer covers specific elements of responsible AI, from performance accountability to ethical considerations. This in-depth focus guarantees that every aspect of the AI life cycle is scrutinized and managed, leaving no room for oversight. By embedding rigorous checks and balances within each layer, the framework effectively mitigates risks and preempts issues that could compromise system integrity. Detailed attention across layers validates the robust framework, with comprehensive safeguards that address potential vulnerabilities and reinforce overall reliability.

System Card+ offers a complementary enhancement to existing frameworks like Google's AI Principles, Microsoft's Responsible AI Strategy, and the NIST AI framework. While these frameworks provide valuable high-level guidelines, they often lack the depth required for thorough life cycle management. The proposed framework's detailed coverage of every stage, from data collection to post-production, fills critical gaps by providing actionable, specific measures that can be integrated with the broader principles established by Google and Microsoft. It offers practical solutions that bolster regulatory compliance and operational rigor by aligning closely with international regulations, such as the EU AI Act. This detailed approach enhances the existing frameworks and supports organizations in implementing their principles more effectively, ensuring that responsible AI practices are seamlessly incorporated throughout the entire life cycle.

The framework's versatility enables its application across various industries, including healthcare, finance, and retail. In healthcare, it ensures that AI systems adhere to rigorous data privacy and accuracy standards, which is crucial for maintaining trust and safeguarding patient outcomes. In the finance sector, the framework enhances fairness and transparency in algorithmic decision-making, helping to prevent biases and protect consumer rights. Retail applications benefit from improved customer experience through more accurate personalization and ethical data u e. The framework is designed with scalability in mind, benefiting small startups and large corporations. For startups, it provides a structured approach to building responsible AI systems from the ground up, ensuring they can compete on a level playing field. Large corporations, on the other hand, can leverage the framework to refine existing systems and integrate new technologies while adhering to comprehensive ethical and regulatory standards. Its flexibility and adaptability across different industry needs are critical to its effectiveness. The framework's modular design allows it to be tailored to specific industry re-

quirements and scaled to the organization's site. This adaptability confirms that whether an organization is just beginning its AI journey or is deeply entrenched in advanced AI deployment, the framework offers relevant, actionable guidance that aligns with industry standards and individual operational needs.

5 Conclusion

Bridging the gap between AI policy regulations and their practical enforcement poses a significant challenge, predominantly due to the lack of a unified, accessible, and straightforward mechanism for compliance and certification. The System Card+ methodology offers a considerable step forward in the ethical assessment of AI systems. This framework addresses this critical gap by introducing a five-layer holistic benchmark to verify that the principles of performance accountability, fairness, inclusivity, and ethical and legal compliance are embedded within AI systems. This framework harmonizes with existing legal and ethical frameworks governing AI accountability. It has universal applicability across various sectors and sizes. The approach will not only enhance responsible AI but will also affirm accountability y. The work on this project includes the step-by-step implementation of the theoretical framework to develop practical metrics for the data, code, model, and system subsections. We plan to validate the framework through application in a variety of industries. Despite our efforts to present a comprehensive system, we acknowledge that no framework can serve as a perfect solution to fully mitigate the complex problems associated with responsible AI and its governance. The inherent limitations of this formative approach emphasize the need for continuous evaluation and adaptation. This recognition does not diminish the value of our contribution; instead, it underscores the importance of flexibility and responsiveness. We recommend a dynamic approach to AI regulation that promotes continuous learning, adaptation, and ethical vigilance.

Acknowledgments: The authors thank their colleagues, Prof. Ryan Kennedy, Prof. Lydia Tiede, Prof. Andrew Michaels, and the CRASA board of advisors for their support and valuable feedback. The National Science Foundation, under Award Number 2131504, supported this research. The views and conclusions expressed in this paper are those of the authors and do not necessarily reflect the official policies or endorsements, either expressed or implied, of the National Science Foundation.

References

[1] S M Yousef. Applications of artificial intelligence in healthcare: A review. *ScienceOpen Preprints*, 2021.

[2] H. Tibebu, I. A. Kakadiaris E. Mekonnen, and V.D Silva. Measuring public policy effectiveness in the age of data and AI: Insights from COVID-19. In

Proc of the 2023 IEEE IAS Global Conference on Emerging Technologies, Loughborough University, London, United Kingdom, May 19–21 2023.

[3] B. Bonnie. *Artificial intelligence in finance*. The Alan Turing Institute, 2019.

[4] C. Bart. AI in criminal law: an overview of AI applications in substantive and procedural criminal law. *Law and artificial intelligence: Regulating AI and applying AI in legal practice*, pages 205–223, 2022.

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[6] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91. New York, 2018.

[7] J. Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2016.

[8] K. Crawford and R. Calo. There is a blind spot in AI research. *Nature*, 538(7625):311–313, 2016.

[9] M.Ienca A. Jobin and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1:389–399, 2019.

[10] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 2016.

[11] T. Hagendorff. The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, 2020.

[12] J. Whittlestone, R. Nyrup, A. Alexandrova, S. Cave, and K. Dihal. Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research, London, 2019.

[13] A. D. Selbst and D. boyd. Understanding algorithmic impact assessments: Towards accountable automation. *UCLA Law Review*, 66:142–176, 2019.

[14] L. Floridi and J. Cowls. A unified framework of five principles for AI in Society. *Harvard Data Science Review*, 1(1), 2019.

[15] D. Leslie. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, London, UK, 2019.

[16] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi. Artificial intelligence and the 'good society': the role of human values in AI design. *AI & Socity*, 33(2):305–320, 2018.

[17] F. Gursoy and I. A. Kakadiaris. System cards for ai-based decision-making for public policy. *arXiv*, 2022.

[18] K.A. Joshua, H. Joanna, B. Solon, F. Edward, R. Joel, R. R. David G, and Y. Harlan. Accountable algorithms. *University of Pennsylvania Law Review*, 165:633, 2017.

[19] R. I. Deborah, S. Andrew, W. Rebecca N., M. Margaret, G. Timnit, H. Ben, S. Jamila, T. Daniel, and B. Parker. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In

Proc Fairness, Accountability, and Transparency, pages 33–44, Barcelona, Spain, 2020.

[20] C. Corinne, W. Sandra, M. Brent, T. Mariarosaria, and F. Luciano. Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 2018.

[21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv*, 2019. [Online]. Available: `https://arxiv.org/abs/1908.09635`.

[22] N. Safiya. U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

[23] E. Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[24] F. Alessandro, M. Stefano, S. Gianmaria, and S. Gian Antonio. Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, 2022.

[25] P. Dana and S. Erez. Algorithmic fairness. In *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pages 867–886. Springer, 2023.

[26] L. Peter, J. David, H. Felicia, and C. PJ. A framework for collecting inclusive design data for the UK population. *Applied Ergonomics*, 46:318–324, 2015.

[27] B. Nicholas P., H. Alyssa M., K. Emily E., T. Richelle L., and W. Kathryn. A push for inclusive data collection in STEM organizations. *Science*, 376(6588):37–39, 2022-04.

[28] S. Ari, O. Kenton P, and T. Alex S. Let's talk about race: Identity, chatbots, and AI. In *Proc CHI Conference on Human Factors in Computing Systems*, pages 1–14, Montréal, 2018.

[29] O. Jessica, M. Leonard, T. Verena, M. Christian, and L. Sven. Perceived algorithmic fairness: An empirical study of transparency and anthropomorphism in algorithmic recruiting. *Information Systems Journal*, 34(2):384–414, 2024.

[30] Z. Wei, L. Lidan, W. Xiaoxue, and C. Xiang. An empirical study on correlations between deep neural network fairness and neuron coverage criteria. *IEEE Transactions on Software Engineering*, 2024.

[31] F. Luciano and C. Josh. A unified framework of five principles for AI in Society. *Harvard Data Science Review*, 1(1):1–16, 2019.

[32] M. Brent D, A. Patrick, T. Mariarosaria, W. Sandra, and F. Luciano. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):1–21, 2016.

[33] I. Marcello J. Anna and V. Effy. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.

[34] B. Reuben. Fairness in machine learning: Lessons from political philosophy. In *Proc Fairness, Accountability, and Transparency*, pages 149–159, New York, USA, 2018.

[35] M. James H. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, 2006.

[36] V. Paul and V. d. Axel. *The EU general data protection regulation (GDPR)*, volume 10. Springer, 2017.

[37] M. Tambiama. Artificial intelligence act. Technical report, European Parliament: European Parliamentary Research Service, 2021.

[38] M. Brent. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11):501–507, 2019.

[39] B. Carsten Stahl. Ethical issues of AI: A european perspective. *IEEE Technology and Society Magazine*, 40(3):72–80, 2021.

[40] L. Floridi. Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, 33:185–193, 2020.

[41] B. Zhang and M. Mildenberger. Automated societies: Challenges and opportunities in the age of AI. *Daedalus*, 150(2):146–158, 2021.

[42] M. Veale and I. Brass. Administration by algorithm? Public management meets public sector machine learning. *The Oxford Handbook of Ethics of AI*, pages 134–150, 2019.