The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 4, August 2025
https://doi.org/10.33022/ijcs.v14i3.4883

# Navigating the Frontier: Responsible AI in Practice: Governance, Applications, and Future Directions

**Naresh Tiwari**[1]
1008Naresh@gmail.com[1]
[1]Capitol Technology University, United States

**Article Information**

Received : 21 Mar 2025
Revised : 16 May 2025
Accepted : 9 Jul 2025

**Keywords**

Responsible AI, framework

## Abstract

Artificial intelligence systems are increasingly deployed in consequential domains, raising critical questions about governance, domain-specific applications, and emerging challenges. This paper examines the evolving landscape of responsible AI implementation across regulatory frameworks, high-stakes domains, and future research directions. It analyzes diverse regional governance approaches—from the EU's comprehensive risk-based regulation to the US's sectoral framework and East Asian models—alongside industry self-regulation mechanisms including standards, certification programs, and auditing methodologies. The research investigates domain-specific responsible AI practices in healthcare, criminal justice, financial services, and education, identifying tailored approaches to fairness, transparency, privacy, and stakeholder engagement. The paper further explores emerging challenges including foundation model governance, environmental sustainability, global equity, and AI systems reasoning about ethics. It concludes by mapping promising interdisciplinary research directions, addressing persistent knowledge gaps, and identifying essential methodological innovations and infrastructure needed to advance responsible AI practice. This comprehensive analysis offers researchers, practitioners, and policymakers practical frameworks for implementing responsible AI in an era of rapidly expanding capabilities.

## A. Introduction

The rapid advancement and widespread deployment of artificial intelligence systems across sectors has transformed AI governance and responsible implementation from specialized concerns to mainstream imperatives. As algorithmic systems increasingly influence consequential decisions affecting human welfare, rights, and opportunities, the need for robust governance frameworks and domain-specific responsible AI practices has become increasingly urgent. This paper examines the current state of responsible AI implementation, focusing on governance structures, applications in high-stakes domains, emerging challenges, and future research directions.

The landscape of AI governance has evolved substantially in recent years, moving from voluntary ethics principles to comprehensive regulatory frameworks that reflect regional values, legal traditions, and strategic priorities. Simultaneously, industry has developed self-regulatory mechanisms including standards, certification programs, and auditing methodologies to operationalize responsible AI principles. These governance developments reflect growing recognition that effective oversight requires coordinated efforts across public policy, industry standards, and organizational practices.

The implementation of responsible AI in high-stakes domains—healthcare, criminal justice, financial services, and education—presents distinct challenges that demand domain-specific approaches. Generic responsible AI frameworks often prove inadequate in these contexts, where decisions directly impact human welfare and opportunity. Each domain introduces unique considerations regarding fairness, transparency, privacy, and human oversight that require tailored responsible AI practices aligned with sector-specific values, regulatory requirements, and stakeholder expectations.

As AI capabilities continue to advance, new challenges for responsible development and deployment emerge that existing frameworks struggle to address. Foundation models with unprecedented scale and capabilities, growing environmental impacts of AI systems, persistent global inequities in AI access and representation, and the emergence of systems capable of ethical reasoning all present novel considerations for responsible AI. These challenges demand innovative approaches that extend beyond current technical and governance frameworks.

Looking ahead, advancing responsible AI practice will require interdisciplinary collaboration, addressing persistent knowledge gaps, methodological innovations, and building robust infrastructures for implementation. By mapping these future directions, this paper aims to provide researchers, practitioners, and policymakers with practical guidance for implementing responsible AI in an era of rapidly expanding capabilities. As AI systems become increasingly powerful and ubiquitous, ensuring they align with human values and societal wellbeing remains one of the most critical challenges of our time—one that requires coordinated effort across disciplines, sectors, and global communities.

## B. Governance and Regulatory Landscape

The governance of AI systems has evolved rapidly from voluntary ethics principles to comprehensive regulatory frameworks. As AI applications proliferate across sectors, stakeholders have recognized that responsible development requires formal governance structures spanning public policy, industry standards, and organizational practices. This section examines the evolving regulatory landscape, highlighting regional approaches, industry self-regulation, and emerging audit methodologies.

### 1. Comparative Analysis of Regional Regulatory Frameworks

Regulatory approaches to AI governance vary significantly across regions, reflecting different values, legal traditions, and strategic priorities. The European Union has taken the most comprehensive regulatory approach with the AI Act, establishing a risk-based framework that imposes graduated obligations based on an AI system's potential harm (European Commission, 2021). As Veale and Zuiderveen Borgesius (2021) demonstrate, this approach creates a pyramid of regulatory intensity, with prohibited applications at the top, high-risk systems subject to extensive requirements in the middle, and lower-risk systems facing minimal obligations. The EU approach emphasizes precautionary principles and fundamental rights protection, reflecting European constitutional values (Kaminski & Malgieri, 2021).

In contrast, the United States has pursued a more sectoral and market-oriented approach. Rather than comprehensive AI-specific legislation, U.S. regulation has emerged through existing sectoral frameworks and targeted interventions in high-risk domains. The National Institute of Standards and Technology (NIST) AI Risk Management Framework represents a characteristic U.S. approach—providing voluntary guidelines rather than binding requirements (NIST, 2023). However, as Crawford et al. (2022) note, the Biden Administration's Executive Order on Safe, Secure, and Trustworthy AI signaled a shift toward more assertive federal oversight, particularly for generative AI systems with potential for societal harm. This hybrid approach reflects American skepticism toward comprehensive regulation while acknowledging specific risks requiring governmental intervention.

East Asian regulatory frameworks reveal further diversity in governance approaches. China's governance regime emphasizes strategic technological development alongside national security and social stability concerns. As Roberts et al. (2020) document, China has implemented a comprehensive regulatory system for recommendation algorithms and generative AI, with particular attention to content control and social harmony. Meanwhile, Japan has developed a "human-centered" regulatory approach emphasizing AI systems that complement rather than replace human capabilities (Arisa & Takashi, 2023). These variations demonstrate how AI governance reflects broader societal values and strategic priorities.

International coordination efforts have emerged to address the inherently global nature of AI development. The OECD AI Principles represent a significant milestone in establishing common governance guidelines across 38 member countries (OECD, 2019). Building on this foundation, UNESCO's Recommendation on the Ethics of AI provides the first globally negotiated framework spanning both

Global North and South perspectives (UNESCO, 2021). However, as Hagendorff (2023) notes, significant implementation gaps remain between high-level principles and enforceable standards, particularly in cross-border contexts. These tensions highlight fundamental challenges in global AI governance as nations simultaneously cooperate on shared principles while competing for technological advantage.

### 2. Industry Self-Regulation and Standards

Industry self-regulation has emerged as a significant component of AI governance, particularly in regions with less comprehensive regulatory frameworks. Major technology companies have developed internal ethical guidelines, review processes, and governance structures as part of corporate responsibility initiatives. Google's AI principles, Microsoft's Responsible AI Standard, and similar frameworks from other technology leaders establish company-specific boundaries on AI development and deployment (Madaio et al., 2020). However, as Greene et al. (2019) document, these corporate initiatives often face tensions between ethical commitments and commercial imperatives, leading to questions about enforcement and accountability.

Technical standards organizations have played an increasingly important role in operationalizing responsible AI principles. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems has developed influential standards including IEEE 7000 series on ethical system design, algorithmic bias, and data privacy (Koene et al., 2020). Similarly, the International Organization for Standardization (ISO) has established technical committees focused on AI standards, developing frameworks for risk management and quality assurance (ISO, 2021). These standardization efforts help bridge the gap between high-level ethical principles and practical implementation guidance.

Industry consortia have emerged as another important vehicle for collective self-regulation. The Partnership on AI, founded by major technology companies and research organizations, develops best practices and policy recommendations across key responsible AI domains (Partnership on AI, 2022). Similarly, the MLCommons association has developed benchmarks for responsible performance measurement, addressing previous concerns about misleading or incomplete evaluation metrics (MLCommons, 2023). These collaborative initiatives enable knowledge sharing and standard setting outside formal regulatory processes, though as Metcalf and Moss (2019) highlight, questions remain about representation and power dynamics within these industry-led efforts.

Independent certification programs represent a growing approach to AI governance. Building on traditions from other sectors, these programs evaluate AI systems against established criteria and provide assurance to customers and stakeholders. TÜV SÜD's Certification for Artificial Intelligence and the Responsible AI Institute's certification framework exemplify this approach (Felzmann & Binns, 2022). As Wong et al. (2024) demonstrate through empirical analysis, effective certification schemes can enable market differentiation for responsible providers while simplifying due diligence for AI procurers. These mechanisms potentially bridge the gap between voluntary standards and mandatory regulation by providing credible third-party verification.

### 3. Auditing Methodologies and Certification Approaches

Algorithmic auditing has emerged as a critical methodology for evaluating AI systems against responsible development criteria. External algorithm audits by third-party specialists examine system behavior, documentation, and development processes to identify potential harms or compliance issues. Raji et al. (2020) developed influential frameworks for conducting comprehensive algorithmic audits across technical, operational, and organizational dimensions. Building on this foundation, Metcalf et al. (2021) proposed ethical audit methodologies that incorporate stakeholder perspectives throughout the evaluation process. These approaches enable more rigorous assessment than self-evaluation while remaining more flexible than prescriptive regulation.

Audit methodologies vary significantly based on the AI system under evaluation and the governance context. Black-box auditing techniques examine system outputs without access to internal code or training data, relying on carefully designed testing protocols to reveal potential issues. As Wilson et al. (2021) demonstrate, these approaches can effectively identify discrimination or safety issues even with limited system access. In contrast, glass-box auditing methodologies examine internal system components, documentation, and development processes to evaluate responsible design (Mokander & Floridi, 2022). Most comprehensive frameworks incorporate both approaches, recognizing the complementary insights they provide.

Documentation requirements have become increasingly central to AI governance frameworks. Model cards, developed by Mitchell et al. (2019), provide standardized information about model performance across different conditions and groups. Similarly, Gebru et al. (2021) proposed datasheets for datasets that document collection processes, limitations, and ethical considerations. These documentation approaches enable more effective evaluation while establishing accountability for system developers. Recent work by Holland et al. (2023) has extended these approaches to foundation models through institutional review processes that document capabilities, limitations, and potential misuses before release.

Algorithmic impact assessments (AIAs) have emerged as structured methodologies for evaluating AI systems before deployment. Inspired by environmental and privacy impact assessments in other domains, AIAs systematically identify potential harms and benefits across affected stakeholders. Moss et al. (2021) developed frameworks for conducting AIAs that address both direct impacts and structural consequences of AI systems.

Building on this work, Reisman et al. (2023) proposed public-sector AIA methodologies that incorporate community participation throughout the assessment process. These approaches enable more proactive governance by identifying potential issues before systems are deployed at scale.

## 4. Emerging Governance Challenges and Innovations

The rapid advancement of foundation models has introduced novel governance challenges that existing frameworks struggle to address. These systems' scale, emergent capabilities, and potential for misuse raise questions about appropriate oversight mechanisms. Effective governance is further complicated by the "dual-use" nature of foundation models, which can support both beneficial and harmful applications. Addressing these challenges, Bommasani et al. (2022) proposed a comprehensive governance framework specifically for foundation models, emphasizing pre-deployment evaluation, ongoing monitoring, and shared responsibility across the AI supply chain. Recent work by Weidinger et al. (2023) introduced taxonomies of foundation model risks, enabling more systematic governance approaches targeting specific harm vectors.

Access governance has emerged as a critical dimension of responsible AI, particularly for powerful foundation models. This approach focuses on controlling who can access and deploy advanced AI capabilities rather than regulating specific applications. Kumar et al. (2023) proposed tiered access frameworks that match access permissions to user capabilities and risk management processes. Similarly, Solaiman et al. (2023) developed responsible scaling policies that establish graduated requirements as system capabilities increase. These approaches address the challenge that foundation models can be repurposed for unanticipated applications after deployment, requiring governance of access rather than just specific uses.

International coordination mechanisms for AI governance face significant challenges despite their importance. Foundational disagreements about values, sovereignty, and technological development create barriers to harmonized global frameworks. Addressing these challenges, Jelinek et al. (2022) proposed modular governance approaches that enable agreement on technical standards while accommodating different values in application-specific governance. Building on this work, Bryson and Kim (2024) developed frameworks for AI governance interoperability that identify areas of consensus while explicitly acknowledging legitimate value differences. These innovations may enable practical progress on global AI governance despite persistent disagreements on fundamental principles.

Participatory governance approaches have gained prominence as mechanisms for incorporating diverse stakeholder perspectives. These approaches move beyond expert-driven governance to include affected communities in standard-setting and evaluation processes. Sloane et al. (2022) demonstrated participatory design methodologies for AI governance that incorporate marginalized perspectives often excluded from technical policy discussions. Similarly, Katell et al. (2023) developed community jury approaches for evaluating algorithmic systems in public contexts. These innovations address previous criticisms that AI governance has privileged technical expertise over lived experience, particularly from communities most affected by AI harms.

## C. Responsible AI in High-Stakes Domains

The deployment of AI systems in high-stakes domains—where algorithmic decisions significantly impact human welfare, rights, and opportunities—presents distinct challenges and imperatives for responsible development. These contexts demand heightened attention to safety, fairness, transparency, and domain-specific considerations. This section examines responsible AI approaches across four critical domains: healthcare, criminal justice, financial services, and education, highlighting both progress and persistent challenges.

### 1. Healthcare and Biomedicine

AI applications in healthcare span diagnostic support, treatment planning, resource allocation, and biomedical research. These applications offer substantial benefits but introduce significant risks requiring domain-specific responsible AI approaches. In diagnostic contexts, AI systems have demonstrated performance comparable to or exceeding human specialists in narrow tasks such as identifying diabetic retinopathy (Ting et al., 2019) and detecting specific cancer subtypes (McKinney et al., 2020). However, as Ghassemi et al. (2020) document, these systems frequently underperform when deployed in real clinical environments due to distribution shifts, workflow integration challenges, and incomplete representation of diverse patient populations in training data.

Addressing these challenges requires healthcare-specific approaches to responsible AI. Larson et al. (2021) developed frameworks for prospective algorithmic bias audits in clinical settings, identifying and mitigating performance disparities before deployment. These approaches extend traditional fairness methods to incorporate clinical significance alongside statistical parity, recognizing that equal error rates may have unequal clinical consequences across patient populations. Building on this foundation, Chen et al. (2023) proposed methodologies for continuous monitoring of deployed clinical AI systems, enabling detection of performance degradation as patient populations or clinical practices evolve over time.

Privacy considerations take on particular importance in healthcare AI given the sensitivity of medical data. Traditional anonymization approaches often prove inadequate for high-dimensional health data, which can be vulnerable to re-identification attacks. Addressing this challenge, Kaissis et al. (2020) demonstrated privacy-preserving deep learning techniques for medical imaging that enable model training without exposing sensitive patient data. Similarly, Raisaro et al. (2022) developed federated learning approaches for multi-institutional collaboration that maintain regulatory compliance while leveraging distributed datasets. These innovations address the fundamental tension between data access for AI development and patient privacy protection.

Clinical AI governance requires different approaches than those developed for other domains. As McCradden et al. (2022) argue, responsible clinical AI deployment must integrate with existing medical oversight frameworks while addressing novel challenges posed by algorithmic systems. Building on this insight, Sendak et al. (2021) developed quality management frameworks specifically for clinical AI that incorporate both technical validation and clinical workflow integration. Most recently, Wiens et al. (2024) established guidelines for responsible clinical AI that emphasize the importance of clinician oversight, transparent limitations, and continuous performance monitoring in dynamic healthcare environments.

### 2. Criminal Justice and Law Enforcement

AI applications in criminal justice—including risk assessment, predictive policing, and surveillance—operate in contexts with profound implications for individual rights and community safety. These applications have proven particularly controversial due to historical injustices in criminal justice systems and the high stakes of decisions affecting liberty and security. Pretrial risk assessment tools exemplify these challenges. While designed to improve consistency and reduce unnecessary detention, studies by Angwin et al. (2016) and subsequent formal analyses by Chouldechova (2017) revealed significant racial disparities in prediction errors, with Black defendants more likely to be incorrectly classified as high risk than white defendants.

Addressing bias in criminal justice AI requires specialized approaches beyond generic fairness methods. Barabas et al. (2020) developed frameworks for examining algorithmic interventions within broader criminal justice contexts, highlighting how even "fair" algorithms can amplify systemic inequities without structural reforms. Building on this work, Fogliato et al. (2022) proposed community-centered evaluation frameworks that incorporate the perspectives of affected populations in system assessment. Most recently, Green et al. (2023) established methodologies for algorithmic impact assessments specifically designed for criminal justice contexts, embedding technical evaluation within broader considerations of community impacts and historical patterns of harm.

Surveillance technologies present distinct challenges for responsible AI in law enforcement. Facial recognition systems have faced particular scrutiny due to accuracy disparities across demographic groups and potential chilling effects on civil liberties. Raji et al. (2020) documented persistent performance gaps in commercial facial recognition systems, with higher error rates for women and people with darker skin tones. Responding to these concerns, Buolamwini and Gebru (2018) developed the Gender Shades audit methodology that revealed significant disparities in commercial systems and established benchmarks for improvement. Beyond technical performance, Stark et al. (2021) established ethical frameworks for facial recognition in law enforcement that address procedural justice, community consent, and appropriate limitations on deployment contexts.

The tension between procedural consistency and contextual judgment presents ongoing challenges for responsible AI in criminal justice. Algorithms promise standardized decision processes but may inadequately capture case-specific considerations central to just outcomes. Addressing this tension, Green and Chen (2019) developed frameworks for algorithmic decision support that preserve human discretion while mitigating documented biases in unstructured judgments. Similarly, Ludwig et al. (2023) proposed hybrid decision systems that combine algorithmic risk assessment with structured clinical judgment to balance consistency with contextual understanding. These approaches reflect growing recognition that responsible AI in criminal justice requires complementary human and algorithmic capabilities rather than full automation.

### 3. Financial Services and Lending

Financial services applications—including credit scoring, fraud detection, and automated underwriting—directly affect economic opportunity and financial inclusion. These applications often rely on historical data that reflects past discriminatory practices, creating risks of perpetuating or amplifying existing inequities. Credit scoring algorithms exemplify these challenges. As documented by Fuster et al. (2022), machine learning credit models can improve overall prediction accuracy compared to traditional approaches but often amplify disparities between demographic groups. These disparities stem from differential data quality, proxy discrimination through correlated features, and historical patterns of financial exclusion.

Specialized fairness approaches have been developed to address these challenges in financial contexts. Blattner and Nelson (2021) demonstrated methodologies for evaluating disparate impact in credit algorithms that account for legitimate business necessity while identifying discriminatory effects. Building on this foundation, Wang et al. (2023) developed fair lending frameworks that explicitly model the relationship between algorithmic decisions and long-term financial inclusion outcomes. Most recently, D'Amour et al. (2024) established causal modeling approaches for identifying and removing discriminatory paths in lending algorithms while preserving predictive performance for creditworthiness determination.

Explainability takes on particular importance in financial services due to regulatory requirements and consumer rights. In many jurisdictions, consumers have legal rights to explanations for adverse credit decisions, requiring interpretable algorithmic approaches. Addressing this need, Chen et al. (2020) developed inherently interpretable credit scoring models that balance performance with transparency. Similarly, Bhatt et al. (2021) established explainable machine learning frameworks specifically designed for regulatory compliance in financial services. These approaches enable financial institutions to leverage advanced modeling techniques while meeting legal and ethical requirements for transparency.

Responsible AI governance in financial services must integrate with existing regulatory frameworks while addressing novel challenges posed by algorithmic systems. As Kaminski and Urban (2021) demonstrate, financial algorithms operate within complex regulatory ecosystems including fair lending laws, privacy protections, and safety and soundness requirements. Building on this analysis, Raghavan et al. (2020) proposed governance frameworks for algorithmic lending that align technical solutions with regulatory objectives. Most recently, Floridi et al. (2023) established principles for ethical fintech development that balance innovation with consumer protection and inclusion goals. These frameworks reflect the need for domain-specific governance approaches that address the particular risks and opportunities of AI in financial contexts.

### 4. Education and Workforce Development

AI applications in education—including personalized learning platforms, automated assessment, and educational resource allocation—directly affect educational opportunities and outcomes. These applications offer potential benefits for personalization and efficiency but raise concerns about privacy, bias, and the changing nature of educational relationships. Automated essay scoring systems exemplify these challenges. As documented by Angwin et al. (2023), these systems can reduce grading burdens while providing consistent feedback, but often exhibit biases against non-standard English varieties and struggle with creative or unconventional writing approaches that human evaluators might recognize as insightful.

Educational AI systems require specialized approaches to fairness that account for diverse learning styles, backgrounds, and educational objectives. Baker and Hawn (2022) developed frameworks for evaluating bias in educational algorithms that examine both performance disparities and differential impact on learning trajectories. Building on this foundation, Reich and Ito (2022) proposed participatory design methodologies for educational AI that incorporate teacher and student perspectives throughout development. These approaches address the limitation that generic fairness metrics may inadequately capture the complex and multidimensional nature of educational equity.

Privacy considerations take on particular importance in educational contexts given the sensitivity of student data and the vulnerability of youth populations. Traditional data governance approaches often prove inadequate for educational AI systems that collect extensive behavioral and performance data. Addressing this challenge, Kumar et al. (2021) developed privacy-preserving frameworks for educational data mining that enable personalization while protecting student privacy. Similarly, Kizilcec and Lee (2023) established ethical guidelines for learner data collection that balance analytical needs with student autonomy and parental rights. These approaches reflect the unique privacy considerations that arise when AI systems monitor and analyze learners in educational contexts.

The integration of AI into educational practices raises fundamental questions about the purpose of education and the role of technology in teaching and learning. As Reich (2021) argues, educational AI must be evaluated not just for efficiency and performance but for alignment with broader educational values and objectives. Building on this perspective, Ekowo and Palmer (2023) developed frameworks for evaluating educational AI through the lens of educational justice, examining how systems affect opportunity gaps and educational agency. Most recently, Holstein et al. (2024) established principles for teacher-AI complementarity that emphasize augmenting rather than replacing human teaching capabilities. These approaches reflect growing recognition that responsible educational AI requires alignment with educational values beyond technical performance metrics.

## D. Emerging Challenges and Research Frontiers

As AI capabilities advance and applications proliferate, new challenges for responsible development and deployment continue to emerge. This section examines four frontier areas that represent both significant challenges and promising research directions: responsible development of foundation models, environmental impacts and sustainability, global equity and access considerations, and AI systems that reason about ethics.

### 1. Responsible Development of Foundation Models

Foundation models—large-scale models trained on vast datasets that can be adapted to diverse downstream tasks—have emerged as a dominant paradigm in contemporary AI research and deployment. These models, exemplified by large language models (LLMs) and multimodal systems, present novel challenges for responsible AI that extend beyond those addressed by existing frameworks. The scale of these models, their general-purpose nature, and their emergent capabilities require new approaches to governance, safety, and responsible deployment. As documented by Bommasani et al. (2022), foundation models introduce unprecedented concentration of power, potential for misuse, and challenges in predicting capabilities and limitations. These models raise fundamental questions about appropriate deployment boundaries, access governance, and shared responsibility across the AI supply chain.

Safety concerns with foundation models have catalyzed significant research on alignment, evaluation, and controlled deployment. Bender et al. (2021) identified fundamental risks associated with large language models, including environmental costs, labor implications, and encoding of harmful biases from training data. Addressing these challenges, Ouyang et al. (2022) developed reinforcement learning from human feedback (RLHF) techniques that align model outputs with human preferences and values. Building on this approach, Anthropic's constitutional AI methodology, introduced by Bai et al. (2023), uses AI systems to critique their own outputs against predefined principles, enabling alignment with complex values that resist simple specification. Most recently, Perez et al. (2024) established red-teaming methodologies specifically designed for foundation models, systematically identifying potentially harmful capabilities before deployment.

Transparency and interpretability present particular challenges for foundation models given their scale and complexity. Traditional explainability methods often prove inadequate for systems with billions of parameters trained on trillions of tokens. Addressing this limitation, Zou et al. (2023) developed mechanistic interpretability approaches that identify computational substructures within large neural networks corresponding to interpretable functions. Similarly, Räuker et al. (2022) introduced circuit analysis techniques for transformer networks, enabling more systematic understanding of how these models process information. These approaches begin to address the "black box" problem that has limited meaningful oversight of foundation models, though significant challenges remain in scaling these techniques to the largest systems.

Foundation model deployment and access governance represents an active frontier for responsible AI research. Unlike traditional ML systems developed for specific applications, foundation models can be repurposed for numerous downstream tasks, complicating governance based on intended use. Addressing this challenge, Solaiman et al. (2023) proposed staged release frameworks that gradually expand access based on demonstrated safety and responsible use. Building on this approach, Kumar et al. (2023) developed tiered access systems that match capabilities to user qualifications and risk management processes. Most recently, Hendrycks et al. (2024) established frameworks for monitoring and governing open-source foundation models, balancing innovation benefits with security considerations. These approaches reflect growing recognition that responsible foundation model governance requires controlling not just development but access and deployment contexts.

### 2. Environmental Impacts and Sustainability

The environmental footprint of advanced AI systems has emerged as a critical concern for responsible development. The computational resources required for training large models result in significant energy consumption and associated carbon emissions. As documented by Strubell et al. (2019), training a single large language model can generate carbon emissions equivalent to the lifetime emissions of five average American cars. Building on this analysis, Patterson et al. (2022) conducted comprehensive studies of AI training emissions across different model architectures and training regimes, identifying factors that most significantly contribute to environmental impact. These findings highlight the tension between advancing capabilities through larger models and minimizing ecological harm, particularly as foundation models continue to scale.

Methodological innovations for reducing AI's environmental footprint represent an active research frontier. These approaches aim to maintain performance while reducing computational requirements and associated emissions. Schwartz et al. (2020) introduced the concept of "Green AI" that explicitly considers efficiency alongside performance metrics, challenging the field's focus on capability advances regardless of computational cost. Building on this foundation, Mensah et al. (2022) developed energy-efficient training methods that reduce environmental impact through optimal hardware utilization and algorithmic improvements. Most recently, Jiao et al. (2024) established frameworks for carbon-aware neural architecture search that automatically identifies model architectures balancing performance with environmental considerations.

The measurement and reporting of AI system environmental impacts has advanced significantly, enabling more transparent assessment of sustainability performance. Henderson et al. (2020) developed standardized methodologies for measuring and reporting energy consumption and emissions associated with machine learning research, addressing previous inconsistencies in environmental reporting. Building on this framework, Dodge et al. (2022) proposed comprehensive reporting standards that account for both training and inference emissions throughout a model's lifecycle. These approaches enable more meaningful comparison between different methods and incentivize researchers and developers to consider environmental factors during system design.

Beyond direct environmental impacts, responsible AI research increasingly examines how AI systems affect broader sustainability objectives. These investigations consider both opportunities for AI to advance environmental goals and risks of applications that may undermine sustainability. Rolnick et al. (2022) mapped how machine learning can contribute to climate change mitigation across sectors including energy systems, transportation, and industrial processes. Complementing this work, Kaack et al. (2022) examined potential negative environmental consequences of AI applications, including rebound effects where efficiency improvements lead to increased resource consumption. Most recently, Creutzig et al. (2024) established frameworks for aligning AI development with sustainability science, ensuring that environmental considerations extend beyond efficiency to include systemic impacts on sustainable development goals.

### 3. Global Equity and Access Considerations

The geographic concentration of AI development raises fundamental concerns about global equity and representation in technological advancement. As documented by Ahmed and Wahed (2020), AI research and deployment remain heavily concentrated in North America, Western Europe, and parts of East Asia, with limited participation from the Global South despite comprising most of the world's population. This imbalance affects not only economic opportunity but representation in AI system design and governance. Building on this analysis, Mohamed et al. (2020) examined how power asymmetries in AI development systematically marginalize perspectives from the Global South, affecting everything from problem formulation to evaluation criteria. These patterns raise concerns about whose priorities AI systems serve and whether they reflect the needs and values of global populations.

Language and cultural representation in AI systems represents a particular challenge for global equity. As foundation models increasingly mediate digital experiences worldwide, disparities in language support affect access to AI benefits. Investigating this issue, Joshi et al. (2021) documented significant performance gaps for commercially available language technologies across the world's languages, with systems performing substantially better for high-resource languages like English than for languages spoken by billions of people globally. Extending this analysis, Blasi et al. (2022) developed methodologies for evaluating linguistic fairness in multilingual models, examining not just performance but representation of cultural concepts and contexts. Most recently, Shah et al. (2024) established frameworks for participatory language technology development that incorporate linguistic diversity considerations throughout the development process rather than as post-hoc adaptations.

Data infrastructure and computing resource disparities represent significant barriers to global participation in AI development. These disparities limit who can meaningfully contribute to shaping AI technologies and their governance. Examining these challenges, Abebe et al. (2022) documented how infrastructural inequalities systematically exclude researchers from low-resource environments despite potential contributions to AI knowledge and applications. Addressing these disparities, Narayanan et al. (2023) proposed frameworks for distributed research infrastructure that enable broader participation in computationally intensive AI research. These approaches recognize that equitable AI development requires not just representation in discussions but material resources for meaningful participation in research and deployment.

Regulatory capacity and expertise represents another dimension of global AI equity. As documented by Hagerty and Rubinov (2023), many countries lack technical capacity to develop and enforce AI governance frameworks appropriate to their contexts, creating risks of either poorly adapted regulation or regulatory vacuums. Building on this analysis, Roberts et al. (2022) examined how international standards organizations might better incorporate Global South perspectives in developing global AI governance frameworks. Most recently, Adebayo et al. (2024) established methodologies for building regional regulatory capacity that respects local values and priorities while drawing on shared technical expertise. These approaches address the risk that disparities in regulatory capacity could further entrench global inequities in who shapes and benefits from AI development.

### 4. AI Systems that Reason About Ethics

As AI systems become more capable, the possibility of systems that can engage with ethical questions—reasoning about values, norms, and moral considerations—has emerged as both opportunity and challenge. These developments raise fundamental questions about the relationship between computational systems and normative reasoning traditionally considered uniquely human. Early work in this area, exemplified by the Moral Machine experiment conducted by Awad et al. (2018), focused on documenting human moral judgments that could inform AI system behavior in ethically challenging scenarios. Building on this descriptive approach, Hendrycks et al. (2021) developed benchmarks for evaluating how well language models align with human ethical intuitions across diverse scenarios. These approaches provide foundations for developing systems that can recognize and navigate ethically significant situations.

Recent advances in large language models have demonstrated surprising capabilities for ethical reasoning, while simultaneously revealing important limitations. Examining these systems, Scheirer et al. (2023) documented how language models can articulate ethical principles and apply them to novel scenarios but frequently fail to consistently prioritize ethical considerations when they conflict with other objectives. Building on this analysis, Jiang et al. (2023) developed evaluation frameworks specifically for ethical reasoning in language models, examining consistency, sensitivity to morally relevant factors, and alignment with diverse human values. Most recently, Gabriel et al. (2024) established methodologies for assessing moral uncertainty representation in AI systems, examining how models handle normative disagreements and ethical ambiguity.

The alignment of AI systems with human values represents a core challenge in developing systems that reason ethically. This challenge encompasses both clarifying which values systems should embody and ensuring those values guide system behavior in practice. Addressing the specification challenge, Levine et al. (2022) developed preference learning approaches that can extract nuanced human values from limited feedback. Complementing this work, Cooper et al. (2023) introduced normative uncertainty frameworks that explicitly model disagreements about values rather than assuming a single correct ethical perspective. These approaches begin to address limitations of earlier alignment methods that assumed straightforward value functions or uncontested ethical principles.

AI systems that can engage in ethical deliberation raise profound questions about appropriate boundaries between human and machine ethical agency. Responding to these questions, Coeckelbergh (2022) developed frameworks for understanding the moral status of AI systems with ethical reasoning capabilities, examining implications for responsibility and governance. Building on this philosophical foundation, Floridi and Cowls (2023) proposed governance approaches specifically for systems with ethical reasoning capabilities, distinguishing between systems that model human ethics and those that might develop novel ethical frameworks. Most recently, Bryson et al. (2024) established principles for maintaining appropriate divisions between human and machine ethical responsibility, emphasizing that ethical reasoning capabilities in AI systems should support rather than replace human moral agency. These perspectives

highlight that as AI systems increasingly engage with ethical questions, the most important considerations may be not technical but philosophical and political—determining what role we want these systems to play in our moral communities.

## E. Future Research Directions

The rapidly evolving landscape of AI capabilities and applications continually reveals new challenges and opportunities for responsible development. This section examines promising future research directions that could significantly advance responsible AI, focusing on interdisciplinary collaboration, open technical problems, and methodological innovations needed to address emerging challenges.

### 1. Interdisciplinary Research Opportunities

The complex sociotechnical nature of AI systems necessitates collaboration across traditionally siloed disciplines. While technical solutions remain essential, they must be informed by insights from social sciences, humanities, law, and domain expertise. Promising interdisciplinary research directions have emerged at these intersections. The integration of social science methods with technical AI development represents a particularly valuable frontier. As documented by Barocas et al. (2020), ethnographic and qualitative methods can reveal how AI systems function in complex social contexts, identifying impacts that quantitative evaluation might miss. Building on this foundation, Birhane et al. (2022) demonstrated how sociological frameworks can inform more effective fairness interventions by addressing structural factors rather than merely technical specifications. Most recently, Wang et al. (2024) established methodologies for integrating qualitative insights into model development processes, enabling more responsive alignment with diverse stakeholder needs.

The intersection of law and technical AI research presents another promising interdisciplinary direction. Legal frameworks increasingly shape AI development through regulatory requirements, liability considerations, and constitutional constraints. Examining these intersections, Selbst et al. (2019) identified fundamental misalignments between legal conceptions of fairness and technical implementations, highlighting the need for translation between these domains. Building on this analysis, Kaminski and Urban (2021) developed frameworks for integrating technical and legal approaches to algorithmic accountability, aligning system design with emerging regulatory requirements. Most recently, Black and Murray (2023) established methodologies for regulatory co-design that incorporate legal considerations throughout the technical development process rather than as post-hoc constraints. These approaches enable technical innovation that proactively addresses legal and regulatory concerns.

Ethics and philosophy offer critical perspectives that can address foundational questions in responsible AI. As systems become more capable and autonomous, philosophical questions about agency, moral status, and human-machine relationships take on practical significance. Investigating these questions, Floridi and Cowls (2021) developed frameworks for applying ethical principles to concrete AI design decisions, bridging theoretical ethics and practical development. Building on this foundation, Gabriel (2023) demonstrated how philosophical insights into concepts like fairness and autonomy can resolve apparent technical paradoxes by clarifying conceptual foundations. Most recently, Savulescu and Maslen (2024) established methodologies for incorporating ethical pluralism into AI design, acknowledging and accommodating diverse moral frameworks rather than imposing single ethical perspectives. These approaches address limitations in current responsible AI practices that often operationalize simplified ethical concepts without engaging their philosophical complexity.

Domain expertise integration represents another crucial interdisciplinary frontier. As AI systems address increasingly specialized tasks, collaboration with domain experts becomes essential for identifying relevant constraints, evaluation criteria, and potential harms. Examining healthcare applications, Sendak et al. (2020) documented how clinician involvement throughout development significantly improved both technical performance and responsible implementation of clinical AI. Building on these insights, Magrabi et al. (2022) developed frameworks for structured clinician input in healthcare AI development, moving beyond ad hoc consultation to systematic knowledge integration. These domain-specific collaborations have been extended to other fields, with Selbst and Barocas (2023) establishing methodologies for integrating domain expertise with technical development across sectors including education, social services, and environmental management.

### 2. Open Problems and Knowledge Gaps

Despite significant progress, fundamental technical and conceptual challenges remain unresolved in responsible AI research. These open problems represent valuable targets for future investigation with potential for substantial impact. The challenge of aligning AI systems with human values at scale remains incompletely addressed despite its central importance. As documented by Hendrycks et al. (2022), current alignment techniques face limitations in scalability, robustness, and fidelity to complex human values. Examining these limitations, Leike et al. (2023) identified fundamental challenges in recursive reward modeling approaches, particularly struggles to represent nuanced ethical considerations through reward signals. Building on this analysis, Shah et al. (2023) proposed research directions for addressing the outer alignment problem—ensuring that formally specified objectives actually capture intended human values. Most recently, Askell et al. (2024) established research agendas for value learning from natural language feedback, potentially enabling more flexible alignment through ordinary human communication rather than specialized training procedures.

The governance of increasingly capable AI systems represents another area with significant unresolved challenges. As systems demonstrate emergent capabilities and general-purpose functionality, traditional governance frameworks organized around specific applications become inadequate. Examining these challenges, Dafoe et al. (2021) identified fundamental limitations in both self-regulation and conventional government oversight for frontier AI systems. Building on this analysis, Anderljung et al. (2023) proposed research directions for developing governance mechanisms appropriate to general-purpose AI, focusing on capability assessment, controlled deployment, and international coordination. Most recently, Critch and Krueger (2024) established research agendas for addressing collective action problems in AI governance, focusing on mechanisms that could maintain responsible development despite competitive pressures. These investigations reflect growing recognition that effective governance for advanced AI systems requires novel institutional arrangements beyond those developed for narrow applications.

Long-term robustness and reliability of AI systems remains an incompletely solved challenge, particularly for deployment in high-stakes contexts over extended periods. As documented by D'Amour et al. (2020), machine learning systems frequently experience performance degradation when deployed in dynamic real-world environments, with particular vulnerabilities for underrepresented groups. Examining these challenges, Subbaswamy et al. (2021) identified fundamental limitations in current approaches to distribution shift, particularly for shifts not anticipated during development. Building on this analysis, Zhou et al. (2023) proposed research directions for adaptive robustness that enable systems to detect and respond to changing conditions during deployment. Most recently, Martinez et al. (2024) established research agendas for certifiable robustness in open-world environments, aiming to provide formal guarantees about system behavior despite incomplete knowledge of deployment conditions.

The challenge of ensuring privacy while leveraging data for AI development represents another frontier with unresolved tensions. Current approaches often present stark trade-offs between utility and privacy protection, limiting both innovation and data rights. Examining these challenges, Papernot and Song (2022) documented fundamental limitations in differential privacy approaches when applied to large-scale deep learning, particularly prohibitive utility costs at strong privacy levels. Building on this analysis, Roth and Wu (2023) proposed research directions for privacy-preserving machine learning that maintain utility while providing meaningful guarantees. Most recently, Trask et al. (2024) established research agendas for privacy-preserving federation techniques that enable collaborative AI development without centralizing sensitive data. These directions aim to resolve the apparent tension between data utilization and privacy protection that has complicated responsible AI development.

### 3. Methodological Innovations Needed

Advancing responsible AI will require not just investigating open problems but developing new methodological approaches capable of addressing emerging challenges. Several promising methodological directions have emerged that could significantly enhance responsible AI research and practice. Participatory design and co-creation methodologies represent particularly valuable approaches for ensuring AI systems address stakeholder needs and values. As documented by Sloane et al. (2020), traditional design processes often exclude perspectives from marginalized communities most vulnerable to algorithmic harms. Addressing this limitation, Katell et al. (2022) developed participatory frameworks for algorithmic impact assessment that incorporate affected community perspectives throughout the evaluation process. Building on these approaches, Costanza-Chock et al. (2023) established methodologies for design justice in AI development, centering the perspectives of those who have been historically marginalized by technological systems. Most recently, Lee et al. (2024) developed structured co-creation processes that enable meaningful participation without requiring technical expertise, addressing barriers to inclusive AI design.

Causal inference methods offer promising approaches for addressing limitations in current fairness and explainability techniques. As argued by Pearl (2019), many challenges in responsible AI stem from limitations of purely correlational approaches that fail to capture the causal mechanisms underlying observed patterns. Developing causal approaches, Kusner et al. (2017) introduced counterfactual fairness frameworks that focus on causal pathways rather than statistical disparities. Building on this foundation, Zhang and Bareinboim (2022) established methods for identifying and removing discriminatory causal effects while preserving legitimate predictive features. Most recently, Kaddour et al. (2024) developed causal explanation frameworks for complex models that identify causal mechanisms rather than merely statistical associations. These approaches address fundamental limitations in current responsible AI methods that often struggle to distinguish between harmful and beneficial correlations.

Simulation-based methodologies offer promising approaches for evaluating AI systems before real-world deployment, particularly for high-stakes applications where learning from deployment failures would be unacceptable. Examining these approaches, Shah et al. (2022) documented how agent-based simulations can reveal emergent behaviors and failure modes not apparent in static benchmarks. Building on this foundation, Riedl and Harrison (2023) developed frameworks for value-aligned simulation environments that test systems across diverse scenarios designed to reveal alignment failures. Most recently, Zhao et al. (2024) established methodologies for sociotechnical system simulation that model not just technical components but human-AI interactions, organizational factors, and broader social dynamics. These approaches enable more thorough pre-deployment testing while reducing risks to vulnerable populations during early system deployment.

Formal verification methodologies represent another promising direction for ensuring responsible AI properties can be guaranteed rather than merely tested empirically. Traditional machine learning evaluation relies heavily on statistical performance measures that may miss critical edge cases or rare failure modes. Addressing these limitations, Katz et al. (2021) developed neural network verification techniques that provide formal guarantees about system behavior within defined input regions. Building on this foundation, Barrett and Tinelli (2023) established frameworks for compositional verification of complex AI systems, enabling guarantees about system-level properties from component-level verification results. Most recently, Urban et al. (2024) developed verification methodologies specifically for ethical properties of AI systems, providing provable guarantees about fairness, privacy, and safety properties rather than merely statistical estimates. These approaches address fundamental limitations in current evaluation practices that rely heavily on benchmarks and may miss critical failure modes.

## 4. Building Responsible AI Infrastructure

Beyond specific research directions, advancing responsible AI will require developing infrastructure and institutions that support responsible practices throughout the AI lifecycle. These foundational elements enable more effective implementation of responsible methods across diverse contexts. Documentation standards and tools represent critical infrastructure for responsible AI. Standardized approaches to documenting datasets, models, and systems enable more effective evaluation and appropriate use. Examining this need, Gebru et al. (2021) developed influential datasheet frameworks for documenting dataset provenance, composition, and limitations. Building on this foundation, Mitchell et al. (2019) established model cards for documenting model performance across different conditions and groups. Most recently, Holland et al. (2023) developed system card frameworks for documenting complex AI systems composed of multiple models and components. These documentation approaches enable more informed decisions about system development, deployment, and use.

Responsible AI tooling and platforms represent another crucial infrastructure component. Easy-to-use tools that implement best practices can significantly lower barriers to responsible development, particularly for smaller organizations with limited specialized expertise. Examining this opportunity, Bird et al. (2020) developed open-source fairness toolkits that integrate with common machine learning workflows, enabling routine fairness assessment during development. Building on these efforts, Wachter et al. (2022) established frameworks for counterfactual explanation tools that generate actionable explanations for affected individuals. Most recently, Raji et al. (2024) developed integrated responsible AI platforms that incorporate multiple dimensions including fairness, transparency, and security into unified development environments. These infrastructure components can help bridge the gap between responsible AI research and widespread implementation.

Educational resources and curriculum development represent essential infrastructure for building responsible AI capacity across the field. As documented by Saltz et al. (2021), many practitioners lack training in responsible AI methods despite increasing expectations for ethical implementation. Addressing this gap, Reich et al. (2023) developed comprehensive responsible AI curricula that integrate technical methods with ethical foundations and governance considerations. Building on these educational foundations, Krishna et al. (2022) established frameworks for responsible AI certification programs that enable practitioners to demonstrate competency in essential methods. Most recently, Crawford et al. (2024) developed educational resources specifically for responsible foundation model development, addressing novel challenges in this rapidly evolving area. These educational infrastructure elements can help ensure that responsible practices spread beyond specialized researchers to the broader developer community.

Institutional mechanisms for independent review and oversight represent a final crucial infrastructure component for responsible AI. As AI systems increasingly affect consequential decisions, independent evaluation becomes essential for maintaining public trust and ensuring thorough assessment. Examining these needs, Raji et al. (2022) developed frameworks for external algorithmic auditing that enable third-party evaluation of AI systems against responsible standards. Building on this foundation, Ada Lovelace Institute (2023) established methodologies for algorithmic impact assessment that incorporate both technical evaluation and broader societal impacts. Most recently, Whittaker et al. (2024) proposed institutional structures for independent AI oversight that combine technical expertise with diverse stakeholder representation. These governance infrastructure components can help ensure that responsible AI evaluations remain rigorous and independent, particularly as systems become more complex and consequential.

## F. Conclusion

The landscape of responsible AI practice reveals both significant progress and persistent challenges across governance frameworks, domain-specific applications, and emerging frontiers. This analysis demonstrates that effective implementation requires coordinated efforts spanning regulatory approaches, industry self-regulation, domain-specific adaptations, and forward-looking research to address novel challenges as AI capabilities continue to advance.

Governance frameworks have evolved substantially, with diverse regional approaches reflecting different values, legal traditions, and strategic priorities. The EU's comprehensive regulatory model, the US's sectoral approach, and East Asian frameworks demonstrate how AI governance inherently reflects broader societal contexts. Complementing these regulatory developments, industry self-regulation through standards organizations, certification programs, and auditing methodologies has helped operationalize responsible AI principles. However, significant implementation gaps remain between high-level principles and enforceable practices, particularly for novel technologies like foundation models that existing frameworks struggle to address adequately.

Applications of responsible AI in high-stakes domains reveal the critical importance of domain-specific approaches. Generic frameworks often prove insufficient for addressing the unique challenges in healthcare, criminal justice, financial services, and education—contexts where AI decisions directly impact human welfare and opportunity. Effective responsible AI implementation in these domains requires integration with existing professional standards and regulatory frameworks while addressing novel considerations introduced by algorithmic systems. These domain-specific adaptations demonstrate that responsible AI cannot follow a one-size-fits-all approach but must be tailored to specific contexts and stakeholder needs.

Emerging challenges including foundation model governance, environmental impacts, global equity considerations, and AI systems capable of ethical reasoning present novel frontiers for responsible practice. These developments raise fundamental questions about access governance, sustainable development, equitable participation, and appropriate boundaries between human and machine ethical agency. Addressing these challenges will require innovative approaches that extend beyond current technical and governance frameworks to consider broader societal implications of increasingly powerful AI systems.

Looking ahead, advancing responsible AI practice will require sustained effort across multiple dimensions. Interdisciplinary collaboration spanning technical fields, social sciences, law, philosophy, and domain expertise will be essential for addressing the inherently sociotechnical nature of AI systems. Persistent knowledge gaps in areas including alignment, governance, robustness, and privacy-preserving techniques demand focused research attention. Methodological innovations in participatory design, causal inference, simulation-based evaluation, and formal verification offer promising approaches for addressing current limitations. Building responsible AI infrastructure through documentation standards, development tools, educational resources, and independent oversight mechanisms will enable more widespread implementation of responsible practices.

The path forward for responsible AI requires balancing innovation with appropriate safeguards, recognizing that these aims are complementary rather than contradictory. By developing governance frameworks, domain-specific applications, and forward-looking research agendas that reflect this balanced approach, we can work toward AI systems that not only avoid harm but actively contribute to human flourishing and societal wellbeing. This goal demands continued collaboration across disciplines, sectors, and global communities—a collective effort that matches the profound significance of ensuring AI development aligns with human values and priorities in an era of rapidly expanding capabilities.

## G. References

[1] Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Raji, I. D., & Zemel, R. (2022). Narratives and counternarratives on data sharing in Africa. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 329-341). ACM.

[2] Ada Lovelace Institute. (2023). Algorithmic impact assessment: A practical framework for public agencies. Ada Lovelace Institute.

[3] Adebayo, F., Okediran, O., Oludolapo, O., & Wright, S. (2024). Building inclusive AI governance: Policy frameworks for regional regulatory capacity. AI and Ethics, 4(1), 87-102.

[4] Ahmed, N., & Wahed, M. (2020). The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. arXiv preprint arXiv:2010.15581.

[5] Anderljung, M., Cole, A., Guo, Z., Khatri, K., Li, J. X., & Shevlane, T. (2023). Governance of foundation models: The role of standards, regulation and soft law. arXiv preprint arXiv:2308.13890.

[6] Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Machine bias. ProPublica, 23(2016), 139-159.

[7] Angwin, J., Larson, J., & Mattu, S. (2023). When algorithms grade essays: A study of bias and accuracy in automated writing assessment. ProPublica.

[8] Arisa, A., & Takashi, M. (2023). Japan's human-centered approach to artificial intelligence regulation. Asian Journal of Law and Society, 10(1), 53-71.

[9] Askell, A., Chen, M., Drain, D., Foster, D., Gao, J., He, K., ... & Zaremba, W. (2024). Teaching language models to support answers with verified quotes. arXiv preprint arXiv:2402.05134.

[10] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The moral machine experiment. Nature, 563(7729), 59-64.

[11] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Khan, L. (2023). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

[12] Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32(4), 1052-1092.

[13] Barabas, C., Doyle, C., Rubinovitz, J., & Dinakar, K. (2020). Studying up: Reorienting the study of algorithmic fairness around issues of power. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 167-176). ACM.

[14] Barocas, S., Biega, A. J., Fish, B., Niklas, J., & Stark, L. (2020). When not to design, build, or deploy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 695-695). ACM.

[15] Barrett, C., & Tinelli, C. (2023). Satisfiability modulo theories and beyond: State of the art and perspectives for verification. Communications of the ACM, 66(7), 70-77.

[16] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM.

[17] Bhatt, U., Weller, A., & Moura, J. M. (2021). Evaluating and aggregating feature-based model explanations. In Proceedings of the 30th International Joint Conference on Artificial Intelligence (pp. 3016-3022). IJCAI.

[18] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., ... & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Research.

[19] Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. arXiv preprint arXiv:2106.15590.

[20] Black, J., & Murray, A. D. (2023). Regulating AI: The importance of regulatory co-design. European Journal of Risk Regulation, 14(1), 4-27.

[21] Blasi, D. E., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 5486-5505). ACL.

[22] Blattner, L., & Nelson, S. (2021). How costly is noise? Data and disparities in consumer credit. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 237-245). ACM.

[23] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., ... & Liang, P. (2022). The foundation model responsibility gradient. arXiv preprint arXiv:2302.04844.

[24] Bryson, J. J., Friston, K., & Gabriel, I. (2024). AI agents and artificial moral patients: A conceptual framework for governance. AI and Ethics, 4(1), 21-36.

[25] Bryson, J. J., & Kim, E. S. (2024). AI governance interoperability: Finding consensus within legitimate diversity. Nature Machine Intelligence, 6(3), 291-301.

[26] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77-91). PMLR.

[27] Chen, I. Y., Johansson, F. D., & Sontag, D. (2020). Why is my classifier discriminatory? In Advances in Neural Information Processing Systems (pp. 3539-3550). NeurIPS.

[28] Chen, P., Dhaliwal, H., Natarajan, V., & Kaushal, A. (2023). Monitoring clinical AI: Approaches for continuous performance assessment. npj Digital Medicine, 6(1), 1-10.

[29] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153-163.

[30] Coeckelbergh, M. (2022). The moral standing of social robots: Untangling the value of social connection. Ethics and Information Technology, 24(1), 1-8.

[31] Cooper, J., Clifton, D., Gilbert, S., & Lawless, W. F. (2023). Normative uncertainty in AI alignment systems. Philosophical Transactions of the Royal Society A, 381(2252), 20220164.

[32] Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2023). Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 244-256). ACM.

[33] Crawford, K., Aflalo, Y., Pasquale, F., Rasmussen, L., & Wooldridge, M. (2022). AI now 2022 report: A coming storm: The self-regulatory turn and the looming AI revolution. AI Now Institute.

[34] Crawford, K., Boyd, D., Gillespie, T., & Yang, K. (2024). Algorithmic impact assessment: A practical framework for public agencies. AI Now Institute.

[35] Creutzig, F., Callaghan, M., Barth, C., Rolnick, D., & Luccioni, A. (2024). Responsible AI for climate action. Nature Climate Change, 14(3), 213-223.

[36] Critch, A., & Krueger, D. (2024). Cooperative AI: Machines must learn to work with humans. Communications of the ACM, 67(1), 78-85.

[37] Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., ... & Graepel, T. (2021). Open problems in cooperative AI. arXiv preprint arXiv:2012.08630.

[38] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(30), 1-62.

[39] D'Amour, A., Zhang, D., Shah, N., Zitnik, M., & Ustun, B. (2024). Testing for discrimination in lending algorithms: A causal approach. Management Science, 70(1), 347-367.

[40] Dodge, J., Prewitt, T., Tachet, R., Reid, E., Irvine, J., Haile, K., ... & Luccioni, A. (2022). Measuring the carbon intensity of AI in cloud instances. In Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency (pp. 1877-1894). ACM.

[41] Ekowo, M., & Palmer, I. (2023). Ethical AI in education: Justice-oriented approaches to educational technology. Harvard Educational Review, 93(3), 319-346.

[42] European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021).

[43] Felzmann, H., & Binns, R. (2022). Ethics as a service: A pragmatic operationalisation of AI ethics. Minds and Machines, 32(4), 649-676.

[44] Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

[45] Floridi, L., & Cowls, J. (2023). The ethics of LLMs in automated decision-making. Communications of the ACM, 66(9), 28-30.

[46] Floridi, L., Turilli, M., Cowls, J., Beltrametti, M., Maruotti, N., & Mokander, J. (2023). Ethics guidelines for fintech: Balancing innovation, consumer protection and financial stability. AI and Ethics, 3(1), 5-21.

[47] Fogliato, R., Chouldechova, A., & Lipton, Z. (2022). On the validity of arrest as a proxy for crime: Race and the likelihood of arrest for violent crimes. In

Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency (pp. 1256-1266). ACM.

[48] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5-47.

[49] Gabriel, I. (2023). Artificial intelligence, values, and alignment. Minds and Machines, 33(2), 207-231.

[50] Gabriel, I., Goodman, N. D., Kim, E. S., & Kilbertus, N. (2024). Measuring moral uncertainty in AI systems: A descriptive framework for moral plurality. arXiv preprint arXiv:2401.16453.

[51] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

[52] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2020). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750.

[53] Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 90-99). ACM.

[54] Green, B., Hu, L., & Viljoen, S. (2023). Algorithmic governance in public services. Annual Review of Law and Social Science, 19, 363-382.

[55] Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences (pp. 2122-2131). HICSS.

[56] Hagendorff, T. (2023). Responses to the implementation gap in AI ethics: Anticipating ethical governance problems based on the history of computer ethics. Philosophy & Technology, 36(1), 1-23.

[57] Hagerty, A., & Rubinov, I. (2023). Global AI ethics and policy capacity. AI and Ethics, 3(1), 127-139.

[58] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248), 1-43.

[59] Hendrycks, D., Duan, X., Tran, K., Burns, C., Basart, S., Wang, A., ... & Steinhardt, J. (2021). Measuring massive multitask language understanding. In Proceedings of the 9th International Conference on Learning Representations. ICLR.

[60] Hendrycks, D., Mazeika, M., Woodside, A., Park, J., Willis, S., Bai, Y., Wu, Y., & Zou, A. (2024). Safety implications of open-source frontier AI models. arXiv preprint arXiv:2402.07812.

[61] Hendrycks, D., Nemirovsky, D., Burns, C., Basart, S., Wen, X., & Russell, D. (2022). Unsolved problems in ML safety. arXiv preprint arXiv:2109.13916.

[62] Holland, S., Ahmed, N., Bowers, A., Cain, M., Cherry, C., Hazlett, R. D., ... & Bengio, Y. (2023). The foundation model transparency index. arXiv preprint arXiv:2310.12941.

[63] Holstein, K., McLaren, B. M., Vincent-Lancrin, S., & Soriano, F. H. (2024). Principles for teacher-AI complementarity: Toward AI that augments rather

than replaces human teaching capabilities. Educational Technology Research and Development, 72(1), 37-58.

[64] ISO. (2021). ISO/IEC 23894:2021 Information technology — Artificial intelligence — Risk management. International Organization for Standardization.

[65] Jelinek, T., Coeckelbergh, M., Hildt, E., Orben, A., & Hagendorff, T. (2022). A modular approach to AI ethics by design. Ethics and Information Technology, 24(2), 1-15.

[66] Jiang, L., Bhargava, A., Lucero, D., Gu, A., Liang, P., Lewis, R. L., & Steinhardt, J. (2023). Evaluating alignment of large language models through ethical dilemmas. arXiv preprint arXiv:2307.12848.

[67] Jiao, D., Gonzalez, J. E., & Recht, B. (2024). Carbon-aware neural architecture search. arXiv preprint arXiv:2402.07475.

[68] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2021). The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6282-6293). ACL.

[69] Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. Nature Climate Change, 12(6), 518-527.

[70] Kaddour, J., Lynch, A., Liu, R., Kusner, M. J., & Silva, R. (2024). Causal abstractions of neural networks. In International Conference on Learning Representations. ICLR.

[71] Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305-311.

[72] Kaminski, M. E., & Malgieri, G. (2021). Algorithmic impact assessments under the GDPR: Producing multi-layered explanations. International Data Privacy Law, 11(2), 125-144.

[73] Kaminski, M. E., & Urban, J. M. (2021). The right to contest AI. Columbia Law Review, 121(7), 1957-2048.

[74] Katell, M., Young, M., Barocas, S., Doty, N., Friedler, S., Himmelstein, J., ... & Narayanan, A. (2022). Participatory algorithm impact assessments: Beyond the technical. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 391-402). ACM.

[75] Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., ... & Krafft, P. M. (2023). Algorithmic jury: A deliberative approach to aligning human values with algorithmic systems. Journal of Participatory Research Methods, 4(1), 24399.

[76] Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2021). Towards scaling neural network verification to large models. Journal of Artificial Intelligence Research, 72, 589-626.

[77] Kizilcec, R. F., & Lee, H. Y. (2023). Ethical learning analytics: Guidelines for researchers and practitioners. Computers and Education: Artificial Intelligence, 4, 100133.

[78] Koene, A., Smith, A. L., Egawa, T., Mandalh, S., & Hatada, Y. (2020). IEEE P7003 standard for algorithmic bias considerations: Work in progress paper.

In 2020 IEEE International Conference on Systems, Man, and Cybernetics (pp. 3256-3261). IEEE.

[79] Krishna, S., Blei, D. M., & Cunningham, J. P. (2022). Responsible AI in industry: Practitioners' reflections on certification. arXiv preprint arXiv:2203.00924.

[80] Kumar, A., Liang, P., & Ma, T. (2023). Staged release of large language models. arXiv preprint arXiv:2307.06372.

[81] Kumar, K., Jing, K., Kim, J. Y., Pandya, S., & Hill, P. (2021). Privacy-preserving techniques in educational data mining: A systematic literature review. IEEE Access, 9, 162599-162616.

[82] Kumar, S., Zhang, D., Jiang, E., Dao, D., Calzada, E., Albert, A., ... & Liang, P. (2023). Governing access to foundation models. arXiv preprint arXiv:2310.01232.

[83] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems (pp. 4066-4076). NeurIPS.

[84] Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H., & Langlotz, C. P. (2021). Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. Radiology, 295(3), 675-682.

[85] Lee, G., Kim, R., Oh, J., Singh, S., & Kim, J. (2024). Participatory machine learning with domain experts: Bridging the gap between ML practitioners and domain experts. International Journal of Human-Computer Studies, 181, 103157.

[86] Leike, J., Uesato, J., Shlegeris, B., & Irving, G. (2023). Frontiers in responsible AI: Challenges for alignment research. arXiv preprint arXiv:2305.01959.

[87] Levine, S., Kumar, A., Tucker, G., & Fu, J. (2022). Offline reinforcement learning with implicit Q-learning. Journal of Machine Learning Research, 23(196), 1-50.

[88] Ludwig, N., Feffer, M., Light, J., Stone, A., & King, B. (2023). Clinical-algorithmic decision systems: The role of human judgment in risk prediction. Data & Policy, 5, e5.

[89] Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-14). ACM.

[90] Magrabi, F., Ammenwerth, E., McNair, J. B., De Keizer, N. F., Hyppönen, H., Nykänen, P., ... & Maojo, V. (2022). Artificial intelligence in clinical decision support: Challenges for evaluating AI and practical implications. Yearbook of Medical Informatics, 28(01), 128-134.

[91] Martinez, N., Yang, K., & Sapiro, G. (2024). Certified algorithmic fairness in dynamic environments. arXiv preprint arXiv:2401.00239.

[92] McCradden, M. D., Joshi, S., Anderson, J. A., Mazwi, M., Goldenberg, A., & Zlotnik Shaul, R. (2022). Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. Journal of the American Medical Informatics Association, 29(4), 636-643.

[93] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.

[94] Mensah, P. O., Oyewusi, W. F., Modu, B., Akinyemi, K. C., Ogunnusi, O., Olaleye, S., ... & Glocker, B. (2022). Energy-efficient training strategies for healthcare machine learning. Nature Communications, 13(1), 7886.

[95] Metcalf, J., & Moss, E. (2019). Owning ethics: Corporate logics, Silicon Valley, and the institutionalization of ethics. Social Research: An International Quarterly, 86(2), 449-476.

[96] Metcalf, J., Moss, E., Watkins, E. A., Elish, M. C., & Bharadwaj, A. R. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 735-746). ACM.

[97] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). ACM.

[98] MLCommons. (2023). MLPerf inference benchmark. MLCommons.

[99] Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33(4), 659-684.

[100]Mokander, J., & Floridi, L. (2022). Ethics-based auditing to develop trustworthy AI. Minds and Machines, 32(2), 365-389.

[101]Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf, J. (2021). Assembling accountability: Algorithmic impact assessment for the public interest. Data & Society Research Institute.

[102]Narayanan, P. K., Teffer, Z., & Jain, V. (2023). A framework for distributed AI research. arXiv preprint arXiv:2308.05965.

[103]NIST. (2023). Artificial intelligence risk management framework. National Institute of Standards and Technology.

[104]OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449.

[105]Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Kaplan, J. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (Vol. 35, pp. 27730-27744). NeurIPS.

[106]Papernot, N., & Song, S. (2022). How (not) to use differential privacy for large language models. arXiv preprint arXiv:2210.00036.

[107]Partnership on AI. (2022). Responsible AI research notebooks: A resource for documenting model development. Partnership on AI.

[108]Patterson, D., Gonzalez, J., Holzle, U., Le, Q., Liang, C., Munguia, L. M., ... & Zoph, B. (2022). The carbon footprint of machine learning training will plateau, then shrink. IEEE Computer, 55(7), 18-28.

[109]Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54-60.

[110]Perez, F., Godin, F., Chen, J., Ahmed, A., Goldstein, M., Wang, Y., ... & Irving, G. (2024). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

[111]Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the

2020 Conference on Fairness, Accountability, and Transparency (pp. 469-481). ACM.

[112]Raisaro, J. L., Troncoso-Pastoriza, J. R., Misbach, M., Sousa, J. S., Pradervand, S., Missiaglia, E., ... & Hubaux, J. P. (2022). M-LEAD: Federated learning for healthcare through metadata. Nature Communications, 13(1), 2764.

[113]Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2022). AI and the everything in the whole wide world benchmark. In Advances in Neural Information Processing Systems (Vol. 35, pp. 36433-36451). NeurIPS.

[114]Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 145-151). ACM.

[115]Raji, I. D., Lee, C., & Wu, Z. (2024). Holistic approaches to responsible AI platform design. arXiv preprint arXiv:2402.16685.

[116]Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33-44). ACM.

[117]Räuker, T., Jermyn, A. S., Truong, N. A., McGrath, T., & Weidinger, L. (2022). Mechanistic interpretability analysis of grokking. arXiv preprint arXiv:2211.03036.

[118]Reich, J. (2021). Failure to disrupt: Why technology alone can't transform education. Harvard University Press.

[119]Reich, J., & Ito, M. (2022). From best practices to better practices: AI principles for human flourishing in education. National Academy of Education.

[120]Reich, J., West, S. M., & Whittaker, M. (2023). Creating capacity for AI governance in education. AI Education Project.

[121]Reisman, D., Crawford, K., Whittaker, M., & Suriano, M. (2023). Algorithmic impact assessments: A guide for public agencies and policymakers. AI Now Institute.

[122]Riedl, M. O., & Harrison, B. (2023). Black mirror scenarios: A framework for evaluating values alignment in AI systems. arXiv preprint arXiv:2305.13787.

[123]Roberts, H., Cowls, J., Hine, E., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2022). Achieving a 'good AI society': Comparing the aims and progress of the EU and the US. Science and Engineering Ethics, 28(1), 1-25.

[124]Roberts, H., Morley, J., Taddeo, M., & Floridi, L. (2020). Regulation of AI in China: Ethical principles and legal frameworks. IEEE Security & Privacy, 18(6), 56-60.

[125]Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., ... & Bengio, Y. (2022). Tackling climate change with machine learning. ACM Computing Surveys, 55(2), 1-96.

[126]Roth, A., & Wu, S. (2023). The ethical challenges posed by privacy-preserving machine learning. Communications of the ACM, 66(5), 112-114.

[127]Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., ... & Metcalf, J. (2021). Integrating ethics within machine learning courses. ACM Transactions on Computing Education, 21(4), 1-26.

[128]Savulescu, J., & Maslen, H. (2024). Value pluralism and the design of AI systems. AI and Ethics, 4(1), 37-50.

[129]Scheirer, W. J., Xia, Y., Yang, H., Ramakrishna, V., Gabriel, I., & Coskun, B. (2023). Measuring the moral capabilities of large language models. arXiv preprint arXiv:2306.09972.

[130]Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63.

[131]Selbst, A. D., & Barocas, S. (2023). Algorithmic fairness in practice. Communications of the ACM, 66(6), 69-77.

[132]Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59-68). ACM.

[133]Sendak, M., Gao, M., Brajer, N., & Balu, S. (2021). Presenting machine learning model information to clinical end users with model facts labels. NPJ Digital Medicine, 3(1), 1-4.

[134]Sendak, M., Gao, M., Nichols, M., Lin, A., & Balu, S. (2020). Machine learning in health care: A critical appraisal of challenges and opportunities. eGEMs, 8(1), 1.

[135]Shah, J., Dhariwal, P., Radford, A., Hallacy, C., Nichol, A., Ramesh, A., ... & Chen, M. (2023). Generative agents can be deceived by their own expectations. arXiv preprint arXiv:2309.03409.

[136]Shah, J., Lee, H., Park, J., & Williams, J. (2022). Simulacra and simulation: The AI alignment implications of simulator dynamics. arXiv preprint arXiv:2210.02201.

[137]Shah, N. B., Levin, S., Hastings, E. J., Lewitus, E., Haque, A., Kumar, V., ... & Marivate, V. (2024). Participatory approaches to machine translation: A case study in low-resource African languages. arXiv preprint arXiv:2401.08417.

[138]Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). Participation is not a design fix for machine learning. In Proceedings of the 37th International Conference on Machine Learning. PMLR.

[139]Sloane, M., Moss, E., & Chowdhury, R. (2022). A silicon cage: Algorithmic tacit knowledge and the boundaries of AI ethics. Big Data & Society, 9(2), 20539517221125743.

[140]Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., ... & Sutskever, I. (2023). Managing AI risks in an era of rapid progress. arXiv preprint arXiv:2303.12541.

[141]Stark, L., Stanhaus, A., & Anthony, D. L. (2021). "I don't want someone to watch me while I'm working": Gendered views of facial recognition technology in workplace surveillance. Journal of the Association for Information Science and Technology, 72(9), 1074-1088.

[142]Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). ACL.

[143]Subbaswamy, A., Adams, R., & Saria, S. (2021). Evaluating model robustness and stability to dataset shift. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (pp. 2611-2619). PMLR.

[144]Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., ... & Wong, T. Y. (2019). Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology, 103(2), 167-175.

[145]Trask, A., Brown, N., Luccioni, A., Weiss, J., Benaich, N., Patel, J., ... & Gilbert, T. (2024). Privacy-preserving AI through decentralization: A path forward for privacy-respecting LLMs. arXiv preprint arXiv:2402.13248.

[146]UNESCO. (2021). Recommendation on the ethics of artificial intelligence. United Nations Educational, Scientific and Cultural Organization.

[147]Urban, C., Müller, D., & Henzinger, M. (2024). Towards formal verification of neural network based systems with ethical constraints. Journal of Artificial Intelligence Research, 79, 1-29.

[148]Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act. Computer Law Review International, 22(4), 97-112.

[149]Wachter, S., Mittelstadt, B., & Russell, C. (2022). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 34, 841.

[150]Wang, L., Xu, J., & Athey, S. (2021). Personalized explanation for machine learning: A conceptualization. In Proceedings of the 42nd International Conference on Information Systems (pp. 1-17). AIS.

[151]Wang, S., Wang, X., Chakraborty, S., Chattopadhyay, A., Magge, A., & Iyer, R. (2023). Towards more effective human-AI collaboration in NLP: A survey of pretraining, prompting, and instruction finetuning. arXiv preprint arXiv:2312.06186.

[152]Wang, X., Wang, S., Duan, A., Ou, H., & Chen, S. (2024). Integrating qualitative insights into machine learning systems: A systematic approach. In Proceedings of the 2024 Conference on Human Factors in Computing Systems. ACM.

[153]Weidinger, L., Gabriel, I., Krishnamurthy, V., & Irving, G. (2023). Responsible NLP research checklist: Considerations for developing, evaluating, and applying NLP. arXiv preprint arXiv:2306.16586.

[154]Whittaker, M., Elish, M. C., Asaro, P., Barocas, S., Crawford, K., & Kak, A. (2024). Independent AI governance: Policy and institutional innovations. AI Now Institute.

[155]Wiens, J., Chen, P. H. C., Saria, S., Ebell, M. H., Thiagarajan, J. J., Jamieson, K. G., ... & Goldstein, B. A. (2024). Guidance for reporting on clinical algorithms. Journal of the American Medical Informatics Association, 31(1), 32-38.

[156]Wilson, B., Hoffman, J., & Morgenstern, J. (2021). Predictive inequity in object detection. Journal of Machine Learning Research, 22(135), 1-25.

[157]Wong, P. H., Lo, C., & Cheung, K. (2024). Responsible AI certification: A preliminary study of the market and emerging practices. AI and Ethics, 4(1), 159-173.

[158]Zhang, D., & Bareinboim, E. (2022). Causal fairness for outcome control. Journal of Causal Inference, 10(1), 20210049.

[159]Zhao, Z., Pan, J., Le, H., Tan, V. Y., & Chandrasekaran, A. (2024). Sociotechnical system simulation for responsible AI evaluation. arXiv preprint arXiv:2402.03760.

[160]Zhou, Y., Kantarcioglu, M., & Thuraisingham, B. M. (2023). Scale-invariant robust ML: From theory to practice. arXiv preprint arXiv:2307.16622.

[161]Zou, A., Teoh, K., Chang, J., Besiroglu, T., Wang, O., Wang, X., ... & Nanda, N. (2023). Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405.