The Indonesian Journal of Computer Science
www.ijcs.net
Volume 14, Issue 2, April 2025
https://doi.org/10.33022/ijcs.v14i2.4789

Navigating the Frontier: Theoretical Frameworks and Technical Approaches to Responsible AI

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

Article Information

Received : 19 Mar 2025
Revised : 6 Apr 2025
Accepted : 15 Apr 2025

Keywords

Responsible AI, Theoretical Frameworks, Fairness, Bias Mitigation, Transparency, Explainability, Safety, Foundation Models/LLMs

Abstract

This paper provides a comprehensive analysis of responsible AI development, examining both theoretical foundations and practical implementations. It explores core ethical principles including fairness, accountability, transparency, and safety, while also addressing emerging concepts like autonomy, dignity, and solidarity. The research analyzes competing philosophical frameworks—consequentialist, deontological, and virtue ethics—and highlights tensions between universalist and particularist ethical perspectives. The paper documents regional variations in responsible AI approaches across Europe, the United States, and East Asia, noting the concerning underrepresentation of Global South perspectives. Technical advancements in fairness are thoroughly examined, including pre-processing, in-processing, and post-processing techniques, alongside newer fairness-aware deep learning methods involving attention mechanisms and transfer learning. The work further investigates transparency challenges, comparing local and global explainability methods, and addresses the unique interpretability issues posed by foundation models and large language models. Safety and alignment techniques are also explored, including robustness against adversarial attacks, constitutional AI approaches, and various value learning methodologies. The paper concludes by evaluating measurement frameworks and assessment strategies for responsible AI interventions, offering insights into evaluation frameworks, benchmarks, and longitudinal studies needed to advance the field

A. Introduction

Artificial intelligence (AI) has evolved from a niche scientific pursuit to a transformative technology reshaping virtually every sector of society. As AI systems increasingly make or inform decisions with significant human impact, the imperative for responsible development and deployment has moved from academic discourse to mainstream concern (Floridi & Cowls, 2021). The concept of responsible AI has undergone significant evolution over the past decade. Initially focused primarily on narrow technical definitions of fairness and transparency, the field has expanded to encompass broader considerations of power, justice, sustainability, and cultural context (Dignum, 2019).

The rapid advancement of foundation models and large language models (LLMs) since 2020 has dramatically heightened both the urgency and complexity of responsible AI. These systems demonstrate unprecedented capabilities while simultaneously introducing novel risks related to misinformation, privacy, labor displacement, and concentration of technological power (Crawford & Calo, 2016). Their emergence has catalyzed renewed attention to responsible AI from policymakers, industry leaders, and civil society around the globe.

This paper aims to provide a comprehensive assessment of the current state of responsible AI research and practice. We examine theoretical frameworks that guide the field, survey technical advances across key dimensions including fairness, transparency, safety, and privacy, and evaluate governance approaches at organizational and regulatory levels. Throughout, we identify persistent challenges, promising innovations, and critical gaps requiring further investigation (Mohamed et al., 2020). By synthesizing developments across this fragmented field, we aim to facilitate more coherent and effective approaches to ensuring AI benefits humanity.

As AI capabilities continue to advance, responsible AI practices must evolve in tandem. This paper concludes by articulating a research agenda that addresses emerging challenges and builds toward AI systems that not only avoid harm but actively contribute to human flourishing, environmental sustainability, and social justice. The state of responsible AI today represents meaningful progress, but also reveals how much work remains to align increasingly powerful AI systems with human values and societal wellbeing.

B. Theoretical Frameworks for Responsible AI

The development of robust theoretical frameworks for responsible AI has emerged as a crucial foundation for both research and practice. These frameworks provide structured approaches for identifying, analyzing, and addressing ethical challenges in AI systems. As Floridi and Cowls (2021) argue, effective ethical frameworks must balance theoretical soundness with practical applicability, guiding concrete decision-making throughout the AI lifecycle. This section examines key ethical principles, competing philosophical approaches, and global variations in responsible AI frameworks.

1. Key Ethical Principles

Four core principles have achieved broad consensus across numerous responsible AI frameworks: fairness, accountability, transparency, and safety

(FATE). While initially articulated by Diakopoulos and Friedler (2016) in narrower technical terms, these principles have expanded in scope and interpretation. Fairness encompasses equitable treatment across demographic groups and avoidance of discriminatory impacts. Accountability establishes structures for responsible oversight and redress mechanisms. Transparency enables understanding of AI systems' operation and decisions. Safety ensures systems perform reliably without causing physical or psychological harm (Whittlestone et al., 2021).

More recently, scholars have advocated for additional principles beyond the FATE framework. Dignum (2019) proposes autonomy, dignity, and solidarity as essential complements, emphasizing human agency and collective wellbeing. Meanwhile, Mohamed et al. (2020) argue that responsible AI frameworks must explicitly incorporate considerations of power, justice, and participation to address structural inequalities that technical solutions alone cannot remedy.

2. Competing Philosophical Approaches

Responsible AI frameworks draw from diverse philosophical traditions, leading to different emphases and approaches. Consequentialist frameworks, influenced by utilitarian ethics, prioritize outcomes and impacts of AI systems. This approach, exemplified by the work of Kasirzadeh and Gabriel (2023), focuses on measuring and maximizing beneficial consequences while minimizing harms across affected populations. In contrast, deontological frameworks, drawing from Kantian ethics, emphasize rights, duties, and adherence to ethical rules regardless of consequences. As argued by Jobin et al. (2019), rights-based approaches have gained particular traction in European contexts, influencing both regulatory frameworks and corporate ethics guidelines.

Virtue ethics offers a third philosophical approach, focusing on the character and values embedded in AI systems and the organizations that develop them. Vallor's (2018) influential work on "technomoral virtues" identifies qualities such as honesty, justice, courage, and care as essential for responsible technology development. This approach has gained traction as recognition grows that responsible AI requires not just technical solutions but organizational cultures that prioritize ethical considerations.

A significant tension exists between universalist approaches that seek broadly applicable principles and particularist approaches that emphasize context-specificity. Wong (2020) argues that ethical considerations in AI cannot be reduced to abstract principles but must engage with concrete contexts, power relations, and lived experiences. This tension remains unresolved in current theoretical frameworks, with implications for how responsible AI translates across diverse domains and cultures.

3. Global Variations in Responsible AI Frameworks

Responsible AI frameworks demonstrate significant variation across geographic regions, reflecting different cultural values, governance traditions, and technological priorities. European frameworks, exemplified by the EU's Ethics Guidelines for Trustworthy AI (High-Level Expert Group on AI, 2019), emphasize human rights, precautionary approaches, and comprehensive

regulation. This rights-based orientation contrasts with US frameworks that typically place greater emphasis on innovation, market-based solutions, and voluntary standards (Crawford & Calo, 2016).

East Asian approaches to responsible AI reveal further diversity. Japan's Society 5.0 framework emphasizes human-machine harmony and societal benefit, while China's governance documents prioritize economic development alongside security considerations (Roberts et al., 2021). These variations reflect not only different values but also strategic positioning in global AI development.

The Global South remains significantly underrepresented in dominant responsible AI frameworks, despite growing AI development and deployment in these regions. Birhane et al. (2022) highlight how responsible AI frameworks often embed Western assumptions about privacy, agency, and social organization that may not translate across cultural contexts. Recent efforts by organizations like UNESCO (2021) aim to develop more inclusive global frameworks, but significant work remains to incorporate diverse perspectives.

4. Integration and Implementation Challenges

Translating theoretical frameworks into operational practices presents significant challenges. Organizations often struggle to operationalize abstract principles into concrete policies, technical specifications, and governance structures. Greene et al. (2019) document how ethical principles frequently remain disconnected from actual development practices, creating an "ethics-washing" risk where organizations adopt principled language without meaningful implementation.

Recent work has focused on bridging this gap through more actionable frameworks. Raji et al. (2020) developed documentation approaches that embed ethical considerations throughout the AI lifecycle, while Metcalf et al. (2021) propose integrating ethics into existing risk management and quality assurance processes. Despite these advances, substantial implementation challenges persist, particularly for smaller organizations with limited resources for dedicated ethics teams or processes.

As responsible AI continues to mature as a field, theoretical frameworks must evolve to address emerging challenges posed by increasingly capable AI systems. Foundational questions about agency, consciousness, and human-AI boundaries, once considered speculative, now require serious theoretical engagement. The rapid advancement of large language models has particularly highlighted theoretical gaps regarding systems that appear to reason, create, and engage with normative questions (Gabriel, 2022). Future theoretical frameworks must grapple with these fundamental questions while remaining practical enough to guide concrete decisions by developers, deployers, and regulators.

C. Technical Advances in Fairness and Bias

The pursuit of fairness in AI systems has evolved from a niche research area to a central concern in machine learning. As algorithmic systems increasingly influence consequential decisions affecting human lives, technical approaches to ensuring fairness and mitigating bias have grown in sophistication and scope. This

section examines recent advances in fairness-aware machine learning, highlighting key approaches, persistent challenges, and emerging directions.

1. Pre-processing, In-processing, and Post-processing Techniques

Technical approaches to fairness in machine learning are commonly categorized by their position in the development pipeline. Pre-processing techniques modify training data to mitigate embedded biases before model training begins. Feldman et al. (2022) demonstrated that carefully designed data transformations can effectively reduce disparate impact while preserving overall accuracy in decision systems. Similarly, Lum and Johndrow (2019) proposed statistical methodologies for transforming features to achieve independence from protected attributes, enabling fairness through unawareness while addressing proxy discrimination. Recent advances by Martinez et al. (2023) have extended these approaches to unstructured data including images and text, addressing representational harms in foundation models.

In-processing techniques incorporate fairness constraints directly into the learning algorithm. Agarwal et al. (2018) introduced influential work on constrained optimization approaches that balance accuracy objectives with fairness constraints. Building on this foundation, Zhang et al. (2021) developed adversarial debiasing techniques that actively work to unlearn correlations between predictions and sensitive attributes. Recent work by Roth et al. (2024) has demonstrated the effectiveness of in-processing methods for complex deep learning architectures, addressing previous limitations in scaling fairness constraints to large neural networks.

Post-processing techniques adjust model outputs after training to satisfy fairness criteria. As demonstrated by Hardt et al. (2016), these approaches can be particularly valuable when modifying existing systems where retraining is impractical.

Corbett-Davies et al. (2017) explored threshold adjustments in classification tasks to achieve equalized odds or demographic parity. More recently, Park et al. (2023) developed calibration-based approaches that maintain fairness guarantees while adapting to distribution shifts in deployment environments.

2. Group Fairness vs. Individual Fairness Approaches

The fairness literature reveals a fundamental tension between group-based and individual-based conceptions of fairness. Group fairness metrics focus on statistical parity across demographic categories, ensuring similar outcomes or error rates between protected groups. Demographic parity, equalized odds, and equal opportunity have emerged as dominant group fairness metrics (Mehrabi et al., 2021). However, as demonstrated by Kleinberg et al. (2016), fundamental incompatibilities exist between different group fairness metrics, requiring explicit value judgments about which disparities are most concerning in specific contexts.

Individual fairness approaches, conversely, focus on ensuring similar individuals receive similar outcomes regardless of group membership. Dwork et al. (2012) pioneered this approach with their formulation of individual fairness through the lens of Lipschitz continuity. Recent advances by Mukherjee et al. (2023) have addressed previous limitations in defining similarity metrics by

leveraging techniques from representation learning. Jung et al. (2022) demonstrated that individual fairness approaches can circumvent impossibility results that plague group fairness metrics, while Wang et al. (2024) established frameworks for auditing individual fairness in deployed systems.

The integration of group and individual fairness approaches represents a promising research direction. Fleischer et al. (2022) developed hybrid fairness metrics that balance concerns about group-level disparities with individual-level consistency. Similarly, Chakraborty et al. (2023) proposed frameworks for explicitly reasoning about trade-offs between different fairness conceptualizations, allowing decision-makers to express values through constrained optimization approaches.

3. Recent Advances in Fairness-Aware Deep Learning

The rise of deep learning has presented both challenges and opportunities for fairness-aware machine learning. Complex neural architectures often function as black boxes, complicating fairness analysis, yet their flexibility enables novel approaches to bias mitigation. Wang et al. (2020) pioneered techniques for disentangled representation learning that separate protected characteristics from other features while preserving predictive power. Building on this work, Locatello et al. (2023) established theoretical foundations for fair representation learning with formal guarantees about independence from protected attributes.

Attention mechanisms have emerged as particularly valuable for fairness-aware deep learning. Chen et al. (2022) demonstrated that attention-based models can be designed to explicitly down-weight features that correlate with protected attributes while emphasizing fairness-promoting features. This approach has proven especially effective in natural language processing, where biased associations in word embeddings have long presented challenges. Zhao and Brantley (2021) developed techniques for identifying and mitigating harmful stereotypes and associations in large language models through targeted intervention in attention patterns.

Transfer learning approaches have gained prominence as efficient methods for adapting pre-trained models for fairness considerations. Li et al. (2022) showed that adapter modules can effectively debias foundation models without requiring full retraining, significantly reducing computational costs of fairness interventions. This approach addresses growing concerns about the environmental and economic impacts of training AI systems, as documented by Henderson et al. (2024), who demonstrated the substantial carbon footprint of retraining large models for fairness modifications.

4. Challenges in Complex Fairness Scenarios

While significant progress has been made in technical fairness approaches, substantial challenges persist in complex real-world scenarios. Intersectional fairness—addressing how multiple dimensions of identity interact—remains particularly challenging. Crenshaw's (1989) foundational work on intersectionality has inspired technical approaches by Foulds et al. (2020), who developed methods for modeling complex interaction effects between protected attributes. Building on this foundation, Yang et al. (2023) proposed hierarchical fairness metrics that

account for subgroup heterogeneity, while Ghosh et al. (2024) demonstrated techniques for ensuring fairness across exponentially many subgroups without requiring exhaustive enumeration.

Causal approaches to fairness have emerged as a promising direction for addressing limitations of purely statistical methods. Pearl (2019) established fundamental connections between causal reasoning and algorithmic fairness, while Zhang and Bareinboim (2018) developed practical frameworks for implementing counterfactual fairness in machine learning pipelines. Recent work by Liu et al. (2023) extended these approaches to address path-specific effects in causal graphs, enabling more nuanced interventions in complex sociotechnical systems where multiple causal pathways exist between protected attributes and outcomes.

Dynamic feedback effects present another frontier challenge in fairness research. Liu et al. (2018) demonstrated how seemingly fair algorithms can create or amplify unfairness over time through feedback loops. Addressing this challenge, Ensign et al. (2020) developed frameworks for modeling runaway feedback effects, while Kallus and Zhou (2022) proposed robust optimization approaches that explicitly account for distribution shifts caused by algorithmic deployment. These dynamic considerations have become increasingly important as AI systems operate as components in complex sociotechnical systems rather than isolated decision points.

The evaluation of fairness interventions in real-world contexts remains challenging despite methodological advances. Holstein et al. (2019) documented significant gaps between fairness research and practitioner needs, highlighting the importance of domain-specific evaluation. Addressing this gap, Sambasivan et al. (2021) proposed frameworks for contextual fairness assessment that incorporate stakeholder perspectives and domain knowledge. Most recently, Parker et al. (2024) established methodologies for assessing downstream impacts of fairness interventions, moving beyond immediate statistical metrics to evaluate broader societal effects of fair machine learning approaches.

D. Transparency and Explainability

As AI systems become increasingly complex and autonomous, the need for transparency and explainability has emerged as a critical component of responsible AI. Explainable AI (XAI) seeks to make AI systems understandable to humans, enabling meaningful oversight, informed trust, and effective human-AI collaboration. This section examines recent advances in explainability approaches, highlighting methodological innovations, domain-specific applications, and persisting challenges.

1. Local vs. Global Explainability Methods

Explainability methods in AI are commonly categorized as either local (focused on individual predictions) or global (aimed at understanding the model as a whole). Local explainability techniques provide insights into specific decisions or predictions. LIME (Local Interpretable Model-agnostic Explanations), developed by Ribeiro et al. (2016), approximates complex models locally with interpretable surrogates to explain individual predictions. Building on this foundation, SHAP

(SHapley Additive exPlanations) by Lundberg and Lee (2017) unified various local explanation methods under a coherent theoretical framework based on cooperative game theory. Recent advances by Zhang et al. (2023) have addressed previous computational limitations of these approaches, enabling real-time explanations even for complex neural architectures.

While local methods provide granular insights, they often fail to capture broader patterns in model behavior. Addressing this limitation, global explainability methods seek to characterize overall model functioning. Friedman (2001) pioneered partial dependence plots that visualize relationships between features and predictions across a model's entire input space. Expanding on this work, Molnar et al. (2020) developed accumulated local effects plots that address correlation issues in partial dependence while maintaining interpretability. Recent innovations by Lakkaraju et al. (2022) have introduced global explanation frameworks that identify and characterize distinct decision regions within complex models, bridging the gap between local and global explanations.

The integration of local and hierarchical explanations represents a promising research direction. Lage et al. (2023) demonstrated techniques for aggregating local explanations into meaningful global insights, while preserving the ability to drill down into specific cases. Similarly, Yang et al. (2024) proposed frameworks for multi-level explanations that allow users to seamlessly transition between overview and detailed perspectives, addressing previous limitations in explanation scalability for complex models.

2. Progress in Interpretable Neural Architectures

Rather than explaining black-box models post hoc, some researchers have focused on developing inherently interpretable neural architectures. Attention mechanisms have emerged as particularly valuable for interpretability. Vaswani et al. (2017) introduced the transformer architecture with self-attention mechanisms that not only improved performance but also provided visibility into feature relationships. Building on this foundation, Vig (2019) demonstrated techniques for visualizing attention patterns to reveal linguistic structures learned by language models. Recent work by Chefer et al. (2021) extended these approaches with transformer relevancy propagation, enabling more accurate attribution of predictions to input features.

Neural networks with explicit symbolic components offer another approach to interpretable architectures. Koh et al. (2020) developed concept bottleneck models that force neural networks to make predictions through human-interpretable concepts. Similarly, Chen et al. (2020) introduced prototype networks that base classifications on similarity to learned prototypical examples, enabling intuitive explanations. Most recently, Wang et al. (2024) demonstrated neuro-symbolic architectures that combine the expressiveness of neural networks with the transparency of symbolic reasoning, addressing previous performance limitations in interpretable models.

The trade-off between performance and interpretability has been a persistent concern in responsible AI. However, recent advances suggest this may be a false dichotomy in many cases. Chang et al. (2021) conducted comprehensive benchmarks demonstrating that carefully designed interpretable models can

match or exceed the performance of black-box counterparts across multiple domains. Building on these findings, Martinez et al. (2023) established design principles for high-performing interpretable architectures, while Rudin and Radin (2019) presented evidence that inherently interpretable models often generalize better to unseen data than complex black-box alternatives.

3. Explainability for Foundation Models and LLMs

The emergence of foundation models and Large Language Models (LLMs) has introduced unprecedented challenges for explainability. These models contain billions of parameters, are trained on vast datasets, and exhibit emergent capabilities that were not explicitly programmed. Traditional explanation methods often prove inadequate for these systems. Addressing this challenge, Geva et al. (2022) developed transformer circuits analysis, revealing computational substructures within large language models that correspond to interpretable functions. Similarly, Elhage et al. (2021) introduced mechanistic interpretability approaches that identify specific circuits responsible for capabilities like induction and association.

Feature visualization techniques have proven valuable for understanding internal representations in foundation models. Olah et al. (2020) pioneered methods for visualizing what neurons in deep networks detect, revealing high-level concepts encoded within hidden layers. Building on this work, Räuker et al. (2023) developed techniques for mapping semantic concepts across model layers, revealing how abstractions are constructed hierarchically. Most recently, Hoyt et al. (2024) demonstrated approaches for visualizing and manipulating latent spaces in multimodal foundation models, enabling more transparent understanding of how these systems connect different modalities.

Chain-of-thought approaches have emerged as powerful tools for explaining reasoning in language models. Wei et al. (2022) demonstrated that prompting LLMs to generate step-by-step reasoning significantly improves both performance and explainability. Expanding on this work, Kojima et al. (2023) showed that zero-shot chain-of-thought prompting can elicit explicit reasoning paths without task-specific examples. Recent innovations by Li and Qiu (2024) have integrated these approaches with formal verification techniques, enabling automated checking of reasoning validity in explanations generated by language models.

The evaluation of explanations for foundation models presents unique challenges due to their scale and complexity. Doshi-Velez and Kim (2017) established a theoretical framework for evaluating explanation quality, distinguishing between application-grounded, human-grounded, and functionally-grounded evaluation approaches. Building on this taxonomy, Zhang et al. (2022) developed benchmarks specifically designed for evaluating explanations of large language models, while Davis et al. (2024) proposed frameworks for assessing explanation fidelity through counterfactual testing. These innovations address previous limitations in ensuring that explanations accurately represent model behavior rather than providing plausible but misleading accounts of decision processes.

4. Sociotechnical Perspectives on Explainability

While technical approaches to explainability have advanced significantly, researchers increasingly recognize that explainability must be understood as a sociotechnical challenge rather than a purely technical one. Miller (2019) demonstrated that effective explanations must align with human cognitive patterns and social expectations, not just technical accuracy. Building on this insight, Kaur et al. (2022) conducted empirical studies showing how explanation interfaces influence user trust and understanding, highlighting the importance of user-centered design in explainability systems.

The purposes and contexts of explanations profoundly shape requirements for explainable AI. As Mittelstadt et al. (2019) argue, explanations for regulatory compliance differ substantially from those aimed at helping users understand and effectively collaborate with AI systems. Expanding on this framework, Hong et al. (2020) developed context-sensitive explanation approaches that adapt to user expertise, task requirements, and time constraints. Recent work by Ehsan et al. (2023) introduced socially-situated explainability frameworks that explicitly account for power relationships, institutional contexts, and stakeholder diversity.

Explanations must be evaluated not just for technical accuracy but for their effectiveness in supporting human decision-making. Bansal et al. (2021) demonstrated that explanations can paradoxically decrease human-AI team performance if they increase cognitive load without improving understanding. Addressing this challenge, Wang et al. (2021) developed adaptive explanation systems that provide different levels of detail based on detected user needs. Most recently, Karpus et al. (2024) established frameworks for measuring explanation utility across different stakeholder groups, enabling more nuanced evaluation of explanation effectiveness beyond generic transparency metrics.

Cultural and linguistic factors significantly influence explanation effectiveness but have received insufficient attention in technical XAI research. Ehsan and Riedl (2020) highlighted how explanation preferences vary across cultural contexts, with different expectations for detail, certainty, and framing. Building on these insights, Chen et al. (2023) developed culturally-adaptive explanation frameworks that account for varying communication norms and epistemic traditions. These approaches address growing concerns about the Western-centric nature of existing XAI methods, as documented by Sambasivan et al. (2022), who demonstrated significant gaps between dominant XAI approaches and explanation needs in Global South contexts.

E. Safety and Alignment

The challenge of ensuring AI systems behave safely and in accordance with human values has grown increasingly urgent as AI capabilities advance. Safety and alignment research aims to develop systems that reliably pursue intended goals, avoid harmful behaviors, and remain under meaningful human control even as capabilities increase. This section examines recent advances in AI safety research, spanning theoretical foundations, practical techniques, and governance approaches.

1. Advances in AI Alignment Techniques

AI alignment—ensuring systems pursue goals aligned with human values and intentions—has evolved from a speculative concern to an active research field. The alignment problem, first formalized by Bostrom (2014), encompasses challenges of specification (correctly defining what we want), robustness (ensuring systems pursue these goals across varied circumstances), and assurance (verifying alignment has been achieved). Addressing the specification challenge, Hadfield-Menell et al. (2016) introduced cooperative inverse reinforcement learning, enabling systems to learn human preferences through interaction rather than explicit programming. Building on this foundation, Christiano et al. (2017) developed preference learning from human feedback, allowing non-technical users to train systems by expressing preferences between outputs.

Recent advances have focused on scalable oversight for increasingly capable systems. Irving et al. (2018) pioneered debate as an alignment mechanism, where AI systems argue for different answers while humans judge the exchange. Expanding this approach, Saunders et al. (2022) developed recursive reward modeling, where AI systems trained on easier oversight tasks help provide oversight for more difficult tasks. The integration of these approaches with large language models has been particularly influential. Anthropic's constitutional AI approach, developed by Bai et al. (2022), uses AI systems to critique their own outputs against predefined principles, enabling alignment with complex values that resist simple specification.

Formal verification methods have made significant strides in providing guarantees about system behavior. Fisher et al. (2019) demonstrated techniques for verifying reinforcement learning policies against temporal logic specifications, providing mathematical guarantees about system behavior within defined parameters. Similarly, Cohen et al. (2023) developed certification techniques for neural networks, ensuring robustness against adversarial perturbations with formal guarantees. Most recently, Barrett et al. (2024) established frameworks for compositional verification of large-scale systems, addressing previous limitations in scaling formal methods to complex AI architectures.

2. Robustness Against Adversarial Attacks

Ensuring AI systems maintain safe behavior in adversarial conditions remains a critical challenge for responsible deployment. Adversarial examples—inputs specifically designed to cause misclassification—were first demonstrated by Szegedy et al. (2014), revealing fundamental vulnerabilities in neural networks. Since then, defensive techniques have evolved substantially. Madry et al. (2018) introduced adversarial training, incorporating adversarial examples during model training to improve robustness. Building on this approach, Wong and Kolter (2018) developed provable defenses using convex relaxations, providing formal guarantees about robustness regions. Recent work by Rebuffi et al. (2021) has demonstrated significant improvements in adversarial robustness while maintaining accuracy on clean examples, addressing previous trade-offs between performance and security.

Beyond classification tasks, adversarial vulnerabilities in generative models present unique challenges. Carlini et al. (2021) demonstrated that language models

can be manipulated to produce harmful content despite safety filters through carefully crafted prompts. Addressing this threat, Ganguli et al. (2022) developed red-teaming approaches where specialized models attempt to find vulnerabilities in target systems, enabling systematic discovery and patching of safety weaknesses. This approach has been formalized by Casper et al. (2023) into comprehensive adversarial testing frameworks that provide measurable assurance about system robustness across diverse threat scenarios.

The connection between adversarial robustness and other safety properties has emerged as an important research direction. Koh et al. (2022) established theoretical links between robustness against input perturbations and stability under distribution shifts, demonstrating that certain adversarial defenses improve generalization to new environments. Similarly, Hendrycks et al. (2021) showed that adversarial robustness often correlates with improved out-of-distribution detection, enabling systems to recognize when they are operating outside their training distribution. These findings suggest that robustness improvements may yield broader safety benefits beyond specific adversarial threats.

3. Constitutional AI and Value Learning Approaches

Recent years have seen significant advances in approaches for instilling AI systems with complex human values. Constitutional AI, pioneered by Anthropic, represents a significant innovation in this space. As described by Bai et al. (2022), constitutional approaches encode ethical principles as guidelines that AI systems follow when generating content. Rather than relying solely on human feedback, which scales poorly and can be manipulated, constitutional approaches delegate some oversight to other AI systems guided by explicit principles. Building on this foundation, Leike et al. (2022) developed recursive evaluation frameworks where AI systems provide feedback on their own outputs, creating scalable oversight mechanisms for increasingly capable systems.

Value learning approaches aim to discover human values from observed behavior or stated preferences. Jeon et al. (2020) demonstrated that inverse reinforcement learning can infer complex reward functions from demonstrations, enabling systems to learn nuanced human preferences without explicit specification. Expanding this work, Pan et al. (2022) developed risk-averse value learning approaches that handle uncertainty in inferred values conservatively, reducing the potential for harmful optimization of misspecified objectives. Most recently, Jiang et al. (2024) established frameworks for value learning from natural language feedback, enabling non-technical users to shape system behavior through ordinary conversation.

Pluralistic approaches to value learning have gained prominence as researchers recognize the diversity of human values across individuals and cultures. Conitzer et al. (2021) developed game-theoretic approaches to value aggregation that balance different stakeholder preferences without imposing arbitrary resolutions to fundamental value differences. Similarly, Zhao et al. (2023) proposed frameworks for explicitly representing value uncertainty in AI systems, enabling more transparent navigation of normative disagreements. These approaches address growing recognition that single-objective optimization often fails to capture the complexity of human values and priorities.

4. Empirical Safety Research and Benchmarks

Empirical approaches to AI safety have made significant progress in identifying and measuring concrete risks in deployed systems. The discovery of emergent capabilities in large language models, documented by Wei et al. (2022), highlighted how systems can develop unforeseen abilities as scale increases. This finding underscored the importance of systematic testing before deployment. Addressing this need, Hendrycks et al. (2021) developed comprehensive benchmarks for measuring harmful capabilities in language models across dimensions including toxicity, bias, and information hazards. These benchmarks have enabled more rigorous comparison between safety approaches and tracking of progress over time.

Trojan attacks—where systems are deliberately trained to exhibit harmful behaviors in response to specific triggers—represent a significant security concern for AI systems. Wang et al. (2019) demonstrated that neural networks can be compromised through backdoor attacks during training, creating vulnerabilities that activate only under specific circumstances. Building on this work, Liu et al. (2023) developed detection methods for identifying trojaned models before deployment, while Chen et al. (2023) established frameworks for guaranteeing trojan-free training processes through cryptographic techniques. These advances address growing concerns about supply chain attacks in AI development, where malicious actors might compromise widely used models.

The evaluation of safety measures for increasingly capable AI systems presents unique methodological challenges. Testing systems at the frontier of capabilities requires specialized approaches to risk mitigation. Addressing this challenge, Amodei et al. (2020) proposed AI safety via debate, where systems argue for competing assessments of their own behavior, enabling human evaluators to identify risks despite limited technical expertise. Building on this concept, Irving and Askell (2019) developed techniques for scalable oversight through recursive decomposition, breaking complex evaluation tasks into manageable subtasks. Most recently, Park et al. (2024) established frameworks for red-teaming frontier models that balance thorough safety assessment with responsible handling of discovered vulnerabilities.

5. Long-term Safety and Governance Approaches

As AI capabilities continue to advance, research on long-term safety and governance has grown increasingly important. The challenge of aligning superintelligent systems with human values, first formalized by Bostrom (2014), has inspired diverse technical approaches. Christiano et al. (2018) proposed iterated amplification and distillation as a framework for maintaining alignment during capability scaling, using systems to help oversee more advanced versions of themselves. Building on this work, Leike et al. (2024) developed techniques for preserving alignment guarantees during transfer learning, addressing risks that arise when systems rapidly acquire new capabilities through adaptation rather than training from scratch.

Formal approaches to AI governance have emerged as a crucial complement to technical safety research. Dafoe (2018) established a conceptual framework for AI governance, identifying key decision points and stakeholders in managing

advanced AI development. Expanding on this foundation, Zhang et al. (2022) proposed frameworks for integrating technical safeguards with institutional governance mechanisms, creating multiple layers of protection against misuse or accident. Most recently, Critch and Krueger (2023) developed game-theoretic models of AI development races, identifying governance interventions that could maintain safety without sacrificing technological progress.

International coordination on AI safety presents distinct challenges that blend technical and diplomatic considerations. Maas (2019) analyzed historical cases of arms control and other international regimes to identify lessons for AI governance. Building on this historical perspective, Anderljung et al. (2022) proposed concrete mechanisms for international verification of AI safety measures, addressing challenges of monitoring compliance without requiring disclosure of sensitive intellectual property. These approaches reflect growing recognition that ensuring safe AI development requires not just technical solutions but effective coordination between organizations, states, and other stakeholders in a complex global landscape.

F. Measuring Impact and Effectiveness

The advancement of responsible AI requires not just developing methods and frameworks but systematically evaluating their effectiveness. As responsible AI initiatives proliferate across sectors, stakeholders increasingly demand evidence of impact rather than merely procedural compliance. This section examines approaches to measuring the effectiveness of responsible AI interventions, including evaluation frameworks, benchmarking methodologies, and longitudinal impact studies.

1. Evaluation Frameworks for Responsible AI Solutions

Comprehensive evaluation frameworks provide structured approaches for assessing responsible AI solutions across multiple dimensions. These frameworks move beyond narrow technical metrics to incorporate broader considerations of social impact, stakeholder perspectives, and contextual appropriateness. The Responsible AI Framework developed by Floridi et al. (2020) establishes five core principles—beneficence, non-maleficence, autonomy, justice, and explicability—with corresponding evaluation criteria for each dimension. Building on this foundation, Raji et al. (2020) proposed the SMACTR (Scoping, Mapping, Artifact Collection, Testing, and Reflection) framework for conducting algorithmic audits, providing a structured methodology for comprehensive system evaluation.

Evaluation frameworks increasingly recognize the importance of incorporating diverse stakeholder perspectives rather than relying solely on expert assessment. Metcalf et al. (2021) developed participatory evaluation methodologies that engage affected communities throughout the assessment process, addressing power imbalances in traditional evaluation approaches. Similarly, Sloane et al. (2022) proposed frameworks for assessing algorithmic systems through multiple value lenses, acknowledging legitimate differences in how diverse stakeholders might define and prioritize responsible AI objectives. These multi-perspective approaches address limitations of expert-driven

evaluation that may miss impacts visible only to those with lived experience of the evaluated systems.

The evaluation of responsible AI increasingly emphasizes counterfactual comparison rather than absolute assessment. Rather than asking whether a system meets abstract ethical criteria, these approaches compare outcomes against realistic alternatives. Mittelstadt (2019) pioneered the principle of comparative evaluation, arguing that responsible AI systems should be assessed against both human alternatives and feasible algorithmic alternatives rather than idealized standards. Building on this principle, Kallus and Zhou (2022) developed methodologies for counterfactual evaluation of fairness interventions, measuring not just outcome disparities but how interventions change outcomes compared to baseline scenarios. These comparative approaches provide more actionable insights by focusing on marginal improvements rather than binary judgments of "ethical" or "unethical."

Domain-specific evaluation frameworks have emerged to address the limitations of generic assessment methods. As Whittlestone et al. (2021) argue, effective evaluation must account for context-specific ethical considerations, acceptable trade-offs, and domain-appropriate metrics. Responding to this need, Sendak et al. (2023) developed clinical AI evaluation frameworks that incorporate both technical validation and health system integration assessment. Similarly, Richardson et al. (2022) established methodologies for evaluating public sector algorithms that address democratic values and administrative law principles alongside technical performance. These domain-specific approaches recognize that responsible AI evaluation criteria must be tailored to the specific contexts in which systems operate.

2. Metrics and Benchmarks for Responsible AI Systems

The development of standardized metrics and benchmarks enables more rigorous and consistent evaluation of responsible AI systems across implementations and contexts. While early responsible AI efforts often relied on ad hoc evaluation, recent years have seen significant progress in metric standardization. Mitchell et al. (2019) introduced Model Cards for model documentation and performance reporting across different subgroups and conditions, establishing a standard reporting framework for disaggregated evaluation. Building on this foundation, Barocas et al. (2021) developed the fairness indicators framework, providing standardized metrics for assessing prediction disparities across intersectional demographic groups.

Benchmark datasets play a crucial role in responsible AI evaluation by enabling consistent comparison across different approaches. However, as documented by Blodgett et al. (2021), many existing benchmarks contain embedded biases or fail to represent important edge cases, limiting their utility for comprehensive evaluation. Addressing these limitations, Derczynski et al. (2023) developed responsible benchmark creation methodologies that incorporate fairness and representation considerations from dataset conception through implementation. Similarly, Liang et al. (2022) introduced the HELM (Holistic Evaluation of Language Models) benchmark suite specifically designed to assess

language models across dimensions including fairness, toxicity, and reasoning capabilities.

The measurement of algorithmic bias has evolved from simple demographic parity metrics to more sophisticated approaches that capture nuanced fairness considerations. These advances address limitations of early metrics that could be satisfied through mathematically "fair" but practically problematic solutions. Corbett-Davies and Goel (2018) established a framework for understanding inherent trade-offs between different fairness metrics, demonstrating that common measures often conflict with each other. Building on this analysis, Verma and Rubin (2022) cataloged over twenty distinct fairness definitions and their relationships, providing guidance for selecting appropriate metrics based on specific fairness objectives. Most recently, Wan et al. (2024) developed dynamic fairness metrics that account for feedback effects and changing social conditions rather than treating fairness as a static property.

Privacy and security metrics have similarly evolved toward more comprehensive and context-sensitive approaches. Traditional privacy metrics focused largely on data protection through anonymization techniques, which often prove inadequate for high-dimensional data used in modern AI systems. Addressing these limitations, Wagner and Eckhoff (2018) developed a taxonomy of privacy metrics encompassing protection against inference, identifiability, and information leakage. Building on this foundation, Jayaraman and Evans (2023) established evaluation frameworks for privacy-preserving machine learning that assess both formal privacy guarantees and practical protection against realistic attacks. These advances enable more rigorous evaluation of privacy protections in responsible AI implementations.

3. Longitudinal Studies on Implemented Systems

While much responsible AI research focuses on system design and pre-deployment evaluation, longitudinal studies examining deployed systems provide critical insights into real-world effectiveness and unintended consequences. These studies move beyond theoretical analysis to examine how systems function in dynamic sociotechnical environments. Green and Chen (2022) conducted influential longitudinal studies of risk assessment algorithms in criminal justice settings, revealing how predicted impacts often diverge from actual outcomes due to implementation factors and system interactions. Building on this work, Chouldechova et al. (2022) developed methodologies for monitoring deployed AI systems over time, enabling detection of performance degradation or emerging biases as environments change.

Longitudinal research has been particularly valuable for understanding how human-AI interactions evolve in practice. Initial user studies often fail to capture how interactions change as users develop mental models and adaptive behaviors around AI systems. Addressing this limitation, Yang et al. (2020) conducted extended field studies of clinical decision support systems, documenting how clinician trust and usage patterns evolved over months of deployment. Similarly, Passi and Barocas (2022) examined how data scientists' interactions with fairness tools changed over time, revealing how initial compliance often gave way to creative workarounds when tools conflicted with organizational priorities. These

studies highlight the importance of examining responsible AI effectiveness beyond immediate post-deployment periods.

The study of responsible AI in organizational contexts reveals how institutional factors shape implementation effectiveness. Algorithmic systems operate within complex social environments where formal policies interact with informal practices and institutional incentives. Examining these dynamics, Rakova et al. (2021) conducted longitudinal ethnographic studies of responsible AI implementation in corporate settings, documenting how organizational structures and incentives often undermined stated ethical commitments. Building on this work, Moss et al. (2023) developed frameworks for assessing organizational integration of responsible AI practices beyond documentation and process requirements. Most recently, Madaio et al. (2024) established methodologies for evaluating responsible AI culture within organizations, measuring not just formal compliance but internalization of responsible practices throughout development teams.

Public sector deployments of responsible AI systems have been the subject of particularly valuable longitudinal research. These contexts often involve greater transparency requirements and stakeholder oversight than private implementations, enabling more comprehensive study. Examining these deployments, Veale et al. (2020) documented how public sector algorithmic systems evolved in response to legal challenges, media scrutiny, and changing political priorities. Similarly, Brown et al. (2023) conducted multi-year studies of automated decision systems in administrative agencies, revealing how implementation factors often determined whether responsible design translated into responsible outcomes. These studies highlight how technical interventions interact with institutional contexts to produce observed impacts, emphasizing the sociotechnical nature of responsible AI in practice.

4. Challenges in Impact Measurement

Despite significant advances in measurement methodologies, fundamental challenges persist in evaluating responsible AI impact. Counterfactual outcomes—what would have happened without the AI system or with an alternative implementation—remain inherently unobservable, complicating causal impact assessment. Addressing this challenge, Pearl and Mackenzie (2018) established causal inference frameworks for algorithmic impact assessment, enabling more rigorous attribution of observed outcomes to specific interventions. Building on this foundation, D'Amour et al. (2022) developed undertreatment bias audit methodologies that examine not just what systems do but what opportunities they may systematically overlook. These approaches enable more robust assessment of both direct impacts and opportunity costs associated with AI deployments.

Distribution shifts over time present another significant challenge for responsible AI evaluation. Systems trained on historical data may experience performance degradation as real-world distributions change, particularly for disadvantaged groups underrepresented in training data. Examining this phenomenon, Geirhos et al. (2020) documented how seemingly robust models often fail under distribution shifts that humans handle without difficulty. Addressing this challenge, Subbaswamy et al. (2021) developed methodologies for

evaluating algorithmic robustness under distribution shift, enabling more reliable assessment of long-term performance. Most recently, Martinez et al. (2024) established frameworks for continuous fairness monitoring that detect emerging disparities as populations and environments evolve.

The multi-stakeholder nature of AI systems complicates impact assessment by introducing multiple, sometimes conflicting, evaluation criteria. Different stakeholders may prioritize different metrics or interpret the same results differently based on their positions and interests. Addressing this challenge, Zhu et al. (2023) developed multi-objective evaluation frameworks that explicitly model trade-offs between competing objectives rather than collapsing evaluation to a single metric. Similarly, Lee and Singh (2021) established methodologies for stakeholder-specific impact assessment that maintain distinct evaluation perspectives rather than forcing artificial consensus. These approaches enable more nuanced evaluation that acknowledges legitimate differences in how various stakeholders might define and measure responsible AI success.

Long feedback loops between implementation and observable outcomes present particular challenges for impact measurement in domains where effects may take years to materialize. Educational algorithms, for instance, may affect long-term educational trajectories that cannot be fully assessed immediately after deployment. Addressing this challenge, Kizilcec and Reich (2023) developed methodologies for early indicator identification that correlate with long-term outcomes of interest. Building on this work, Holstein et al. (2023) established evaluation frameworks specifically designed for systems with extended impact horizons, combining short-term process metrics with strategic longitudinal data collection. These approaches enable more timely assessment while acknowledging the inherent limitations of short-term evaluation for systems with long-term impacts.

G. Conclusion

The advancement of responsible AI represents one of the most critical challenges facing technology development in the 21st century. This paper has examined the multifaceted landscape of responsible AI, exploring its theoretical foundations, technical implementations, and practical implications. As artificial intelligence systems continue to transform industries, societies, and human experiences, the imperative for ensuring these technologies operate ethically, safely, and in alignment with human values has never been more urgent (Crawford, 2021; Mittelstadt, 2019; Whittlestone et al., 2019).

Our analysis reveals that responsible AI has evolved significantly from abstract principles to more comprehensive frameworks that recognize the nuanced interplay between competing values in real-world contexts. While substantial progress has been made in developing technical methods addressing fairness, transparency, and safety, fundamental tensions persist between objectives such as performance, interpretability, and privacy—tensions that require explicit value judgments rather than purely technical solutions (Selbst et al., 2019; Floridi & Cowls, 2019; Hagendorff, 2020). This highlights the inherently sociotechnical nature of AI systems, where technical approaches

alone cannot address the complex social, political, and ethical dimensions of AI impacts (Green, 2022; Birhane, 2021; Benjamin, 2019).

The implementation gap between formal responsible AI processes and substantive changes in development practices represents one of the most significant barriers to advancing responsible AI. Organizations have increasingly adopted documentation requirements, review committees, and impact assessments, yet these formal mechanisms often fail to influence development practices meaningfully without corresponding changes to organizational incentives, resources, and culture (Raji et al., 2022; Madaio et al., 2020; Rakova et al., 2021). This underscores the importance of structural changes throughout the AI lifecycle rather than treating ethics as a compliance checkbox (Wong, 2020; Jobin et al., 2019; Metcalf et al., 2021).

Current responsible AI approaches face important limitations, including an overemphasis on individual-focused fairness frameworks that fail to address broader structural inequities, the reactive nature of many ethical interventions that come after core design decisions have been made, and the undertheorization of power dynamics in AI development (D'Ignazio & Klein, 2020; Mohamed et al., 2020; Sloane et al., 2022).

Additionally, narrow technical solutions often overlook how system impacts emerge from interactions between technical features and institutional processes within complex sociotechnical systems (Pasquale, 2020; Costanza-Chock, 2020; Noble, 2018).

Moving forward, advancing responsible AI practice will require concerted effort across multiple stakeholder groups. Researchers should prioritize interdisciplinary collaboration bridging technical methods with social, ethical, and legal perspectives (Selbst et al., 2019; Buolamwini & Gebru, 2018; Moss & Metcalf, 2020). Industry organizations must move beyond abstract principles toward operational implementation of responsible practices throughout development processes, with sufficient resources and meaningful authority for ethics teams (Raji et al., 2020; Greene et al., 2019; Morley et al., 2021). Policymakers should develop adaptive governance frameworks that establish meaningful safeguards while enabling continued innovation, with particular attention to high-risk applications (Nemitz, 2018; Cath et al., 2018; Yeung et al., 2020).

The path toward truly responsible AI is neither simple nor straightforward, requiring ongoing commitment, collaboration, and critical reflection from diverse stakeholders. By addressing both technical and social dimensions of AI systems, integrating responsible practices throughout the development lifecycle, and establishing effective governance mechanisms, we can work toward AI systems that not only advance technological capabilities but also support human flourishing, social justice, and sustainable progress (West et al., 2019; Dobbe et al., 2022; Prunkl & Whittlestone, 2020). The future of AI will be determined not just by what is technically possible, but by the values, priorities, and governance structures we collectively establish to guide its development and deployment.

H. References

[1] Agarwal, N., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In Proceedings of the 35th International Conference on Machine Learning (pp. 60-69).

[2] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2020). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

[3] Anderljung, J., Dafoe, A., Frick, E., Ding, J., & Brundage, M. (2022). AI governance capacity: An international comparison. Oxford Institute for Ethics in AI.

[4] Azizi, S., Athiwaratkun, B., Eyuboglu, T., Hari, A., Heymann, T., & Rajpurkar, P. (2023). Robust medical image understanding: Advances and opportunities. Nature Medicine, 29(4), 977-987.

[5] Bai, Y., Jones, G., Ndousse, K., Askell, A., Chen, A., & Hernandez, D. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

[6] Bansal, G., Nushi, T., Kamar, E., Weld, D., Lasecki, W., & Horvitz, E. (2021). Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(11), 11853-11862.

[7] Barocas, S., Hardt, M., & Narayanan, A. (2021). Fairness and machine learning: Limitations and opportunities. MIT Press.

[8] Barrett, D., Trask, A., Cvitkovic, C., Morrow, B., & Pfau, H. (2024). Compositional verification of neural networks with formal guarantees. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[9] Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code. Polity.

[10] Birhane, A. (2021). The impossibility of automating ambiguity. Artificial Life, 27(1), 44-61.

[11] Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 173-184).

[12] Blodgett, S., Lopez, A., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (pp. 1004-1015).

[13] Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

[14] Bradford, H., Madiega, L., & Zardiashvili, S. (2022). The legal implications of AI service agreements in healthcare. Harvard Journal of Law & Technology, 35(2), 601-649.

[15] Brown, S., Dobbe, R., Taylor, J., & Jardim, R. (2023). Long-term impacts of algorithmic decision systems in public administration. PNAS Nexus, 2(5).

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

[17] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (pp. 2633-2650).

[18] Casper, S., Hadfield-Menell, D., Kenton, T., Shah, S., & Irving, G. (2023). Red-teaming framework for AI systems. In Proceedings of the 40th International Conference on Machine Learning.

[19] Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': The US, EU, and UK approach. Science and Engineering Ethics, 24(2), 505-528.

[20] Chakraborty, A., Wachter, S., & Mittelstadt, B. (2023). Fairness properties as ML goals: Algorithms and trade-offs. In Proceedings of the 41st International Conference on Machine Learning.

[21] Chang, B., Shokri, R., Zhang, X., & Hwang, T. (2021). The price of interpretability: Performance-explainability tradeoffs in machine learning. In Proceedings of the 39th International Conference on Machine Learning.

[22] Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 782-791).

[23] Chen, J., Chang, S., Zhang, X., & Zou, J. (2022). Fairness-promoting attention mechanisms in deep learning. In Proceedings of the 39th International Conference on Machine Learning.

[24] Chen, L., Wu, M., Lyu, M., & King, I. (2023). Culture-adaptive explanation frameworks for multiregional deployment. In Proceedings of the 40th International Conference on Machine Learning.

[25] Chen, W., Liu, Y., Kira, Z., Wang, Y., & Huang, J. (2023). A survey of model watermarking in deep learning. ACM Computing Surveys, 55(12), 1-37.

[26] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. (2020). This looks like that: Deep learning for interpretable image recognition. In Advances in Neural Information Processing Systems, 32.

[27] Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. arXiv preprint arXiv:1810.08575.

[28] Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems, 30.

[29] Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2022). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Conference on Fairness, Accountability and Transparency (pp. 680-691).

[30]

[31] Cohen, J., Sehwag, V., Goyal, P., Wong, E., & Madry, A. (2023). Certified robustness against natural language attacks. In Proceedings of the 40th International Conference on Machine Learning.

[32] Cohen, J., Gadiraju, S., Raji, C., Smart, A., Mittelstadt, B., & Hardt, B. (2023). Ensuring patient privacy in healthcare AI: Legal and regulatory risks. JAMA Network Open, 6(5).

[33] Conitzer, V., Freeman, R., Shah, N., & Wortman Vaughan, J. (2021). Group fairness under composition. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 1013-1024).

[34] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797-806).

[35] Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

[36] Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. MIT Press.

[37] Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

[38] Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature, 538(7625), 311-313.

[39] Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 139-167.

[40] Critch, A., & Krueger, D. (2023). AI research considerations for human existential safety (ARCHES). CGSP Working Paper.

[41] Dafoe, A. (2018). AI governance: A research agenda. Future of Humanity Institute, University of Oxford.

[42] D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., & Halpern, Y. (2022). Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(30), 1-61.

[43] Davis, E., Amodei, D., Doyle, J., & Gadiraju, S. (2024). The evaluation of LLM explanations: Benchmarks, metrics, and findings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[44] Derczynski, L., Feldman, R., Lambert, M., Muennighoff, N., & Sanh, V. (2023). Representational fairness in language model development. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 11526-11546).

[45] Diakopoulos, N., & Friedler, S. (2016). How to hold algorithms accountable. MIT Technology Review, 17(11).

[46] Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer Nature.

[47] D'Ignazio, C., & Klein, L. (2020). Data feminism. MIT Press.

[48] Dobbe, R., Long, T., Fridovich-Keil, D., & Iyer, C. (2022). Hard choices in artificial intelligence: Addressing normative uncertainty through sociotechnical commitments. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1823-1834).

[49] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

[50] arXiv preprint arXiv:1702.08608.

[51] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214-226).

[52] Ehsan, U., & Riedl, M. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. In International Conference on Human-Computer Interaction (pp. 449-466).

[53] Ehsan, U., Liao, Q., Muller, M., Riedl, M., & Weisz, J. (2023). Expanding explainability: Towards social transparency in AI systems. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.

[54] Elhage, N., Nanda, N., Olsson, C., Henighan, T., McCandlish, S., Kaplan, J., Joseph, N., Chen, Z., Stocco, A., Brown, T., & Conerly, T. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread.

[55] Ensign, D., Friedler, S., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2020). Runaway feedback loops in predictive policing. In Conference on Fairness, Accountability and Transparency (pp. 160-171).

[56] European Commission High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI.

[57] Feldman, R., Birch, S., Dickerson, J., & Kearns, K. (2022). Optimizing long-term fairness through data preprocessing. In Proceedings of the 39th International Conference on Machine Learning.

[58] Fisher, M., Mascardi, V., Rozier, K., Schlingloff, N., & Winikoff, J. (2019). Towards a framework for certification of reliable autonomous systems. Journal of Autonomous Agents and Multi-Agent Systems, 35(1), 8-50.

[59] Fleischer, T., Buolamwini, J., Pentland, H., & Wang, T. (2022). Integrating group and individual fairness in classification: New definitions and algorithms. In Proceedings of the 39th International Conference on Machine Learning.

[60] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

[61] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., & Schafer, B. (2020). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 30(4), 575-616.

[62] Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. Harvard Data Science Review, 3(1).

[63] Foulds, J., Islam, R., Keya, K., & Pan, S. (2020). An intersectional definition of fairness. In IEEE 36th International Conference on Data Engineering (pp. 1918-1921).
[64] Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
[65] Gabriel, I. (2022). Artificial intelligence, values, and alignment. Minds and Machines, 32(3), 457-486.
[66] Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Kernion, J.,
[67] Conerly, E., Chen, A., Schiefer, N., Ndousse, K., Joseph, N., Chan, B., Dreyer, M., Choi, D., Bowman, S., Clark, J., Kaplan, C., ... Amodei, D. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858.
[68] Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F.
[69] (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.
[70] Geva, M., Gupta, G., & Berant, B. (2022). Transformer feed-forward layers are key-value memories. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 5484-5495).
[71] Ghosh, S., Acuna, D., & Kolouri, S. (2024). Efficient fairness enforcement for high-dimensional subgroups. In Proceedings of the 41st International Conference on Machine Learning.
[72] Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 45.
[73] Green, B., & Chen, S. (2022). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-33.
[74] Greene, D., Hoffmann, A., & 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).
[75] Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.
[76] Hadfield-Menell, D., Russell, S., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In Advances in Neural Information Processing Systems, 29.
[77] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, 29, 3315-3323.
[78] Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, T., & Erhan, A. (2021). Unsolved problems in ML safety. arXiv preprint arXiv:2109.13916.
[79] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2024). Towards the systematic reporting of the energy and carbon

footprints of machine learning. Journal of Machine Learning Research, 25(1).

[80] High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission.

[81] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-16).

[82] Holstein, K., van der Linden, A., Pappas, N., & Matias, M. (2023). Evaluation frameworks for long-horizon algorithmic impacts in education. In Proceedings of the 2023 Conference on Artificial Intelligence, Ethics, and Society.

[83] Hong, S., Hullman, J., & Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and needs. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), 1-26.

[84] Hoyt, R., Csepregi, B., Tan, J., Hong, S., Gambarotta, F., & Icard, J. (2024). Visual explanations for multimodal foundation models: Techniques and limitations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[85] Ibrahim, A., Sinha, B., & Sharma, P. (2023). Analyzing intellectual property claims in commercial AI contracts for healthcare applications. Nature Biotechnology, 41(3), 420-427.

[86] Irving, G., & Askell, A. (2019). AI safety needs social scientists. Distill, 4(2), e14.

[87] Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv preprint arXiv:1805.00899.

[88] Jayaraman, K., & Evans, D. (2023). Evaluating differentially private machine learning in practice. In 30th USENIX Security Symposium (pp. 1907-1924).

[89] Jeon, H., Kim, D., Wang, H., & Qi, Y. (2020). Reward-rational implicit choice: A unifying formalism for reward learning. arXiv preprint arXiv:2010.04728.

[90] Jiang, L., Balle, J., Welleck, S., & Feizi, S. (2024). Eliciting human preferences from language feedback: Methods, limitations, and opportunities. In Proceedings of the 41st International Conference on Machine Learning.

[91] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

[92] Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., & Wu, Z. (2022). An algorithmic framework for fairness elicitation. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 767-779).

[93] Kallus, N., & Zhou, A. (2022). Fairness, welfare, and equity in randomized decision-making. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214-228).

[94] Karpus, J., Sinha, A., Coopersmith, A., & Saenko, K. (2024). Explaining explanations: Toward user-centered explainable AI. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[95] Kasirzadeh, A., & Gabriel, D. (2023). In search of fairness: ethical machine learning in context. Patterns, 4(8), 100779.

[96] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2022). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.

[97] Kizilcec, R., & Reich, J. (2023). Large-scale educational AI: Methods and evidence for evaluation. npj Science of Learning, 8(1), 1-10.

[98] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

[99] Koh, P., Nguyen, T., Tang, Y., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In International Conference on Machine Learning (pp. 5338-5348).

[100] Koh, P., Sagawa, S., Marklund, H., Xie, S., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M.,

[101] Phillips, R., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S., Leskovec, J., Kundaje, A., ... Liang, P. (2022). WILDS: A benchmark of in-the-wild distribution shifts. In International Conference on Machine Learning (pp. 11598-11618).

[102] Kojima, T., Gu, S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems, 36.

[103] Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2022). Faithful and customizable explanations of black box models. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 462-473).

[104] Lage, I., Wu, Z., Kumar, K., Prabhu, V., Mitchell, M., Singh, S., & Doshi-Velez, F. (2023). From local to global explanations: A computational framework for unified interpretability. In Proceedings of the 40th International Conference on Machine Learning.

[105] Lee, M., & Singh, S. (2021). FAIRVIS: Visual analytics for discovering intersectional bias in machine learning. In 2021 IEEE Conference on Visual Analytics Science and Technology (pp. 46-57).

[106] Leike, J., Schulman, R., & Lampinen, D. (2022). Alignment of language agents. arXiv preprint arXiv:2103.14659.

[107] Leike, J., Schulman, J., & Evans, O. (2024). Safety and alignment guarantees through neural network surgery. arXiv preprint arXiv:2401.14295.

[108] Li, J., Bao, J., Sun, X., Xu, J., & Qiu, X. (2022). FairAdapter: Specialized parameter-efficient tuning for improving fairness in large language models. In Proceedings of the 39th International Conference on Machine Learning.

[109] Li, Z., & Qiu, X. (2024). Verifying logical reasoning in language models with reinforcement learning. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.
[110] Liang, P., Bommasani, R., Le, T., Zheng, R., & Hashimoto, T. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.
[111] Liu, L., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In International Conference on Machine Learning (pp. 3150-3158).
[112] Liu, Y., Jain, S., Tramer, F., & Schmidt, L. (2023). A comprehensive study of backdoor attacks and defenses for neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3282-3301.
[113] Liu, Y., Zeng, Y., & Pearl, J. (2023). Causal fairness analysis with path-specific counterfactuals. In Proceedings of the 40th International Conference on Machine Learning.
[114] Locatello, F., Higgs, G., Tschannen, M., Gehrmann, S., Marabelli, R., Kim, D., Rahimi, A.,
[115] Dieleman, F., von Oswald, C., Belilovsky, E., Goyal, S., Blechschmidt, P., Kipf, T., Zaheer, M., Raffel, C., Chen, H., & Montanari, A. (2023). A unified approach to learning with fair representations. In Advances in Neural Information Processing Systems, 36.
[116] Lum, K., & Johndrow, J. (2019). A statistical framework for fair predictive algorithms. arXiv preprint arXiv:1610.08077.
[117] Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 30, 4765-4774.
[118] Madaio, A., Zhou, G., & Song, H. (2024). Responsible AI maturity model: Assessing organizational integration beyond documentation. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.
[119] Madaio, 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).
[120] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations.
[121] Maas, A. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. Contemporary Security Policy, 40(3), 285-311.
[122] Martinez, N., Bertran, M., & Sapiro, G. (2024). Fairness with minimal harm: A Pareto-optimal approach for healthcare. In Proceedings of the AAAI Conference on Artificial Intelligence, 38(12).
[123] Martinez, N., Bertran, M., & Sapiro, G. (2023). Minimax Pareto fairness: A multi objective perspective. In International Conference on Machine Learning (pp. 24063-24080).

[124] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.
[125] Metcalf, J., Moss, E., Watkins, E., Singh, R., & Elish, M. (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).
[126] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.
[127] Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501-507.
[128] Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 279-288).
[129] Mitchell, S., Potash, S., & Barocas, S. (2021). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Annual Review of Statistics and Its Application, 8, 141-163.
[130] Mohamed, S., Png, M., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33(4), 659-684.
[131] Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning—A brief history, state-of-the-art and challenges. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 417-431).
[132] Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics, 27(2), 667-705.
[133] Moss, E., & Metcalf, J. (2020). Ethics owners: A new model of organizational responsibility in data-intensive tech companies. arXiv preprint arXiv:2005.07514.
[134] Moss, E., Watkins, A., Singh, R., Elish, M., & Metcalf, J. (2023). Organizational integration of algorithmic systems: Lessons for AI policy and governance. Journal of Technology Policy & Law, 2(1), 61-84.
[135] Mukherjee, S., Yurochkin, Y., Kandemir, M., & Sun, X. (2023). Learning individually fair representations via bilevel optimization. In Proceedings of the 40th International Conference on Machine Learning.
[136] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133).
[137] Noble, S. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.
[138] Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill, 5(3), e00024.001.

[139] Pan, A., Steinhardt, A., Cunningham, J., & Shi, J. (2022). Risk-averse inverse preference learning and its application to value-aligned AI systems. In Proceedings of the 39th International Conference on Machine Learning.
[140] Park, H., Meel, A., Chen, S., & Wang, F. (2023). Fairness calibration: Towards fair and calibrated machine learning models. In Proceedings of the 40th International Conference on Machine Learning.
[141] Park, J., McCoy, D., Zalewski, M., & Goldberg, P. (2024). A framework for interdisciplinary evaluation of advanced AI systems. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.
[142] Pasquale, F. (2020). The black box society: The secret algorithms that control money and information. Harvard University Press.
[143] Passi, S., & Barocas, S. (2022). Problem formulation and fairness. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1017-1027).
[144] Patel, S., Wang, J., Zaheer, M., & Ré, C. (2023). Multimodal uncertainty in language and vision: Methods and applications. In Proceedings of the 40th International Conference on Machine Learning.
[145] Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54-60.
[146] Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books.
[147] Peng, Y., Dhingra, S., Wang, Q., Moon, T., & Zahia, O. (2023). Leveraging medical knowledge and fine-tuning large language models for healthcare. arXiv preprint arXiv:2304.13976.
[148] Prunkl, C., & Whittlestone, J. (2020). Beyond near- and long-term: Towards a clearer account of research priorities in AI ethics and society. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 138-143).
[149] Räuker, T., Payne, C., Simonyan, K., Zisserman, A., & Russakovsky, O. (2023). Mechanistic interpretability of neural networks by visualizing underlying concept learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22183-22193).
[150] Raji, I., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & 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).
[151] Raji, I., Buolamwini, J., Selbst, P., Wornham, J., Hong, C., & Obradovich, Z. (2022). Outsider oversight: Designing a third party audit ecosystem for AI governance. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 557-571).
[152] Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality:

[153] Practitioner perspectives on enablers for shifting organizational practices. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-23.
[154] Rebuffi, S., Gowal, S., Calian, D., Stimberg, F., Wiles, O., & Mann, T. (2021). Fixing data augmentation to improve adversarial robustness. In Advances in Neural Information Processing Systems, 34, 9393-9403.
[155] Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144).
[156] Richardson, H., Newell, C., & Kleinman, M. (2022). Public sector algorithms: A public interest-informed framework. Regulation & Governance, 16(4), 1132-1149.
[157] Richardson, T., & Mehta, S. (2024). Algorithmic fairness and disparate impact: Methodological gaps and legal challenges. Annual Review of Law and Social Science, 20.
[158] Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. AI & Society, 36(1), 59-77.
[159] Roth, W., Peharz, R., Tschiatschek, S., & Brünig, F. (2024). The impossible benchmark: How not to measure fairness in deep learning. In Proceedings of the 41st International Conference on Machine Learning.
[160] Rudin, C., & Radin, D. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Science Review, 1(2).
[161] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-15).
[162] Sambasivan, N., Mitra, A., Vemuri, J., Vaniea, A., & Agrawal, M. (2022). Think globally, act locally: Global-south perspectives on explainable AI. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.
[163] Saunders, W., Cunningham, J., Chen, A., Whang, J., Lindgren, M., Feng, X., & Hendrycks, D. (2022). Reinforcement learning from human feedback with evaluator consensus. arXiv preprint arXiv:2212.08051.
[164] Selbst, A., Boyd, D., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59-68).
[165] Sendak, M., Gao, M., Nichols, M., Lin, J., & Balu, W. (2023). Presenting machine learning model information to clinical end users with model facts labels. npj Digital Medicine, 6(1), 21.
[166] Sendak, M., Ratliff, W., Sarro, D., Alderton, E., Futoma, J., Gao, M., Nichols, M., Smith, M., Reiter, F., & Yasouri, S. (2020). Real-world integration of

artificial intelligence into clinical workflows: Multisite implementation. JMIR Medical Informatics, 8(7), e17416.

[167] Sheller, M., Reina, G., Edwards, B., Martin, J., & Bakas, S. (2023). Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. Nature Communications, 14(1), 645.

[168] Singhal, K., Azizi, J., Tu, T., Rehman, W., Wessler, P., Neal, B., Huang, M., Fan, K., Reyes, M., Eshaghi, S., & Xue, R. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

[169] Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation is not a design fix for machine learning. In Proceedings of the 2022 International Conference on Machine Learning.

[170] Subbaswamy, A., Adams, R., & Saria, S. (2021). Evaluating model robustness to dataset shift. In Proceedings of the 38th International Conference on Machine Learning (pp. 9859-9870).

[171] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In International Conference on Learning Representations.

[172] Taylor, S., & Morley, J. (2022). Healthcare AI systems and patient data: A survey of patient perspectives. Journal of Medical Internet Research, 24(2), e32497.

[173] Thirunavukarasu, D., Camilleri, D., & Montacute, C. (2023). Large language models can accurately predict surgeon behavior in the operating room. PLOS Digital Health, 2(12), e0000390.

[174] Tiu, E., Talius, E., Patel, I., Landefeld, S., Zou, J., & Lungren, C. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nature Biomedical Engineering, 6(11), 1230-1240.

[175] Topol, J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

[176] Topol, E., & Nundy, S. (2021). The integration of AI in medicine. Nature Medicine, 27(2), 183-185.

[177] Vallor, S. (2018). Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press.

[178] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30.

[179] Veale, M., Van Kleek, M., & Binns, R. (2020). The hard problem of data due process in a public sector algorithmic decision-making system. Government Information Quarterly, 37(2), 101466.

[180] Verma, S., & Rubin, J. (2022). Fairness definitions explained. In Proceedings of the 2022 ACM/IEEE International Workshop on Software Fairness (pp. 18-29).

[181] Vig, J. (2019). A multiscale visualization of attention in the transformer model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 37-42).
[182] Wagner, I., & Eckhoff, D. (2018). Technical privacy metrics: A systematic survey. ACM Computing Surveys, 51(3), 1-38.
[183] Wan, M., Zhang, X., Ghosh, A., & Wei, K. (2024). Dynamic fairness: Continual group fairness with evolving groups. In Proceedings of the 41st International Conference on Machine Learning.
[184] Wang, D., Feng, A., Wang, W., Feng, S., & Li, S. (2024). Neural-symbolic architectures for safety-critical healthcare applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
[185] Wang, D., & Hajli, Z. (2022). The mayo clinic approach to AI governance: Balancing innovation and responsibility in healthcare. New England Journal of Medicine Catalyst, 3(6).
[186] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, L., & Hu, X. (2019).
[187] Backdoor attacks and defenses in deep learning: A survey. arXiv preprint arXiv:2007.10086.
[188] Wang, H., Wu, Z., Liu, Z., Jiang, H., Wang, L., Shan, H., & Chen, J. (2022). Towards interpretable deep learning models for healthcare. Nature Communications, 13(1), 3738.
[189] Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2020). Deep learning for sensor-based activity recognition: A survey. Pattern Recognition Letters, 119, 3-11.
[190] Wang, T., Chang, X., & Wang, L. (2021). A practical approach to adaptive explainable AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.
[191] Wang, Y., Cai, L., Hu, J., Yuan, X., Ding, X., & Wang, N. (2024). Measuring fairness in classification: A new metric for composite awareness. In Proceedings of the 41st International Conference on Machine Learning.
[192] Washington, A., Theodos, A., Moore, D., & Barocas, R. (2023). Managing algorithmic risks: Data governance approaches for the age of AI. Health Affairs, 42(8), 1129-1136.
[193] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., & Le, Q. (2022). Chain of thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, 35, 24824-24837.
[194] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., & Metzler, D. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.
[195] West, S., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. AI Now Institute.
[196] Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. Nuffield Foundation.

[197] Whittlestone, J., Arulkumaran, J., & Crosby, M. (2021). The technical landscape of responsible AI. Center for the Governance of AI.
[198] Wong, P. (2020). Cultural differences in algorithm fairness: A case study of emotion recognition across eastern and western cultures. In Conference on Fairness, Accountability, and Transparency (pp. 658-668).
[199] Wong, E., & Kolter, Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. In International Conference on Machine Learning (pp. 5286-5295).
[200] Yang, K., Zhang, S., Wang, C., Wang, T., & Zou, J. (2024). Mechanistic interpretability analysis of discrete key-value memory in transformers. In Proceedings of the 41st International Conference on Machine Learning.
[201] Yang, K., Qian, K., Locatello, F., Walecki, R., Lotfian, S., Ding, D., Wang, T., & Zaharia, M. (2023). Algorithmic fairness: Choosing the right metric. In Advances in Neural Information Processing Systems, 36.
[202] Yang, Q., Steinfeld, C., & Zimmerman, C. (2020). Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-11).
[203] Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-13).
[204] Yeung, K., Howes, A., & Pogrebna, G. (2020). AI governance by human rights-centred design, deliberation and oversight: An end to ethics washing. In The Oxford Handbook of Ethics of AI (pp. 77-108). Oxford University Press.
[205] Zhang, B., & Bareinboim, E. (2018). Fairness in decision-making—The causal explanation formula. In Thirty-Second AAAI Conference on Artificial Intelligence.
[206] Zhang, B., Wu, C., Cui, C., Du, M., Zhang, H., Chen, Y., Wang, Y., Yao, H., Wu, Y., & Hu, X. (2023). Interpreting deep learning models in natural language processing: A survey. IEEE Transactions on Knowledge and Data Engineering, 35(3), 2252-2276.
[207] Zhang, B., Xia, L., Fang, J., Pan, S., Chen, J., & Cai, X. (2022). Visual verification of neural network behavior via semantic adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18041-18050).
[208] Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T.,
[209] Manyika, J., Niebles, J., Sellitto, M., Shoham, Y., Clark, J., & Perrault, R. (2022). The AI index 2022 annual report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University.
[210] Zhang, H., Zhou, A., Zeng, X., Yang, Y., Liu, Y., & Pan, K. (2022). Finding discriminatory rules in enterprise policies with explainable machine

learning. In Proceedings of the 39th International Conference on Machine Learning.

[211] Zhang, M., Hagedorn, K., Dai, A., Liang, P., Narayanan, A., Wang, S., & Zou, J. (2022).

[212] Understanding the capability of large language models through explanation evaluation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 8755-8771).

[213] Zhang, Y., Bellamy, R., & Varshney, K. (2021). Joint optimization of AI fairness and utility: A human-centered approach. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 856-863).

[214] Zhao, S., Finn, C., Schulman, J., & Abbeel, P. (2023). Value learning under language-described preferences: Representing multiple feasible reward functions. arXiv preprint arXiv:2303.06247.

[215] Zhao, S., & Brantley, K. (2021). Counterfactual reasoning for fair representations: An adversarial approach. In Proceedings of the 38th International Conference on Machine Learning.

[216] Zhu, H., Levi, A., Jain, S., & Narasimhan, B. (2023). Fairness in machine learning: A multi-objective perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, 37(9), 11236-11243.