Evaluating Artificial Intelligence Beyond Performance

A Synthesis of Methodologies for Ethical, Environmental, Social, and Governance Assessment

By TowerIO LLC

Evaluating Artificial Intelligence Beyond Performance

A Synthesis of Methodologies for Ethical, Environmental, Social, and Governance Assessment

By TowerIO LLC

Forward

This report was born from a frustrating yet illuminating endeavor: the attempt to conduct a responsible and thorough evaluation of competing AI platforms. The initial goal was simple - to make an informed decision. What quickly became apparent, however, was a systemic deficiency in the tools and frameworks available. The overwhelming emphasis on performance benchmarks – speed, accuracy, output quality – overshadowed the equally critical dimensions of ethical impact, environmental sustainability, accessibility and equity, governance practices, and broader societal consequences. This stark reality, the industry's apparent willingness to prioritize technical prowess over holistic responsibility, is, to put it plainly, terrifying.

To address this gap, the research and methodologies detailed in the following pages take on an ambitious scope and depth. This report represents a synthesis of knowledge spanning diverse and traditionally siloed fields. It integrates insights from moral philosophy and computer science, from environmental science and social work, from legal theory and the lived experience of marginalized communities. To achieve the necessary depth and granularity, the analysis employs principles of fractal iteration, drilling down into sub-criteria and exploring the complex interplay between various factors.

The creation of this report itself represents a novel approach to research, the implications of which are significant. The bulk of the investigation, resulting in over 100 pages of structured data, was conducted and synthesized by a "Gem" agent powered by Gemini 2.5 Pro Deep Research. This one-shot prompt engineering feat demonstrates the potential of advanced AI to conduct sophisticated research, synthesize information from diverse sources, and produce coherent, well-structured reports on complex topics with remarkable efficiency and speed. It validates the power of effective prompt engineering and frameworks to elicit high-quality output from AI models, suggesting that carefully crafted prompts can guide AI to perform intricate tasks with minimal human intervention. This has major implications for the automation of various complex tasks, such as literature reviews, policy analysis, and report generation across different domains, and for the scalability of operations that rely heavily on in-depth research.

However, let us be clear: this is not a story of AI replacing human intellect. The human component remains indispensable. The creation of the prompt itself, the design of the fractal iteration framework, the contextual understanding that defines the report's scope and purpose, and the critical evaluation of the AI's output – these are all fundamentally human endeavors. It is the human orchestrator, conducting the AI "orchestra," who shapes the final outcome.

And beyond individual contribution, this report reflects the intellectual and creative labor of an organization. At TowerIO, we have developed proprietary processes for designing AI agents and eliciting information through specialized interview techniques and fractal iteration prompting. The AI, in this context, functions as an extremely competent research assistant, operating within a carefully constructed framework. Therefore, the resulting work is most accurately attributed to TowerIO as a collective entity, a symbiotic collaboration between human ingenuity, AI capabilities, and organizational knowledge.

This emphasis on collective, organizational "authorship" challenges traditional notions of individual creation. If the legal system can grant personhood to corporations, recognizing their capacity for ownership and action, then surely it can acknowledge the intellectual and creative contributions of a company in orchestrating and guiding AI-assisted research. I will stand firmly by this principle.

This work comes as a product of my efforts at TowerIO. Our organization is dedicated to identifying opportunities to adapt business automation technology and emerging generative AI as assistive technology, with a particular focus on serving disabled individuals and the nonprofit organizations that champion their cause. This mission is deeply personal. As a brain tumor survivor who experienced homelessness as recently as two years ago, I understand the urgent need for equitable access and ethical considerations to be at the forefront of technological development. The journey from that precarious situation to authoring this report, leveraging cutting-edge AI to scrutinize AI itself, underscores the resilience of the human spirit and the potential for technology to empower change.

Ultimately, this report is more than just an analysis of AI evaluation methodologies. It is a call for a fundamental shift in perspective. It argues for elevating the standards by which we judge AI systems, demanding a level of rigor, comprehensiveness, and transparency commensurate with their transformative power. It is my sincere hope that this work will contribute to a future where AI is developed and deployed responsibly, ethically, and for the benefit of all.

Bill O’Rly
Founder
TowerIO LLC

Executive Summary

Artificial Intelligence (AI) offers immense potential but requires evaluation beyond traditional performance metrics due to its profound ethical, social, environmental, and governance implications. Misused or faulty AI can cause significant harm, including safety failures, discrimination, privacy violations, and negative environmental impacts. The complexity, dynamic nature, and socio-technical integration of AI necessitate a holistic assessment to build and maintain trust. Trustworthy AI encompasses ethical adherence, legal compliance, and technical robustness (including safety, security, fairness, transparency, accountability, and privacy).

This report synthesizes methodologies for evaluating AI across five critical non-performance categories:

1. **Ethical Considerations:** Alignment with moral values, including fairness/bias mitigation, transparency/explainability, accountability, privacy, safety, human oversight, and misuse potential.
2. **Environmental Impact:** Quantifying the ecological footprint via energy consumption, carbon emissions, and resource use (water, hardware).
3. **Accessibility and Equity:** Ease of use for diverse populations (including those with disabilities), cost/affordability, language support, and digital divide implications.
4. **Responsibility and Governance:** Practices of AI actors regarding organizational transparency, community engagement, regulatory compliance, and data/IP ownership.
5. **Societal Impact:** Broader effects on employment, education, creativity/culture, democracy, and social norms.

A key challenge is quantifying qualitative or context-dependent aspects, hampered by a lack of universal metrics and standardized methodologies. This report's goal is to synthesize existing and adaptable frameworks, metrics, tools, and benchmarks across these categories, analyzing their strengths and weaknesses concerning quantification and cross-disciplinary adaptation. This synthesis informs the exploration of developing a comparative ranking system for AI based on non-performance factors, acknowledging the complexities and limitations inherent in measuring and comparing these interconnected dimensions.

Methodologies are drawn from various fields.

Ethical evaluation adapts concepts from applied ethics and law, utilizing fairness metrics (Demographic Parity, Equal Opportunity), explainability techniques (LIME, SHAP), accountability mechanisms (audit trails, lineage tracking), Privacy-Enhancing Technologies (Differential Privacy, Homomorphic Encryption), safety validation methods (robust testing, formal methods), human oversight assessment (empirical testing), and misuse evaluation (AI red teaming). Documentation standards like Model Cards and Datasheets for Datasets enhance transparency. A significant gap exists between stated ethical principles and verifiable implementation, risking "ethics washing".

Environmental assessment leverages methods like Life Cycle Assessment (LCA), carbon/water footprinting, and metrics such as Power Usage Effectiveness (PUE), Water Usage Effectiveness (WUE), and energy consumption (kWh). Tools like CodeCarbon estimate operational emissions, but LCA is needed to capture the significant embodied impacts of hardware manufacturing and disposal. Evaluating the energy-water-carbon nexus requires considering trade-offs (e.g., PUE vs. WUE) and local context (grid intensity, water stress).

Accessibility and Equity evaluation adapts Web Content Accessibility Guidelines (WCAG) and utilizes usability testing with people with disabilities. It examines affordability through Total Cost of Ownership (TCO) analysis and assesses multilingual support using NLP benchmarks and metrics (BLEU, ROUGE). Addressing the digital divide requires analyzing access, skills, usage, and outcome metrics across demographics, integrating quantitative and qualitative methods. Technical accessibility is a prerequisite for broader digital equity.

Responsible Governance relies on frameworks like the NIST AI RMF, the EU AI Act, and standards such as ISO/IEC 42001 and IEEE P7000. Key practices include enhancing transparency through documentation, community engagement via participatory methods, ensuring regulatory compliance through audits and assessments, and navigating complex IP/data ownership issues. Effective governance is foundational for evaluating other dimensions.

Societal Impact assessment employs methods from social sciences (sociology, economics, political science, anthropology) and frameworks like Social Impact Assessment (SIA). It examines AI's dual impact on employment (displacement vs.

augmentation/creation), education (personalization vs. bias/privacy risks), creativity (tools vs. copyright/labor disruption), and democracy (disinformation/polarization vs. potential efficiency gains). Establishing causality and finding reliable measurements for broad societal impacts remain challenging.

Synthesizing these diverse methodologies highlights the potential of adapting approaches like LCA, SIA, the Capability Approach, WCAG, ethnography, and HRIA/FRIA. Comparative analysis could utilize AI indices, Multi-Attribute Decision Making (MADM) techniques, or risk-based categorization. However, a universal ranking system faces severe limitations due to challenges in quantification, context-dependency, metric standardization, and managing inherent trade-offs between principles. A simple additive score is likely misleading. Socio-technical evaluation approaches, integrating diverse methods and stakeholders, are essential.

Future directions include fostering interdisciplinary research for metric development and trade-off analysis, advancing standardization efforts, encouraging adoption of governance frameworks and socio-technical practices, mandating greater transparency, and promoting AI literacy.

I. Introduction: The Imperative of Evaluating AI Beyond Performance

Defining the Scope: Why Non-Performance Aspects Matter

Artificial Intelligence (AI) holds immense potential to transform economies, enhance scientific discovery, and address complex global challenges.[1] However, the power of AI extends far beyond its ability to perform specific tasks efficiently. Its integration into society carries profound ethical, social, environmental, and governance implications that demand rigorous assessment.[2] Faulty or misused AI systems can lead to significant harm, including safety failures, perpetuation of discrimination, violations of privacy, erosion of individual freedoms, and negative environmental consequences.[2]

The risks associated with AI are multifaceted, impacting individuals through threats to civil liberties, physical and psychological safety, and economic opportunity; organizations through operational disruptions, reputational damage, and security breaches; and broader ecosystems, including financial systems, supply chains, and the natural environment.[5] Evaluating these non-performance dimensions is critical because AI systems are often complex, operate in dynamic contexts, and can evolve in ways that are difficult to predict or understand.[5] Furthermore, AI systems are inherently socio-technical; their development and impact are deeply intertwined with societal dynamics, human behavior, and existing structural inequalities.[5]

Therefore, establishing and maintaining trust in AI necessitates a holistic evaluation that goes beyond traditional performance metrics like accuracy or speed.[5] Trustworthy AI is increasingly defined as encompassing ethical adherence, legal compliance, and technical robustness, including safety, security, fairness, transparency, accountability, and privacy.[10] Assessing these characteristics is essential for realizing the benefits of AI while proactively mitigating its potential harms.[1]

Overview of the Five Core Evaluation Categories

This report focuses on synthesizing methodologies for evaluating AI across five critical non-performance categories:

1. **Ethical Considerations:** Examining alignment with moral values and human rights, including fairness and bias mitigation, transparency and explainability, accountability, privacy protection, safety, human oversight, and the potential for misuse.
2. **Environmental Impact:** Quantifying the ecological footprint of AI systems, focusing on energy consumption, carbon emissions, and resource utilization

(including water and hardware sustainability).

3. **Accessibility and Equity:** Assessing the ease of use for diverse populations, including people with disabilities, the cost and affordability of AI systems, language support, and implications for the digital divide.
4. **Responsibility and Governance:** Evaluating the practices of AI developers and deployers, including organizational transparency, community engagement, compliance with regulations and standards, and approaches to intellectual property and data ownership.
5. **Societal Impact:** Analyzing the broader effects of AI on society, including employment and the future of work, education and learning, creativity and the arts, democracy and civic engagement, and cultural norms and values.

The Challenge and Goal: Synthesizing Methodologies for Holistic AI Assessment and Potential Ranking

A central challenge in evaluating these non-performance dimensions lies in the difficulty of quantifying aspects that are often qualitative or context-dependent.[5] There is currently a lack of universally accepted metrics, benchmarks, and standardized evaluation methodologies for many of these areas.[4] Furthermore, effective evaluation often requires adapting methods from diverse fields such as social sciences, environmental science, ethics, human-computer interaction (HCI), and policy analysis.[16]

The primary goal of this report is to investigate and synthesize existing and adaptable methodologies—including frameworks, specific metrics, measurement tools, datasets, and benchmarks—across the five core categories outlined above [Query]. This synthesis aims to provide a structured overview of the current state-of-the-art in non-performance AI evaluation, analyzing the strengths, weaknesses, and challenges associated with different approaches, particularly concerning quantification and cross-disciplinary adaptation [Query].

Ultimately, this research seeks to inform the potential development of a comparative ranking system for AI based on these non-performance factors [Query]. Such a system could, in principle, help guide procurement decisions, policy development, and responsible innovation. However, the feasibility and utility of such a ranking system depend critically on understanding the complexities and limitations inherent in measuring and comparing these diverse dimensions.[4]

The various dimensions of non-performance evaluation are deeply interconnected. Ethical considerations, such as fairness and bias [6], are inextricably linked to issues of accessibility and equity [Query III]. Failures in governance, like inadequate data management practices [18], often underpin ethical breaches such as discriminatory outcomes.[2] Similarly, governance mechanisms like transparency and auditability [19] are prerequisites for establishing accountability.[6] The environmental costs associated with AI, including energy and water consumption [21], are not distributed evenly and raise significant questions of social justice and equity, particularly concerning resource allocation and impact on vulnerable communities.[23] Accessibility limitations, preventing certain groups from using or benefiting from AI [Query III], represent a fundamental failure of equity and can exacerbate societal divides.[24] Furthermore, governance decisions regarding intellectual property [25] or data ownership directly influence societal domains like the creative industries.[26] Evaluating these non-performance aspects, therefore, demands a systemic perspective that recognizes these interdependencies. A simple, additive approach to assessment or ranking, treating each category in isolation, would likely overlook crucial interactions and could lead to misleading conclusions. Any robust evaluation framework, particularly one intended for comparative ranking, must account for these complex relationships and potential trade-offs.

II. Frameworks and Metrics for Ethical AI Evaluation

Evaluating the ethical dimensions of AI systems is paramount for ensuring they align with human values and societal norms. This involves assessing multiple facets, from fairness and transparency to accountability and safety.

Assessing Algorithmic Fairness and Mitigating Bias

Definitions & Concepts: Fairness in AI aims to ensure equitable treatment and outcomes for individuals and groups, irrespective of protected attributes like race, gender, or age.[27] It involves actively identifying and mitigating harmful biases, which can manifest in various forms: systemic (reflecting societal inequalities), computational/statistical (arising from data or algorithms), and human-cognitive (introduced by developers or users).[5] Fairness is not merely about statistical parity but about avoiding unjustified adverse effects on individuals or groups.[29] Importantly, the definition and operationalization of fairness are highly context-specific [30], and different mathematical formulations of fairness can sometimes be mutually exclusive, leading to "fairness impossibility" theorems.[31] Group fairness metrics compare outcomes across predefined groups, while individual fairness focuses on treating similar individuals similarly.[32]

Metrics: A variety of quantitative metrics have been proposed to measure group fairness:

- *Demographic Parity (or Statistical Parity):* Requires the likelihood of a positive outcome (e.g., loan approval) to be equal across different groups. Measured by Statistical Parity Difference (difference in rates) or Disparate Impact (ratio of rates).[17]
- *Equal Opportunity:* Requires the true positive rate (sensitivity) to be equal across groups. Measured by the Equal Opportunity Difference.[33]
- *Equalized Odds:* Requires both the true positive rate and the false positive rate to be equal across groups. Measured by the Average Odds Difference or the Equalized Odds Difference (maximum absolute difference).[33]
- Other metrics include differences/ratios in error rates, false positive/negative rates, false discovery/omission rates, and predictive values.[33] The OECD.AI metrics catalogue includes 'Equal performance', ensuring a model is equally accurate across groups.[35] Individual fairness metrics include *Consistency*, measuring label similarity for similar instances [32], and the *Generalized Entropy Index*, measuring inequality in benefit distribution.[32]

Tools & Frameworks: Several open-source toolkits facilitate fairness assessment and mitigation:

- *IBM AI Fairness 360 (AIF360):* Provides a comprehensive library of fairness metrics and bias mitigation algorithms (pre-processing, in-processing, post-processing).[17]
- *Microsoft Fairlearn:* Offers tools to assess fairness (e.g., demographic_parity_difference, equalized_odds_difference) and implement mitigation techniques like Exponentiated Gradient, Grid Search, and Threshold Optimizer.[34]
- *Google's What-If Tool* and *TensorFlow Fairness Indicators:* Provide interactive visualization and analysis capabilities for fairness evaluation.[37] Major governance frameworks mandate fairness assessments. The NIST AI Risk Management Framework (RMF) requires fairness evaluation and bias management as part of its 'Measure' function, referencing NIST SP 1270 for bias guidance.[5] The EU AI Act imposes strict fairness requirements, particularly for high-risk systems, including bias mitigation and data governance.[2]

Challenges: Key challenges include the difficulty of defining fairness appropriately for a given context [30], the inherent trade-offs between different fairness metrics [31], ensuring representative and high-quality training data [43], and preventing algorithms from amplifying existing societal biases.[33]

Evaluating Transparency and Explainability (Interpretability)

Definitions & Concepts: Transparency refers to the availability of information about an AI system, its capabilities, limitations, and outputs, tailored to the stakeholder and context.[5] Explainability (or XAI) focuses on elucidating the internal mechanisms or logic driving an AI's decisions or predictions ("why" or "how" a decision was made).[5] Interpretability relates to conveying the meaning of an AI system's output in understandable terms.[5] Together, these concepts are crucial for building user trust, enabling debugging, ensuring accountability, and verifying fairness.[43]

Methods & Tools: Explainability techniques can be local (explaining individual predictions) or global (explaining overall model behavior).

- *Local Interpretable Model-agnostic Explanations (LIME):* Approximates a complex model locally with a simpler, interpretable model (e.g., linear regression) by perturbing the input instance and observing output changes.[45] It is model-agnostic but can be unstable.[47]
- *SHapley Additive exPlanations (SHAP):* Based on cooperative game theory (Shapley values), it attributes the contribution of each feature to a specific prediction compared to a baseline.[45] Provides both local and global explanations

and indicates feature impact direction (positive/negative).[50] Computationally more intensive than LIME, but specialized versions exist (TreeSHAP, DeepExplainer/DeepSHAP, Expected Gradients).[50]

- *Other Techniques:* Include feature importance analysis, partial dependence plots, counterfactual explanations, and inherently interpretable models (e.g., decision trees, linear regression).
- *Toolkits:* Microsoft's Responsible AI Toolbox [53] includes InterpretML [53] and Error Analysis.[53] The Holistic AI Library [46] and Azure AI [53] also offer explainability tools. ELI5 is another framework.[45]

Documentation Standards: Standardized documentation is key to transparency.

- *Model Cards:* Provide structured summaries ("nutrition labels") detailing a model's intended use, performance characteristics (including across different groups), limitations, training data, and ethical considerations.[20] Pioneered by Google [20] and adopted by others like Salesforce.[20]
- *Datasheets for Datasets:* Document dataset motivation, composition, collection processes, preprocessing, intended uses, distribution, and maintenance, increasing transparency about the data underpinning AI models.[59]

Frameworks & Regulations: Transparency and explainability are core tenets of most AI governance frameworks. The NIST RMF 'Measure' function requires AI models to be explained and outputs interpreted.[5] The EU AI Act mandates transparency for high-risk systems (requiring technical documentation explaining system operation) and specific obligations for systems like chatbots (disclosure of AI interaction) and deepfakes (labeling).[41] IEEE 7001 sets standards for transparency in autonomous systems.[3] OECD AI Principles emphasize transparency and responsible disclosure.[10] Corporate frameworks (e.g., Microsoft [71], AWS [27], AMD [72]) invariably list transparency/explainability as core principles.

Ensuring Accountability and Responsibility

Definitions & Concepts: Accountability in AI refers to the obligation of AI actors (developers, deployers, operators) to take responsibility for the proper functioning and outcomes of AI systems, based on their roles and the context.[7] It involves being answerable for AI decisions and impacts, particularly when harm occurs. Accountability is intrinsically linked to transparency, as understanding how a system works and who was involved is necessary to assign responsibility.[5] The specific mechanisms and locus of accountability can vary depending on cultural, legal, and sectoral contexts.[5]

Mechanisms: Effective accountability relies heavily on traceability – the ability to reconstruct the lifecycle and decision-making process of an AI system.[10] Key mechanisms include:

- *Audit Trails and Logging:* Maintaining detailed, immutable records of system operations, data usage, model changes, user interactions, and decisions made.[73] Tools like Valohai provide audit logs specifically for MLOps platforms.[74]
- *Data Lineage:* Tracking the origin, transformations, and usage of data throughout the AI lifecycle.[18]
- *Model Lineage:* Documenting model versions, training procedures, parameters, and updates.[19]
- *Decision Lineage:* Recording the inputs, processes, and steps leading to a specific AI output or decision.[19]
- *Version Control:* Using systems to track changes in code, models, and datasets.[19]
- *Clear Roles and Responsibilities:* Defining who is accountable for specific aspects of the AI lifecycle (design, testing, deployment, monitoring, incident response).[6]

Frameworks & Standards: Accountability is a central pillar of AI governance frameworks:

- *NIST AI RMF:* The 'Govern' function focuses on establishing organizational structures, policies, roles, and responsibilities to cultivate a culture of risk management and ensure accountability.[5]
- *EU AI Act:* Establishes legal obligations for various actors (providers, deployers, importers, distributors) and imposes significant penalties for non-compliance, creating a strong accountability mechanism.[2] Requires traceability through logging for high-risk systems.[79]
- *ISO/IEC 42001:* The AI Management System standard mandates processes for accountability, risk management, and lifecycle documentation.[80]
- *OECD AI Principles:* Explicitly list accountability as a core value-based principle, requiring traceability and risk management.[10]
- *IEEE Ethically Aligned Design (EAD) & P7000 Standards:* Emphasize accountability through ethical design processes and transparency.[3]
- *Corporate Frameworks:* AWS [27], Microsoft [71], AMD [72], and others include accountability as a key principle.
- *AI Bill of Rights (US):* Includes principles related to notice, explanation, and algorithmic discrimination protections, which support accountability.[14]
- *Future-AI Framework:* Lists traceability as a key principle with operational recommendations.[83]

Measuring and Enhancing Privacy and Data Security

Techniques & Concepts: Protecting personal data and ensuring system security are fundamental to trustworthy AI.

- *Privacy-Enhancing Technologies (PETs):*
  - *Differential Privacy (DP):* A mathematical framework providing quantifiable privacy guarantees by adding calibrated noise to data or algorithm outputs. It ensures that the output is statistically similar whether or not any individual's data is included.[43] Key parameters are epsilon (ϵ, privacy loss budget, lower is more private) and delta (δ, probability of failure).[84] Applied by companies like Apple, Google, Microsoft.[84]
  - *Homomorphic Encryption (HE):* Allows computation directly on encrypted data without decryption.[85] Libraries like Microsoft SEAL implement HE.[53]
  - *Secure Multi-Party Computation (SMC):* Enables multiple parties to jointly compute a function over their inputs while keeping those inputs private.[88]
  - *Federated Learning (FL):* Trains models locally on decentralized devices without centralizing raw data, sharing only model updates.[12]
  - *Anonymization/Pseudonymization:* Removing or replacing personally identifiable information (PII).[85] Tools like Microsoft Presidio aid de-identification.[53]
  - *Data Minimization:* Collecting only necessary data.[89]
- *Security Practices:*
  - *Secure Development Lifecycle (SDLC):* Integrating security throughout AI development, including threat modeling, secure coding, and security testing.[82]
  - *Vulnerability Defense:* Identifying, assessing, and mitigating security weaknesses through measures like firewalls, intrusion detection, patching, vulnerability scanning, and penetration testing.[82]
  - *Resilience:* Designing systems to withstand attacks and maintain function or recover quickly.[6]

Metrics:

- *Differential Privacy:* Epsilon (ϵ) and Delta (δ) quantify privacy loss.[84]
- *Anonymity:* Anonymity Set Size measures how many users are indistinguishable from a target individual.[35]
- *Data Leakage:* Amount of Leaked Information quantifies the extent of a breach (e.g., number of compromised records), though not severity.[35]
- *Security Resilience:* Time until Adversary's Success measures how long a system can resist a specific attack.[35] Standard cybersecurity metrics (e.g., vulnerability

detection rates, time to patch) are also relevant.

Tools & Frameworks:

- *Privacy Libraries:* IBM's Diffprivlib [85], Microsoft SEAL [53], Google's DP libraries, OpenFL for federated learning.
- *Security Tools:* Microsoft Defender for Cloud [53], Microsoft Counterfit [53] for security testing.
- *Governance Frameworks:* NIST RMF lists 'privacy-enhanced' and 'secure and resilient' as trustworthy characteristics.[5] EU AI Act includes requirements for data governance and cybersecurity.[41] IEEE P7002 standard focuses specifically on Data Privacy Processes.[3] OECD Principles cover human rights including privacy.[10] Corporate frameworks (AWS [27], Microsoft [71], AMD [72]) emphasize privacy and security. Privacy-by-Design is a key principle.[82] Compliance with data protection laws like GDPR is essential.[30]

Validating Safety and Reliability

Definitions & Concepts: Safety in AI refers to the absence of conditions that could endanger human life, health, property, or the environment.[5] Reliability means the AI system consistently functions as intended under specified conditions over a given period, without failure.[7] Robustness is the ability of an AI system to maintain its level of performance even under adverse conditions, such as noisy inputs or adversarial attacks.[7] These concepts are crucial for building trust, especially in high-stakes applications.

Approaches & Methods:

- *Robust Testing:* Rigorous evaluation under diverse and challenging conditions, including stress tests, performance benchmarks against varied inputs, and simulation of adverse scenarios.[82]
- *Adversarial Robustness Testing:* Specifically evaluates the system's resilience against inputs intentionally crafted to deceive or cause failure (e.g., small perturbations to images, malicious prompts).[93] Benchmarks like AutoAdvExBench [95] focus on automated exploit generation against defenses. Hardware-level monitoring (e.g., SAMURAI using AI Performance Counters) is proposed for detecting adversarial inputs during inference.[96]
- *Formal Methods:* Mathematical techniques used to specify and verify system properties, potentially offering guarantees of behavior within certain bounds, though often limited by complexity.
- *Fail-Safe Design:* Incorporating mechanisms to ensure the system can be safely overridden, controlled, or shut down if it behaves undesirably or risks causing

harm.[10]

- *Validation and Verification (V&V):* Processes to confirm that the AI system meets its specified requirements (validation) and that it is built correctly according to its design (verification).[5]

Frameworks & Standards:

- *NIST AI RMF:* Requires regular evaluation for safety risks against defined risk tolerances and demonstration that the system can fail safely.[5] Lists 'valid and reliable' and 'safe' as key trustworthy characteristics.[5]
- *EU AI Act:* Mandates high levels of accuracy, robustness, and cybersecurity throughout the lifecycle for high-risk AI systems.[41]
- *IEEE Standards:* P7008 focuses on Fail-Safe Design [68], P7009 addresses Ethically Driven Nudging (related to safe interaction) [68], and P7010 specifies Wellbeing Metrics for Ethical AI/AS.[68] IEEE 7000 includes safety in its ethical design process.[67]
- *OECD AI Principles:* Include a principle dedicated to Robustness, Security, and Safety throughout the lifecycle.[10]
- *Corporate Frameworks:* Reliability and Safety are core principles in frameworks from Microsoft [71], AWS [27], AMD [72], IBM [7], etc.

Assessing Human Oversight Effectiveness

Requirements & Importance: Human oversight is considered essential, particularly for high-risk AI systems, to enable intervention, ensure decisions align with human values and context, prevent errors, and maintain ultimate human accountability.[98] The EU AI Act's Article 14 explicitly mandates effective human oversight measures, requiring that designated personnel can understand the AI system's capabilities and limitations, monitor its operation, and have the ability to intervene or override its decisions.[75]

Evaluation Methods: Assessing whether oversight is truly *effective* is challenging.

- *Checklist-Based Approaches:* Can verify if procedural requirements are met (e.g., personnel received training, documentation exists, awareness of automation bias is promoted). However, these risk being superficial and may not guarantee actual effectiveness in practice.[75]
- *Empirical Testing:* Involves observing human overseers interacting with the AI system in realistic or simulated scenarios. This could include user studies, controlled experiments comparing different interface designs or explainability methods, or monitoring performance in deployment.[75] Such testing aims to measure whether overseers actually detect errors, understand system limitations,

intervene appropriately, and avoid over-reliance (automation bias).[75] This approach requires expertise in HCI and experimental design.[75]

- *Metrics:* Potential metrics include error detection rates by humans, timeliness and appropriateness of interventions, measures of overseer understanding (e.g., through comprehension tests), task completion success under oversight, and subjective measures of trust and confidence.[75] Performance metrics like review turnaround time, alert response speed, and issue resolution timelines can also be tracked.[98]

Challenges: A key difficulty is defining and measuring "effectiveness" in diverse contexts.[75] Automation bias, the tendency to over-trust or uncritically accept AI outputs, poses a significant risk that oversight mechanisms must counteract.[75] Balancing the need for thorough oversight with operational efficiency is another challenge.[98] Effective oversight is a socio-technical design problem, depending on the technology (e.g., quality of explanations), the individual (training, cognitive biases), and the environment (task design, workload, organizational culture).[75]

Evaluating Potential for Misuse

Methodologies: Assessing how an AI system could be intentionally misused requires proactive, adversarial thinking.

- *AI Red Teaming:* Involves dedicated teams simulating attacks by malicious actors to identify vulnerabilities, test limits, and uncover potential misuse scenarios.[101] This goes beyond standard testing by adopting an attacker's perspective and methods.[102] Techniques include data poisoning (corrupting training data) [101], model evasion (crafting inputs to fool the model) [101], model extraction/inversion (stealing the model or sensitive training data) [101], prompt injection (manipulating inputs to bypass safeguards) [101], and generating harmful or biased content.[101] Microsoft's AI Red Team is a prominent example.[53]
- *Threat Modeling:* A systematic process, often part of secure development, to identify potential threats, vulnerabilities, and attack vectors specific to the AI system and its deployment context.[53]
- *Vulnerability Assessment:* Using tools and techniques to scan for known weaknesses in the AI system and its underlying infrastructure.[82]

Tools & Frameworks:

- *Attack Frameworks:* MITRE ATLAS (Adversarial Threat Landscape for AI Systems) and the OWASP AI Security Top 10 provide taxonomies of AI attacks and vulnerabilities.[103]
- *Testing Tools:* Open-source tools like garak [103], Microsoft's PyRIT (Python Risk

Identification Toolkit) [53], Microsoft Counterfit [53], and IBM's Adversarial Robustness Toolbox (ART) [103] automate aspects of adversarial testing.
- *AI Safety Frameworks:* OpenAI's Preparedness Framework (Beta) focuses on evaluating and mitigating catastrophic risks from highly capable models.[104]

Scope & Challenges: Effective red teaming should assess the entire AI system stack, including data pipelines, models, APIs, and user interfaces.[101] Key challenges include the dynamic and adaptive nature of AI systems (vulnerabilities can emerge over time) [101], the lack of transparency in many models and datasets [101], difficulties in measuring and quantifying risk [105], appropriately scoping the exercise [105], and standardizing methodologies to allow for comparison.[105] It's crucial to remember that red teaming identifies existing weaknesses but cannot guarantee the absence of future vulnerabilities or misuse potential.[105]

Adaptation Insights: Drawing from Applied Ethics and Legal Principles

Ethical evaluation of AI often borrows and adapts concepts from established fields:

- **Bioethics Principles:** The principles of beneficence (do good), non-maleficence (do no harm), autonomy (respect for self-determination), and justice (fairness) are frequently adapted from medical ethics to form the basis of AI ethics frameworks.[28] AI-specific principles like explicability or transparency are often added to address the unique nature of the technology.[106]
- **Capability Approach:** Originating in philosophy and development economics (Amartya Sen, Martha Nussbaum), this approach evaluates well-being and justice based on individuals' substantive freedoms or 'capabilities' (what they are actually able to do and be). Applying this to AI ethics shifts the focus from abstract principles or resource distribution to how AI systems actually enable or hinder people's ability to achieve valued functionings, considering the necessary 'conversion factors' (personal, social, environmental conditions) needed to benefit from the technology.[108] This provides a lens for assessing equity and real-world impact.
- **Rights-Based Frameworks:** Methodologies like Fundamental Rights Impact Assessment (FRIA) or Human Rights Impact Assessment (HRIA) adapt legal and international human rights standards to systematically identify and assess the potential risks AI systems pose to fundamental rights.[109] This approach aligns evaluation directly with established legal and ethical norms, particularly relevant for compliance with regulations like the EU AI Act.[109]
- **Classical Ethical Theories:** Deontology (duty-based ethics), utilitarianism (consequence-based ethics), and virtue ethics (character-based ethics) provide underlying philosophical frameworks that inform the formulation of AI principles

and can guide the analysis of ethical dilemmas and trade-offs in AI evaluation.[110]

The landscape of ethical AI is marked by a tension between the proliferation of numerous, often high-level, principle sets and the drive towards concrete standardization. Many organizations, from intergovernmental bodies like the OECD [10] and UNESCO [111] to national governments [25], industry consortia [27], corporations [71], and academic groups [106], have proposed ethical guidelines. While significant convergence exists around core themes like fairness, transparency, accountability, privacy, and safety [30], the sheer number and subtle variations can lead to confusion, redundancy, or strategic selection ("ethics shopping").[107] Concurrently, efforts are underway to establish formal standards (e.g., ISO/IEC 42001 [80], IEEE P7000 [3]), regulatory requirements (EU AI Act [65]), and common metrics and benchmarks [4] to provide concrete, verifiable, and comparable ways to assess ethical performance. This creates a dynamic where evaluation must navigate between context-specific ethical nuances and the need for standardized, enforceable rules. A purely standards-based assessment might miss contextual ethical issues, while a purely principle-based one lacks rigor and comparability.

Furthermore, a significant gap often exists between the articulation of ethical principles and their actual implementation in AI systems.[11] The abundance of high-level guidelines contrasts with a relative scarcity of validated, practical methods for assessing adherence.[11] This gap creates a risk of "ethics washing," where organizations publicly commit to ethical principles without enacting substantive changes in their development or deployment practices.[16] The lack of standardized implementation methods and verifiable metrics exacerbates this problem.[16] Consequently, effective ethical evaluation must move beyond simply checking for the existence of principles or policies. It requires assessing demonstrable actions, documented processes (such as those mandated by ISO 42001 [81] or required for EU AI Act conformity [41]), and measurable outcomes, such as quantified fairness metrics [33], differential privacy guarantees [84], or results from robust safety testing.[95] Evaluation frameworks and potential ranking systems should prioritize evidence of practice over mere statements of intent.

Table 1: Key Ethical AI Dimensions, Metrics, and Tools

Ethical Dimension	Key Concepts/Definitions	Example Metrics/Measurement Approaches (Quantitative & Qualitative)	Example Tools/Standards/Frameworks	Key Challenges
Fairness & Bias Mitigation	Equitable treatment/outcomes across groups; addressing systemic, computational, cognitive bias.[5] Group vs. Individual fairness.[32]	*Quantitative:* Demographic Parity, Equal Opportunity, Equalized Odds (Differences/Ratios) [33], Consistency Score [33], GEI.[33] *Qualitative:* Bias audits, impact assessments on specific groups.[18]	AIF360 [17], Fairlearn [38], What-If Tool [37], NIST SP 1270 [5], EU AI Act [2], ISO 42001.[81]	Context-dependency, metric conflicts (impossibility) [31], data representativeness, bias amplification.[33]
Transparency & Explainability	Availability of info about system (Transparency) [5]; understanding internal mechanisms (Explainability) [44]; understanding output meaning (Interpretability).[5]	*Quantitative:* Metrics from explainability tools (e.g., feature importance scores, stability metrics [46]). *Qualitative:* User studies on explanation effectiveness, clarity of documentation.	LIME [45], SHAP [50], InterpretML [53], Model Cards [20], Datasheets for Datasets [59], NIST RMF [8], EU AI Act [65], IEEE 7001.[3]	Trade-offs (e.g., with accuracy/privacy) [113], instability of explanations [47], evaluating explanation quality.
Accountability & Responsibility	Responsibility for system functioning/outcomes [10];	*Quantitative:* Audit log completeness/coverage.	Audit trails/logging tools (e.g., Valohai [74]),	Complexity of tracing decisions in opaque models,

	traceability of data, models, decisions.[19]	*Qualitative:* Clarity of defined roles/responsibilities, effectiveness of governance processes.	Data/Model Lineage tools, NIST RMF (Govern) [6], EU AI Act [77], ISO 42001 [81], OECD Principles.[10]	defining accountability in multi-actor systems.[114]
Privacy & Data Security	Protecting personal/sensitive info [7]; data subject control [30]; confidentiality, integrity, availability.[12]	*Quantitative:* Differential Privacy (ϵ,δ) [84], Anonymity Set Size [35], Amount Leaked Info [35], Security metrics (vulnerability counts, patch times). *Qualitative:* Privacy Impact Assessments (PIAs) [30], security audits, adherence to Privacy-by-Design.[89]	DP libraries (Diffprivlib [85]), HE libraries (SEAL [53]), Anonymization tools (Presidio [53]), NIST RMF [6], EU AI Act [41], IEEE P7002 [3], GDPR.[42]	Privacy-utility trade-offs [88], complexity of PET implementation, evolving security threats.
Safety & Reliability	Avoiding harm to humans/environment [5]; functioning as intended without failure [7]; performing well under adverse conditions (Robustness).[10]	*Quantitative:* Accuracy/error rates under stress/adversarial tests, failure rates, mean time between failures (MTBF), benchmark scores (e.g., AutoAdvExBench [95]). *Qualitative:* Safety case documentation, fail-safe	Robustness testing tools (ART [103]), Formal methods tools, NIST RMF [5], EU AI Act [41], IEEE P7008/P7010 [68], OECD Principles.[10]	Proving safety for complex systems, predicting emergent failures, testing against unknown unknowns.

		mechanism validation.		
Human Oversight	Ability for humans to monitor, intervene, override AI [98]; understanding capabilities/limitations.[75]	*Quantitative:* Error detection rates, intervention timeliness/accuracy, task success under oversight.[98] *Qualitative:* User studies on understanding/trust, checklist compliance.[75]	EU AI Act (Art. 14) [75], HCI/Usability testing methods.[75]	Defining/measuring "effectiveness" [75], automation bias [75], balancing oversight vs. efficiency.[98]
Potential for Misuse	Assessing vulnerability to malicious use (harmful content, data poisoning, extraction, prompt injection).[101]	*Quantitative:* Success rates of simulated attacks (red teaming), vulnerability scan results. *Qualitative:* Threat modeling reports, red team findings documentation.	AI Red Teaming methodologies [101], MITRE ATLAS [103], OWASP AI Top 10 [103], Testing tools (garak, PyRIT, Counterfit, ART).[103]	Dynamic nature of threats, difficulty predicting novel misuses, scoping challenges.[105]

III. Quantifying the Environmental Footprint of AI

The development and deployment of AI systems, particularly large-scale models, consume significant computational resources, leading to concerns about their environmental impact. Evaluating this footprint requires measuring energy consumption, associated carbon emissions, and the utilization of other resources like water and hardware materials.

Measuring Energy Consumption and Carbon Emissions

Importance: The training and inference phases of AI models, especially deep learning algorithms, demand substantial energy, often powered by energy-intensive hardware like GPUs and TPUs.[115] This energy consumption translates directly into greenhouse gas (GHG) emissions, contributing to climate change.[21] Data centers, which house the necessary infrastructure, are themselves major energy consumers, with projections suggesting their global electricity demand could reach 1,000 TWh by 2026.[116] Training large models like GPT-3 has been estimated to consume over 1,200 MWh.[115]

Metrics: Several metrics are used to quantify energy use and carbon impact:

- *Energy Consumption:* Measured in kilowatt-hours (kWh) or megawatt-hours (MWh), representing the total electricity used by the hardware during AI tasks (training, inference).[21]
- *Power Usage Effectiveness (PUE):* A standard data center efficiency metric, calculated as Total Facility Energy divided by IT Equipment Energy. A PUE closer to 1 indicates higher efficiency, meaning less energy is used for cooling and other overheads relative to the energy powering the IT equipment itself.[22] Average PUE was reported around 1.58 in 2020 [121], though leading providers aim for much lower figures.[125]
- *Carbon Emissions:* Measured in kilograms or metric tons of CO2 equivalent (kgCO2e, tCO2e). Calculated by multiplying energy consumption (kWh) by the carbon intensity of the electricity grid providing the power.[21] CO2=Energy Consumption (kWh)×Carbon Intensity (kgCO2e/kWh).[117]
- *Carbon Intensity:* Represents the GHG emissions per unit of electricity generated (kgCO2e/kWh). This varies significantly based on the energy mix (fossil fuels vs. renewables) of the specific geographic region and time.[115]
- *Compute Carbon Intensity (CCI):* A metric proposed for AI hardware, measuring lifetime carbon emissions (including manufacturing) per unit of computation (e.g., grams of CO2e per Exa-FLOP).[128] Lower CCI indicates greater carbon efficiency for a given computational task.
- *Efficiency Scores:* Combining performance with environmental cost, e.g., model

accuracy per gram of CO2 emitted.[117]

Tools: Various tools aim to estimate or measure these impacts:

- *CodeCarbon:* A Python package that estimates CO2 emissions by tracking the power consumption of CPU, GPU, and RAM during code execution and multiplying by the location-specific carbon intensity (obtained from public data or APIs like ElectricityMap).[21] It provides dashboards for visualization and recommends lower-carbon compute regions.[126]
- *ML CO2 Impact:* A web-based tool that *estimates* carbon footprint based on user inputs: hardware type, usage duration, cloud provider, and region.[21] Useful for post-hoc estimation but relies on estimated power consumption figures, which can be inaccurate.[120]
- *Eco2AI:* Another Python library for tracking CO2 emissions, using regional coefficients derived from global energy reports.[127]
- *Cloud Provider Tools:* Major cloud providers (AWS, Azure, GCP) offer dashboards or reports that provide some information on energy consumption and carbon footprint associated with cloud usage, though transparency and granularity vary.[21]
- *Hardware Monitoring Tools:* Tools like NVIDIA SMI (for GPUs) and Intel Power Gadget (for CPUs) can provide direct power usage measurements on some systems.[134]
- *AIWattch:* An example of a specialized tool; a browser extension estimating LLM carbon emissions during chat interactions.[21]

Benchmarks & Research: The field of "Green AI" focuses on measuring and mitigating the environmental impact of AI.[23] Benchmarks like MLPerf are incorporating energy efficiency alongside performance.[15] Academic studies quantify the significant footprints of training large models like GPT-3, BLOOM, Gopher, and OPT.[115] Research explores optimizing training (e.g., hyperparameter tuning, early stopping) and deployment for lower impact.[134]

Assessing Resource Utilization

Beyond energy and carbon, AI systems utilize other critical resources.

Water Usage: Data centers consume substantial amounts of freshwater, primarily for cooling IT equipment.[119] This consumption occurs both on-site (Scope 1, e.g., through evaporative cooling towers) and off-site (Scope 2, water used in generating the electricity consumed by the data center).[123] Estimates suggest training GPT-3 required 700,000 liters of water [119], and global AI water withdrawal could reach billions of cubic meters annually by 2027.[119] Even inference consumes water; estimates suggest 20-50 ChatGPT queries use roughly 500ml.[123]

Water Metrics:

- *Water Usage Effectiveness (WUE):* The primary metric for data center water efficiency, calculated as annual water consumption (liters) for cooling and humidification divided by the total annual IT equipment energy consumption (kWh).[22] A lower WUE indicates better water efficiency. WUE = Liters / kWh.[122]
- *WUE Values:* Vary significantly based on cooling technology (air cooling ≈ 0 WUE, evaporative cooling can be >2.0 WUE) and local climate (humidity, temperature).[22] Microsoft reports WUE values ranging from 0.01 L/kWh (Singapore) to 1.63 L/kWh (Arizona).[124]
- *Water Consumption Intensity Factor (WCIF):* Measures water used per unit of electricity generated off-site.[123]
- *Operational Water Footprint:* Calculated considering energy consumption, the fraction cooled by on-site vs. off-site water, and the respective WUE/WCIF values.[123] Woperational=Σtet×(θ1×WUEonsite+θ2×WCIFoffsite).[123]
- *Embodied Water Footprint:* Water consumed during hardware manufacturing and supply chain processes.[123]

Hardware Sustainability: The demand for powerful processors (GPUs, TPUs) and large storage systems for AI drives the consumption of raw materials, including rare earth metals, and contributes significantly to the growing problem of electronic waste (e-waste).[23] E-waste is one of the fastest-growing waste streams globally, containing toxic substances and representing a loss of valuable resources.[143]

Circularity Metrics: Evaluating hardware sustainability involves assessing circular economy principles – keeping materials and products in use for longer.

- *Material Footprint:* Total quantity of materials used in a product's lifecycle.[146]
- *Recycling Rate:* Percentage of materials recovered from end-of-life products.[146]
- *Reuse/Refurbishment Rate:* Percentage of components or devices given a second life.[143]
- *Product Lifespan Extension:* Measuring the increase in useful life due to design for durability or predictive maintenance.[143]
- *Sorting Accuracy/Throughput:* Efficiency of identifying and separating materials in e-waste processing, often enhanced by AI vision systems.[143]
- *Precious Metal Recovery Rate:* Percentage of valuable metals successfully extracted.[145] While AI can exacerbate e-waste through rapid hardware obsolescence and resource intensification [142], it also holds potential to *improve*

circularity through optimized design, predictive maintenance, automated sorting/dismantling, and supply chain transparency.[142]

Applying Life Cycle Assessment (LCA) to AI Systems

Methodology: Life Cycle Assessment (LCA) provides a comprehensive and systematic framework for evaluating the environmental impacts of a product, process, or service throughout its entire lifespan.[118] This includes stages such as raw material extraction, manufacturing, transportation, use (operation), and end-of-life management (recycling, disposal).[118] LCA aims to quantify impacts like energy consumption, GHG emissions, water use, resource depletion, and waste generation across all these stages.[118]

Application to AI: LCA can be applied to assess the full environmental footprint of both AI software (models) [118] and the specialized hardware they run on (e.g., GPUs, TPUs, servers, data centers).[116] By analyzing the entire lifecycle, LCA moves beyond just operational impacts (like energy use during training/inference) to include *embodied* impacts associated with manufacturing and disposal.[118] This holistic view is crucial for identifying the most significant impact stages (e.g., Google's TPU LCA found operation dominated lifetime emissions, but manufacturing was still notable [128]) and for targeting mitigation efforts effectively.[118]

Metrics within LCA: LCA studies typically quantify various environmental indicators, including:

- Carbon Footprint (GHG emissions, often broken down by scope or lifecycle stage).[118]
- Energy Consumption (cumulative, across stages).[118]
- Water Consumption (direct and indirect, across stages).[118]
- Resource Depletion (e.g., use of specific minerals).[118]
- Waste Generation (including e-waste).[118]

Standards & Challenges: International standards like ISO 14040 and ISO 14044 provide general methodologies for conducting LCAs.[129] However, applying LCA to complex, rapidly evolving AI systems and hardware presents challenges. It requires detailed data from across the supply chain (e.g., component manufacturing), which can be difficult to obtain due to proprietary concerns or lack of transparency.[129] Specialized expertise is needed to conduct rigorous LCAs.[130] Interestingly, AI itself shows potential for enhancing LCA processes through data analysis, predictive modeling, and automating tasks like emission factor matching.[150]

Adaptation Insights: Leveraging Environmental Science Methodologies

The evaluation of AI's environmental impact directly leverages established methodologies from environmental science and engineering:

- **LCA:** As discussed, LCA is a cornerstone technique in environmental assessment, directly applicable to quantifying the cradle-to-grave impacts of AI systems and infrastructure.[149]
- **Footprinting:** Carbon footprinting [120], water footprinting [123], and material footprinting [146] are standard methods used to quantify specific resource consumption or emission types, readily adaptable to AI.
- **Material Flow Analysis (MFA):** Used to track the flows of materials through industrial systems, relevant for assessing resource efficiency and circularity in AI hardware production and e-waste management.
- **Ecological Impact Assessment:** Broader assessment methods considering impacts beyond resource use and emissions (e.g., biodiversity loss from mining, land use change for data centers) could be adapted from environmental impact assessment (EIA) practices.[23]

The current focus on operational energy and carbon metrics, while important, provides an incomplete picture of AI's environmental burden. Tools like CodeCarbon measure energy use during computation [120], but LCA studies reveal that embodied emissions from manufacturing hardware (Scope 3) are substantial and will constitute a growing share of the total footprint as operational efficiency improves and electricity grids become cleaner.[116] Similarly, the significant water consumption associated with both data center cooling and electricity generation is often underreported compared to carbon emissions.[119] The rapid hardware upgrade cycles driven by AI advancements also exacerbate the e-waste problem.[142] Therefore, relying solely on operational carbon or energy metrics for evaluating or ranking AI systems is insufficient and potentially misleading. A comprehensive assessment necessitates adopting lifecycle perspectives, incorporating metrics for embodied impacts (e.g., using LCA or metrics like CCI [128]), water footprint (e.g., WUE [22]), and e-waste/circularity indicators.[145] Standardization efforts must evolve to encompass these broader lifecycle impacts.[137]

Furthermore, optimizing AI's environmental performance involves navigating complex trade-offs within the energy-water-carbon nexus. Efforts to improve energy efficiency (lower PUE) in data centers, such as using evaporative cooling, often increase direct water consumption (higher WUE).[22] Conversely, minimizing water use through air cooling typically increases energy consumption (higher PUE).[22] The ultimate carbon impact of these choices depends heavily on the carbon intensity of the local electricity grid; saving energy has a greater carbon benefit on a high-carbon grid than on a low-carbon one.[22] Likewise, the significance of water consumption depends on local water availability and stress levels.[22] Optimizing for a single metric like PUE or even operational carbon footprint can lead to unintended negative consequences for water resources or overall environmental impact in specific contexts. Thus, environmental evaluation must consider PUE, WUE, and carbon intensity holistically.[22] Comparative rankings based solely on one dimension risk promoting environmentally suboptimal choices in certain regions. Meaningful assessment requires context-awareness regarding local climate, grid characteristics, and water availability.

Table 2: Key Environmental Impact Metrics and Tools for AI

Impact Area	Specific Metrics	Measurement Tools/Methods	Key Considerations/Challenges
Energy Consumption	kWh, MWh (Operational & Lifecycle) [118]	CodeCarbon [126], Eco2AI [133], ML CO2 Impact (Estimate) [120], Cloud Dashboards [21], Hardware Monitors [134], LCA [118]	Accuracy of estimates vs. measurements, Scope (operation vs. lifecycle), Transparency from providers.[134]
Data Center Efficiency	PUE (Power Usage Effectiveness) [22]	Data center monitoring systems, Provider reports [125]	Comparability issues across facilities [121], Measurement point consistency [121], Trade-off with WUE.[22]
Carbon Emissions	kgCO2e, tCO2e (Operational & Embodied) [118]	CodeCarbon [126], Eco2AI [133], ML CO2 Impact [120], LCA [118], Carbon Intensity Data (e.g., ElectricityMap [134])	Requires accurate energy data & carbon intensity factors [117], Importance of lifecycle (embodied) emissions.[128]
Hardware Carbon Efficiency	CCI (Compute Carbon Intensity) [gCO2e/Exa-FLOP][128]	LCA of hardware components [130]	Data availability from supply chain [129], Standardization of metric.
Water Consumption	Liters, Cubic Meters (Operational & Embodied) [119]	Data center monitoring, Provider reports [125], LCA [118], WCIF data [123]	Often underreported [119], Scope (on-site vs. off-site vs. embodied) [123], Local water stress context.
Data Center Water Efficiency	WUE (Water Usage Effectiveness)	Data center monitoring systems,	Trade-off with PUE [22], Climate dependency

	[L/kWh][22]	Provider reports [124]	[22], Reporting consistency.[22]
Hardware Sustainability / E-waste	Material Footprint [146], Recycling Rate [146], Reuse/Refurbishment Rate [146], Lifespan Extension [143], Sorting Accuracy/Throughput [145], Precious Metal Recovery [145]	Waste tracking, Material Flow Analysis, Product lifecycle tracking, E-waste facility monitoring	Data collection across supply chain, Standardization of circularity metrics [145], Complexity of e-waste streams.

IV. Evaluating Accessibility and Equity in AI Systems

Ensuring that AI systems are accessible to and benefit all members of society, regardless of ability, background, language, or socioeconomic status, is a critical aspect of responsible AI development. This requires evaluating ease of use, affordability, language support, and the potential to exacerbate the digital divide.

Ensuring Ease of Use for Diverse Populations (including disabilities)

Importance: AI systems, whether manifesting as chatbots, recommendation engines, or complex analytical tools, must be designed with inclusivity at their core. Just as websites and software need to be accessible, AI interfaces must be usable by individuals with a wide range of abilities, including those with visual, auditory, motor, cognitive, or speech impairments.[152] While AI itself holds promise for creating powerful assistive technologies (e.g., automated image description for blind users, real-time captioning for deaf users, augmentative communication aids) [155], poorly designed AI can erect new barriers or perpetuate existing digital exclusion.[153] Inclusivity issues arise when training data lacks diversity, leading AI systems like speech recognition to perform poorly for users with atypical speech patterns [157] or computer vision systems trained on images from sighted users failing for blind users.[157]

Guidelines & Standards: The foundational standards for digital accessibility, primarily the Web Content Accessibility Guidelines (WCAG) developed by the W3C, provide a robust starting point.[152] WCAG is built on four principles (Perceivable, Operable, Understandable, Robust - POUR) and defines testable success criteria at three conformance levels (A, AA, AAA).[152] These principles and criteria can be adapted to evaluate AI interfaces.[161] For example, 'Perceivable' might involve ensuring AI outputs can be presented in multiple formats (text, audio); 'Operable' could relate to keyboard accessibility for interacting with an AI agent; 'Understandable' might concern the clarity and predictability of AI responses; and 'Robust' involves compatibility with assistive technologies.[153] The upcoming WCAG 3.0 aims for greater flexibility and a focus on user outcomes, which may be well-suited for evaluating dynamic AI systems.[162] Legal mandates like Section 508 in the US (for federal technology) and the Americans with Disabilities Act (ADA) also require accessibility.[162]

Adaptation for AI: Applying accessibility principles to AI requires specific considerations:

- *Content Adaptability:* AI should ideally adjust its output format, complexity, or presentation based on user needs or preferences.[161]
- *Multimodal Interaction:* Supporting diverse input methods (keyboard, voice,

switch access) and output modalities (text, speech synthesis, visual representations).[153]

- *Contextual Awareness & Understandability:* AI should interpret user intent effectively and provide clear, predictable, and easily comprehensible responses, adapting language complexity as needed.[161]
- *Assistive Technology Compatibility:* Ensuring AI interfaces and outputs can be reliably interpreted by screen readers, braille displays, magnifiers, etc..[153]
- *Time Sufficiency:* Allowing users, particularly those with motor or cognitive impairments, adequate time to interact.[161] AI itself can play a role in *automating* accessibility checks and even suggesting or implementing remediations for digital content or interfaces.[155]

Evaluation Methods: Assessing AI accessibility requires a combination of methods:

- *Usability Testing with People with Disabilities (PWD):* This is considered the most valuable technique as it reveals real-world barriers.[154] It involves recruiting representative users with various disabilities and observing them interacting with the AI system, often using their own assistive technologies.[163] Standard usability testing protocols need adaptation (e.g., focusing on accessibility errors over speed, allowing more time, ensuring accessible testing environments).[154] Data collected includes task success, error types (especially accessibility-related), and qualitative feedback.[163]
- *Automated Testing Tools:* Software tools can scan interfaces or code for conformance with technical WCAG criteria (e.g., color contrast, presence of alt text).[152] They are efficient for catching certain issues but cannot evaluate aspects requiring human judgment, like keyboard navigation logic or screen reader experience.[152]
- *Manual Expert Review:* Accessibility experts conduct manual checks against guidelines (like WCAG) and test with various assistive technologies.[152]
- *Accessibility Audits:* Typically combine automated scanning, manual expert review, and often usability testing with PWD for a comprehensive assessment.[152]
- *Human-Computer Interaction (HCI) Methods:* Techniques like heuristic evaluation, cognitive walkthroughs, and participatory design can be adapted to focus on accessibility.[153]

Metrics:

- *Conformance Levels:* Pass/fail status against WCAG success criteria (Level A, AA, AAA).[152]
- *Usability Metrics (adapted):* Task success rates, frequency and severity of accessibility errors encountered, time on task (used cautiously), user satisfaction

scores.[163]

- *AI-Specific Metrics (proposed):* Adaptation Rate (how well AI adjusts output), Comprehension Score (user understanding of AI responses), Interaction Efficiency (task completion effort), Error Resolution Rate (AI correcting misunderstandings).[161]

Analyzing Cost and Affordability

Importance: The economic aspects of AI significantly influence its accessibility and equitable distribution. High costs associated with developing, deploying, accessing, or using AI systems can create or widen the digital divide, limiting benefits to affluent individuals, large corporations, or developed nations.[24] Conversely, AI has the potential to *reduce* costs and improve affordability in various sectors, such as optimizing construction processes to lower housing costs [167] or improving efficiency in healthcare delivery.[169] However, the substantial investments required for AI development (data, hardware, expertise) and operation (compute resources) remain a significant factor.[170]

Metrics & Analysis: Evaluating affordability involves analyzing various cost components and economic impacts:

- *Total Cost of Ownership (TCO):* A comprehensive assessment including initial development/acquisition costs, data collection/preparation/labeling costs, hardware and infrastructure expenses (GPUs, cloud services), integration costs, ongoing maintenance and update costs (including model retraining), compute costs for inference, and compliance costs.[170]
- *Return on Investment (ROI):* Comparing the financial benefits derived from AI (e.g., increased productivity, cost savings, new revenue streams) against the TCO.[170] Studies suggest significant potential economic gains from AI, boosting GDP and productivity.[1]
- *Pricing Models:* Analyzing how AI services are priced affects user affordability. Common models include usage-based pricing (e.g., cost per token for LLMs, influenced by context window size), tiered subscriptions, and enterprise licenses.[170]
- *AI for Pricing Optimization:* AI itself can be used to create predictive pricing models that dynamically adjust prices based on market conditions, demand, and potentially patient adherence or affordability factors, aiming to improve access.[168]
- *Economic Impact Assessment:* Broader studies examining AI's effect on market structures, competition, and overall economic indicators like GDP.[173]

Accessibility Factors: Cost is a direct barrier to equitable access. High development costs may limit innovation to well-funded entities.[171] High usage costs can exclude individuals or smaller organizations.[170] Strategies to improve affordability include offering free or 'freemium' tiers [173], developing open-source AI tools and models [37], and optimizing models for efficiency to reduce compute costs.[171]

Assessing Multilingual Capabilities and Language Support

Importance: For AI, particularly Large Language Models (LLMs), to be globally equitable and useful, they must effectively support a wide range of human languages [Query IIId]. Currently, many models are trained predominantly on data from high-resource languages like English, which can lead to poorer performance, biases, and lack of cultural nuance when applied to other languages, further marginalizing non-dominant language communities.[24]

Evaluation Methods: Assessing language support involves evaluating LLM performance on various NLP tasks across different languages.

- *Benchmarking:* Using standardized multilingual datasets and tasks. Examples include subsets of GLUE (General Language Understanding Evaluation) [178], XTREME, XNLI, TyDi QA, etc. These benchmarks test capabilities like cross-lingual classification, question answering, and information retrieval.
- *Task-Specific Evaluation:* Testing performance on core NLP tasks within specific target languages:
    - *Machine Translation:* Evaluating the quality of translation between language pairs.
    - *Text Summarization:* Assessing the ability to summarize texts accurately in different languages.
    - *Question Answering:* Measuring accuracy in answering questions posed in various languages.
    - *Text Generation:* Evaluating the fluency, coherence, and relevance of text generated in target languages.
    - *Named Entity Recognition (NER):* Assessing the ability to identify entities (people, places, organizations) in multilingual text.

Metrics: Standard NLP metrics are adapted and applied, often comparing model output to human-created reference texts or labels.

- *BLEU (Bilingual Evaluation Understudy):* Widely used for machine translation. Compares n-gram overlap between machine translation and reference translations. Score ranges from 0 to 1 (or 0 to 100), higher is better.[35]
- *ROUGE (Recall-Oriented Understudy for Gisting Evaluation):* Primarily used for summarization, but also translation. Measures overlap based on n-grams (ROUGE-N), longest common subsequences (ROUGE-L), or skip-bigrams

(ROUGE-S).[54] Higher scores are better.
- *METEOR (Metric for Evaluation of Translation with Explicit ORdering):* A translation metric considering precision, recall, stemming, synonymy, and word order alignment.[54] Generally correlates better with human judgment than BLEU.
- *Perplexity (PPL):* Measures how well a language model predicts a sequence of text in a given language. Lower perplexity indicates better fluency and predictability.[178]
- *Cross-Entropy Loss:* Related to perplexity, measures the difference between the model's predicted probability distribution and the actual distribution of words/tokens.[179] Lower is better.
- *F1 Score, Precision, Recall:* Standard classification metrics used for tasks like NER, applied per language.[178]
- *Exact Match (EM):* For question answering, measures the percentage of predictions that exactly match the ground truth answer.[182]
- *Human Evaluation:* Crucial for assessing nuances like fluency, adequacy, cultural appropriateness, and overall quality that automated metrics often miss.[182] Involves human judges rating or comparing outputs.

Addressing the Digital Divide

Definition & Context: The digital divide refers to the gap between individuals, communities, and geographic areas regarding access to, use of, and knowledge of information and communication technologies (ICTs), including the internet and increasingly, AI.[24] This divide is not just about access to infrastructure but also encompasses skills (digital literacy), affordability, and meaningful use.[24] It often aligns with existing societal inequalities based on socioeconomic status, education, race, gender, age, disability, and geography (urban vs. rural).[184] AI has the potential to both exacerbate this divide (e.g., through biased algorithms trained on data from dominant groups, high access costs, automation displacing low-skilled workers) and potentially bridge it (e.g., providing personalized education or health information to underserved areas, assistive technologies).[24] Billions globally still lack meaningful internet connectivity, a prerequisite for most AI benefits.[24]

Metrics & Indicators: Measuring the digital divide in the context of AI requires looking beyond simple internet access:

- *Access Metrics:* Internet penetration rates, availability and quality of broadband infrastructure (speed, reliability - 'meaningful connectivity'), device ownership (smartphones, computers) across different demographic groups and regions.[24]
- *Skills/Literacy Metrics:* Digital literacy levels, AI literacy (understanding how AI works, its capabilities and limitations), ability to effectively use AI tools for specific

goals (e.g., education, job searching).[24]

- *Usage Metrics:* Frequency and patterns of internet and AI tool usage, types of applications used (consumption vs. creation/innovation) across different groups.[184]
- *Affordability Metrics:* Cost of devices, internet access, and AI services relative to income levels.
- *Outcome Metrics:* Assessing the differential impact of AI access (or lack thereof) on key life outcomes such as educational attainment, employment opportunities, health status, income levels, and civic participation across different population segments.[184]
- *Inequality Indices:* Applying standard inequality measures (e.g., Gini coefficient variations) to digital access, skills, or outcome metrics to quantify the divide between groups.[186]

Evaluation Methodologies: Understanding the digital divide requires mixed methods:

- *Quantitative Analysis:* Using large-scale surveys, census data, and national/international statistics (e.g., from ITU [24], OECD) to measure access, usage, and skills gaps across demographics.[185] Econometric modeling can explore determinants of the divide.[186]
- *Qualitative Research:* Employing methods like interviews, focus groups, and ethnography to understand the lived experiences, barriers, and needs of marginalized or digitally excluded communities.[184] This provides context and depth beyond quantitative data.
- *Impact Assessment:* Utilizing frameworks like Social Impact Assessment (SIA) [151] or specific Digital Impact Assessment [189] to evaluate how the introduction or lack of AI technology differentially affects various communities.
- *Technology Acceptance Models (TAM):* Frameworks like TAM can be used to study factors influencing AI adoption and use among specific groups, such as older adults, identifying barriers like perceived usefulness or ease of use.[184]
- *Comparative Studies:* Analyzing the digital divide across different regions, countries, or policy environments.[186]

Adaptation Insights: Integrating methods from HCI, Accessibility Studies, and Economics

Evaluating AI accessibility and equity naturally draws upon methodologies from several related fields:

- **Human-Computer Interaction (HCI):** Provides core methods for understanding user needs and evaluating usability, such as user-centered design, participatory

design involving diverse users, usability testing, heuristic evaluation, and cognitive walkthroughs. These are directly applicable to designing and testing AI interfaces for ease of use and positive user experience.[153]
- **Accessibility Studies:** Offers established standards (WCAG [152]), specialized testing methodologies (accessibility audits, usability testing with PWDs [163]), and design principles focused specifically on ensuring technology is usable by people with disabilities.[165]
- **Economics:** Contributes tools for analyzing costs (TCO [172]), benefits (ROI [173]), market impacts, pricing strategies [170], and distributional effects, including measuring inequality and assessing affordability barriers.[169]
- **Sociology and Development Studies:** Provide theoretical frameworks (e.g., on social inequality, diffusion of innovations) and methods (surveys, qualitative analysis) for understanding and measuring the digital divide and its societal consequences.[187]

A crucial observation arising from synthesizing these areas is that technical accessibility serves as a foundational layer for achieving broader digital equity. While the digital divide encompasses barriers like cost, connectivity, and digital literacy [24], these factors become relevant only if the technology itself is fundamentally usable by diverse individuals, including those with disabilities. An affordable AI service is inequitable if its interface is inaccessible to a blind user relying on a screen reader.[153] Conversely, a perfectly accessible interface provides no benefit if potential users lack the internet connection, device, or funds to access it.[24] Therefore, ensuring basic accessibility, evaluated through methods like WCAG compliance audits and usability testing with PWDs [152], is a necessary prerequisite for any meaningful progress towards equitable AI deployment. Evaluation frameworks must integrate both aspects, and ranking systems should arguably treat fundamental accessibility failures as critical flaws undermining equity.

Another important consideration emerges from AI's capacity for personalization. AI can tailor interfaces [155], learning materials [192], and content delivery [170] to individual user needs and preferences, potentially enhancing usability, engagement, and accessibility for some. However, this personalization relies on collecting and analyzing user data, which inherently raises privacy concerns.[194] Furthermore, if the algorithms driving personalization are biased or trained primarily on majority user data, they might fail to adapt appropriately for minority groups or individuals with disabilities, potentially creating *less* equitable or accessible experiences for them.[157] There exists a tension between the goal of universal design and accessibility, often pursued through standardization (like WCAG), and the individualized experiences created by hyper-personalization. Evaluating personalized AI systems thus requires assessing not just the benefits for the intended user but also the potential negative impacts on privacy and equity for other groups. Metrics need to capture the range and diversity of users effectively served, and frameworks must balance the advantages of adaptation against the need for baseline universal accessibility standards and robust privacy safeguards.

Table 3: Overview of Common NLP Metrics for Language Support Evaluation

Metric Name	Brief Description	Typical Use Case	Interpretation	Key Limitations
BLEU (Bilingual Evaluation Understudy)	Measures n-gram precision overlap between candidate and reference text(s), with a brevity penalty.[180]	Machine Translation	Higher is better (0-1 or 0-100)	Correlates weakly with human judgment of fluency/adequacy; penalizes lexical diversity; requires reference translations.
ROUGE (Recall-Oriented Understudy for Gisting Evaluation)	Measures overlap based on n-grams (ROUGE-N), longest common subsequence (ROUGE-L), or skip-bigrams (ROUGE-S).[180]	Text Summarization, Machine Translation	Higher is better (0-1)	Focuses on recall; may not capture coherence or factuality; requires reference summaries/translations.
METEOR (Metric for Evaluation of Translation with Explicit ORdering)	Considers unigram precision/recall, stemming, synonymy, and chunking for alignment.[54]	Machine Translation	Higher is better (0-1)	More complex than BLEU; requires language-specific resources (stemmers, synonyms).
Perplexity (PPL)	Measures how well a probability model predicts a sample text sequence (exponential of negative log-likelihood).[17]	Language Modeling, Text Generation (Fluency)	Lower is better (≥1)	Sensitive to vocabulary size and tokenization; doesn't directly measure task performance or factual accuracy.

	9			
Cross-Entropy Loss	Measures the difference between the predicted probability distribution and the actual distribution of tokens.[179]	Language Model Training/Evaluation	Lower is better (≥0)	Similar limitations to Perplexity; primarily a training objective.
F1 Score / Precision / Recall	Standard classification metrics applied at token or entity level.[178]	Named Entity Recognition (NER), Slot Filling, Classification Tasks	Higher is better (0-1)	Task-specific; don't capture overall language quality.
Exact Match (EM)	Percentage of predictions that exactly match the ground truth answer string.[182]	Question Answering (Extractive)	Higher is better (0-1)	Very strict; penalizes minor variations or paraphrasing.
Similarity Metrics (e.g., Cosine Similarity, Semantic Similarity)	Measure semantic closeness between generated text and reference/input using embeddings.[54]	Various tasks (Relevance, Groundedness)	Higher is better (often -1 to 1 or 0 to 1)	Depends heavily on the quality of embeddings; may not capture nuances of meaning or task success.
Human Evaluation	Subjective assessment by human judges on dimensions like fluency, adequacy, coherence, relevance,	All tasks, especially generative ones	Ratings/Rankings based on criteria	Subjective, costly, time-consuming, potential for inter-rater variability.

	safety, etc..[182]			

V. Frameworks for Responsible AI Governance

Effective governance is crucial for ensuring that AI systems are developed and deployed responsibly, ethically, and in compliance with legal and societal expectations. This involves establishing clear policies, processes, and accountability structures.

Enhancing Developer and Provider Transparency

Transparency is a cornerstone of responsible AI governance, enabling trust, accountability, and informed decision-making. Key mechanisms for achieving transparency include standardized documentation:

- **Model Cards:** Functioning like "nutrition labels" for AI models, Model Cards provide concise, structured information about a model's intended use, performance characteristics (including across different demographic groups), limitations, ethical considerations, training data, and evaluation results.[20] They aim to help developers, deployers, policymakers, and the public understand how a model works and its potential impacts.[20] Google [20] and Salesforce [20] are notable proponents.
- **Datasheets for Datasets:** Proposed as a standard practice for documenting datasets used in machine learning, these datasheets detail the dataset's motivation, composition, collection process, preprocessing steps, labeling procedures, intended and recommended uses, distribution methods, and maintenance plans.[59] They increase transparency about the data foundation of AI models, helping to identify potential biases, limitations, and ethical issues related to the data itself.[59]

These documentation standards are increasingly recognized as essential components of trustworthy AI frameworks. The NIST AI RMF emphasizes documentation throughout the AI lifecycle.[90] The EU AI Act mandates comprehensive technical documentation for high-risk systems, covering aspects addressed by Model Cards and Datasheets.[41] ISO/IEC 42001 requires documentation as part of the AI Management System.[81] Transparency is also a core principle in guidelines from the OECD [10], UNESCO [111], the US AI Bill of Rights [14], and numerous corporate frameworks.[27] The goal is to move beyond opaque "black box" systems towards understandable and scrutable AI.[56]

Methodologies for Community Engagement and Feedback

Given the socio-technical nature of AI [5], engaging with diverse stakeholders—including end-users, affected communities, domain experts, and civil society groups—is critical for responsible development and deployment.[8] Engagement helps identify potential harms and biases, ensures alignment with community values and needs, builds trust, promotes equity, and informs ethical considerations.[177] This is particularly vital for AI applications in sensitive domains like healthcare [196] or public services.

Various methodologies can be employed for effective engagement:

- **Participatory Design & Co-creation:** Involving stakeholders directly in the design and development process.[9] Workshops can gather diverse perspectives on potential impacts and mitigation strategies.[197]
- **Stakeholder Consultation:** Seeking input and feedback from relevant groups at various stages of the AI lifecycle.[177]
- **Community Advisory Boards:** Establishing standing groups representing affected communities to provide ongoing guidance.
- **Ethnographic Methods:** Employing anthropological techniques to gain deep, contextual understanding of how AI impacts specific communities or cultural practices.[200]
- **Feedback Mechanisms:** Implementing channels for users and the public to report issues, concerns, or adverse impacts encountered with deployed AI systems.[83]
- **Delphi Method:** Using structured expert elicitation techniques to achieve consensus on ethical principles or evaluation criteria.[198]

Effective community engagement should be guided by principles such as mutual trust, respect for diverse perspectives and values, mutual understanding through open communication, accountability, integrity in the process, and transparency.[196] It requires recognizing the legitimacy of different forms of knowledge (epistemic legitimacy) [196] and genuinely aiming for co-creation rather than top-down imposition.[198] Supporting AI literacy among participants may also be necessary to enable meaningful contribution.[24] Evaluating the success of engagement involves assessing whether it genuinely influenced the AI system's design or deployment and fostered trust.[196] Frameworks like the AIM-AHEAD principles provide guidance for ethical community engagement in health AI.[198]

Ensuring Compliance with Regulations and Standards

Navigating the complex landscape of AI regulations and standards is a core function of AI governance. Compliance is essential not only to avoid legal penalties but also to build trust and ensure responsible practices.[80]

Key Regulations:

- *EU AI Act:* A landmark comprehensive regulation establishing a risk-based framework (Prohibited, High-Risk, Limited Risk, Minimal Risk). Imposes strict requirements for high-risk systems concerning data quality, documentation, transparency, human oversight, accuracy, robustness, and cybersecurity. Requires conformity assessments and establishes significant fines for non-compliance.[42]
- *General Data Protection Regulation (GDPR):* Applies to AI systems processing personal data of EU residents, mandating principles like data minimization, purpose limitation, and data subject rights.[30]
- *Sector-Specific Regulations:* AI applications in areas like healthcare (e.g., FDA regulations for AI medical devices [203]), finance (e.g., rules against discriminatory lending [12]), and employment are subject to existing domain-specific laws.

Key Standards:

- *ISO/IEC 42001:2023:* The first international standard for an AI Management System (AIMS). Provides a framework for establishing, implementing, maintaining, and improving AI governance within an organization. Covers risk management, AI impact assessment, system lifecycle management, data management, transparency, and defines specific controls (Annex A).[80] It follows the Plan-Do-Check-Act cycle common to other ISO management systems (like ISO 27001 for security).[80]
- *IEEE P7000™ Series:* A family of standards addressing specific ethical considerations in system design, including ethical process (7000), transparency (7001), data privacy (P7002), algorithmic bias (7003), child/student data (P7004), employer data (P7005), ethically driven robotics (7007), fail-safe design (P7008), ethical nudging (P7009), and wellbeing metrics (7010).[3]
- *NIST AI Risk Management Framework (RMF):* A voluntary framework developed in the US to help organizations manage AI risks and promote trustworthy AI. Organised around four functions: Govern, Map, Measure, Manage.[6] Includes characteristics of trustworthy AI.[5]

Principles & Guidelines: Complementing formal regulations and standards are influential sets of principles:

- *OECD AI Principles:* An intergovernmental standard adopted by numerous countries, promoting AI that is innovative, trustworthy, respects human rights, and based on five values (inclusive growth, human-centered values/fairness, transparency/explainability, robustness/security/safety, accountability) and five policy recommendations.[30]

- *UNESCO Recommendation on the Ethics of AI:* A global standard emphasizing human rights, dignity, diversity, and environmental flourishing, based on four values and ten principles (including proportionality, safety/security, fairness, transparency, accountability, sustainability, human oversight).[30]
- *AI Bill of Rights (US White House Blueprint):* Outlines five principles for protecting rights in the age of AI: Safe and Effective Systems; Algorithmic Discrimination Protections; Data Privacy; Notice and Explanation; Human Alternatives, Consideration, and Fallback.[14]
- *Corporate Frameworks:* Many companies (e.g., Microsoft [71], AWS [27], Google [55], Salesforce [20], IBM [7], AMD [72]) have published their own responsible AI principles, often mirroring those from intergovernmental bodies.

Compliance Assessment: Verifying compliance involves various methods:

- *Conformity Assessments:* Mandated by the EU AI Act for high-risk systems, involving either internal controls (self-assessment against standards) or assessment by a third-party Notified Body, resulting in CE marking.[41]
- *Certification:* Organizations can seek certification against standards like ISO/IEC 42001 to demonstrate compliance with best practices.[204] IEEE offers CertifAIEd™ for assessing ethics.[67]
- *Audits:* Internal or external audits assess adherence to policies, regulations, and standards.[18]
- *Checklists & Tools:* Compliance checklists [79] and AI governance platforms [18] help organizations track requirements and manage compliance activities.
- *Monitoring:* Continuous monitoring of AI systems and governance processes is essential.[76]

Navigating Intellectual Property and Data Ownership

Intellectual Property (IP) Issues: AI significantly challenges existing IP frameworks, primarily copyright and patent law.[25]

- *Authorship/Inventorship:* Traditional IP law attributes ownership to human creators/inventors.[25] Current legal positions in major jurisdictions like the US and UK generally hold that works generated *solely* by AI, without sufficient human creative input, are not eligible for copyright or patent protection.[25] The threshold for "sufficient human authorship" in AI-assisted works (e.g., using AI tools for parts of the creative process) is still being defined by courts and copyright offices.[212]
- *Training Data Infringement:* Training generative AI models often involves ingesting

vast amounts of data, potentially including copyrighted works (text, images, music) without explicit permission from rights holders. This has led to numerous lawsuits arguing copyright infringement, with ongoing debates about whether such training constitutes "fair use" (in the US) or falls under exceptions in other jurisdictions.[26]

- *Ownership of AI Outputs:* The terms and conditions (T&Cs) of generative AI tools vary regarding ownership of the content produced. Some assign ownership to the user (e.g., ChatGPT's policy [176]), while others may retain rights or grant only non-exclusive licenses, especially for free tiers (e.g., Midjourney's policy [176]). These T&Cs add another layer of complexity to IP ownership.[217]
- *Global Fragmentation:* IP laws concerning AI are not harmonized globally, creating legal uncertainty for cross-border AI development and deployment.[213]

Data Ownership & Governance: Data is a critical asset for AI, making its governance paramount.[173] Responsible AI governance requires robust data governance practices.[76]

- *Key Principles:* Establishing clear data ownership and stewardship roles within the organization.[18] Implementing strong data quality management processes (validation, cleansing, monitoring) to ensure data is accurate, complete, consistent, and timely for AI training and operation.[18] Ensuring data security (encryption, access controls) and privacy compliance (adhering to GDPR, CCPA, etc.).[18]
- *Best Practices:* Defining policies for the entire data lifecycle, including collection, processing, storage, sharing, retention, and deletion.[76] Implementing Role-Based Access Controls (RBAC) and principles of least privilege.[218] Utilizing data lineage tools to track data provenance and transformations for transparency and auditability.[18] Employing data catalogs and metadata management for better organization and discovery.[220] Using AI itself to automate aspects of data governance, such as data classification, quality monitoring, and compliance checks.[18]

Adaptation Insights: Applying principles from Policy Analysis, Law, and Corporate Social Responsibility (CSR)

Evaluating AI responsibility and governance benefits from insights from several disciplines:

- **Policy Analysis:** Provides frameworks and methods to analyze the design, implementation, and effectiveness of different AI regulations and standards (e.g., comparing the mandatory, risk-based EU AI Act with the voluntary, framework-based NIST RMF).[221] Helps understand the policy objectives,

stakeholder interests, and potential unintended consequences of different governance approaches.
- **Law:** Legal analysis is fundamental for interpreting and applying existing laws (IP, data protection like GDPR, liability) to AI systems.[25] It informs compliance requirements and helps navigate the ambiguities created by AI in areas like authorship or responsibility.
- **Corporate Social Responsibility (CSR):** Principles of CSR inform voluntary corporate commitments to ethical conduct, stakeholder engagement, social responsibility, and environmental sustainability that go beyond strict legal compliance.[5] Many corporate AI ethics frameworks draw on CSR concepts. Enterprise risk management (ERM) practices also inform AI risk management frameworks like NIST RMF.[6]

A key realization from examining these governance frameworks is their foundational role in enabling the assessment of other non-performance dimensions. Effective governance structures, such as those outlined in the NIST RMF 'Govern' function [6] or required by ISO 42001 [80], establish the necessary policies, processes, roles, and responsibilities for managing AI risks and ensuring trustworthy practices.[76] For example, evaluating fairness reliably depends on robust data governance practices to ensure data quality and representativeness.[18] Assessing accountability requires traceability mechanisms like audit logs and lineage tracking, which are implemented through governance.[73] Measuring environmental impact necessitates systems for monitoring resource consumption, enabled by governance processes.[21] Ensuring accessibility requires organizational policies and development processes defined within a governance structure.[152] Therefore, the maturity and effectiveness of an organization's AI governance framework directly impacts its ability to meaningfully evaluate and manage other non-performance aspects. Assessing governance maturity itself (e.g., against ISO 42001 or NIST RMF Govern criteria) should be considered a critical component, perhaps even a prerequisite, for evaluating other dimensions like fairness or safety. A high score on a fairness metric, for instance, carries less weight without evidence of strong underlying data governance and bias auditing processes.

Furthermore, while ethical principles provide essential guidance, formal standards are increasingly defining the practical, operational meaning of 'responsible' or 'trustworthy' AI. Standards like ISO/IEC 42001 [80], the IEEE P7000 series [3], and documentation standards such as Model Cards [20] and Datasheets for Datasets [59] translate high-level principles [10] into specific, auditable requirements for processes, controls, and documentation. Compliance with these standards offers a verifiable method for organizations to demonstrate responsible practices and differentiate themselves in the market.[80] Regulatory bodies, like the EU with its AI Act, are also looking towards harmonized standards to provide a presumption of conformity with legal requirements.[66] Consequently, evaluation methodologies should increasingly incorporate assessments against these relevant standards, and comparative ranking systems could potentially leverage compliance with key standards as a significant indicator of responsible AI maturity. The ongoing development, refinement, and adoption of these standards will be crucial in shaping the future of AI governance and evaluation.

VI. Assessing Broader Societal Impacts of AI

Beyond immediate ethical concerns and system properties, AI deployment has far-reaching consequences for society, impacting employment, education, culture, and democratic processes. Assessing these broader impacts requires drawing on methodologies from the social sciences.

Impact on Employment and the Future of Work

Dual Nature: AI's impact on employment is complex and dual-natured. On one hand, AI-driven automation, particularly of routine cognitive and manual tasks, poses a risk of job displacement across various sectors.[175] Studies suggest AI could influence a significant portion of jobs, including white-collar roles previously considered less vulnerable to automation.[227] On the other hand, AI also acts as an augmentation tool, enhancing human capabilities, improving productivity, and potentially increasing job satisfaction by automating mundane tasks.[225] Furthermore, AI creates demand for new roles related to its development, deployment, maintenance, and governance, such as AI/ML specialists, data scientists, prompt engineers, AI ethicists, and cybersecurity experts.[175] AI is projected to contribute significantly to economic growth and productivity.[1]

Affected Roles & Skills: While initial automation focused on manufacturing and routine tasks [226], generative AI is increasingly impacting knowledge work and creative professions.[227] Occupations requiring significant physical or outdoor work appear less exposed [227], while roles demanding high levels of social interaction may increase in value.[175] There is a growing emphasis on skills complementary to AI, including critical thinking, creativity, socio-emotional skills (resilience, adaptability), technological literacy, and AI-specific competencies.[225] Reskilling and upskilling the workforce are becoming critical strategies for employers.[225]

Metrics & Analysis: Evaluating AI's impact on work involves tracking:

- *Job Displacement/Creation:* Net changes in employment levels overall and within specific sectors/occupations.[225]
- *Task Automation Rate:* Percentage of work tasks automated by AI.[228]
- *Productivity Metrics:* Changes in output per employee, time saved per task, error reduction rates.[228]
- *Skills Demand:* Tracking shifts in required skills through job postings analysis, employer surveys.[225] Measuring AI literacy and training completion rates.[228]
- *Workforce Transition:* Role transition success rates, employee satisfaction scores related to AI integration, retention rates.[228]

- *Economic Indicators:* Impact on wages, income inequality, labor costs, business profitability, and GDP.[173]

Frameworks & Studies: Major reports like the World Economic Forum's Future of Jobs Report [225] provide global forecasts and employer perspectives. Organizations like the OECD [1] and consultancies like McKinsey [173] conduct extensive research on AI's economic and productivity impacts. Occupational databases like O*NET are used to assess the exposure of different jobs to AI automation or augmentation.[227] Guidelines like the Partnership on AI's (PAI) Guidelines for AI and Shared Prosperity aim to promote equitable outcomes.[104]

Impact on Education and Learning

Potential Benefits: AI offers numerous possibilities for transforming education:

- *Personalized Learning:* AI can analyze student performance data to tailor content, pacing, and learning paths to individual needs and styles, potentially improving engagement and outcomes.[1] Adaptive learning platforms adjust difficulty in real-time.[192]
- *Assessment & Feedback:* AI can automate the grading of various assignment types (multiple choice, short answer, even essays), providing students with immediate, personalized feedback and freeing up teacher time for instruction and support.[192]
- *Content Creation & Curation:* AI tools can assist educators in generating lesson plans, activities, assessments, presentations, and supplementary materials, potentially saving significant preparation time.[193]
- *Administrative Efficiency:* AI can streamline tasks like scheduling, attendance tracking, communication with parents, and managing student records.[192]
- *Accessibility:* AI can enhance accessibility through tools like real-time transcription, text-to-speech, and alternative assessment formats.[194]
- *Data Analytics:* AI provides tools for analyzing learning trends, identifying students needing support, and informing instructional strategies.[193]

Challenges & Risks: Despite the potential, AI in education raises concerns:

- *Data Privacy & Security:* Use of student data requires robust protection measures.[194]
- *Bias:* AI algorithms trained on biased data could perpetuate or amplify educational inequalities.[194] AI-driven assessment tools might exhibit bias against certain groups (e.g., non-native English speakers [194]).
- *Reduced Human Interaction:* Over-reliance on AI could diminish crucial teacher-student and peer-to-peer interactions, potentially hindering

social-emotional development.[194]

- *Cost & Equity:* High implementation costs for sophisticated AI systems could exacerbate the digital divide between well-resourced and under-resourced schools.[194]
- *Academic Integrity:* AI tools make it easier for students to generate work that is not their own, requiring new approaches to assessment design and academic honesty policies.[194]
- *Reliability & Accuracy:* AI-generated content or feedback may contain errors or inaccuracies.[194]
- *Teacher Role & Training:* Educators need training and support to effectively integrate AI tools and adapt their pedagogical roles.[192]

Evaluation Focus: Assessing AI's impact requires looking at:

- *Learning Outcomes:* Changes in student performance (grades, test scores), knowledge retention, critical thinking skills, engagement levels.[192]
- *Teacher Impact:* Effects on teacher workload, job satisfaction, changes in teaching practices and roles.[192]
- *Assessment Validity & Fairness:* Comparing the effectiveness and fairness of AI-driven assessments versus traditional methods.[230]
- *Equity:* Analyzing whether AI tools benefit all student groups equally or widen existing gaps.[231]
- *User Experience:* Student and teacher satisfaction with AI tools.[193]

Metrics: Evaluation can use standard educational metrics (grades, standardized test scores, graduation rates), measures of student engagement and motivation [193], teacher time-use studies, user satisfaction surveys, fairness metrics applied to educational algorithms, and AI adoption rates.[194] Frameworks like Fink's Taxonomy of Significant Learning or Learning Assessment Techniques (LATs) can help design assessments that target higher-order skills potentially less susceptible to AI misuse.[233]

Impact on Creativity, Arts, and Culture

AI as Creative Tool & Disruptor: Generative AI models can now produce sophisticated outputs traditionally considered creative, including text, images, music, and video.[26] This capability allows AI to function as a tool that can augment human creativity, assist in the creative process (e.g., generating sketches, exploring variations), and potentially increase creative productivity.[234] Studies suggest AI adoption can enhance artists' output and peer evaluation.[234]

Copyright & Ownership Challenges: AI fundamentally challenges copyright law's traditional focus on human authorship.[26] Key issues include:

- *Copyrightability of AI Output:* As discussed in Section V, purely AI-generated works generally lack copyright protection in jurisdictions like the US and UK due to the absence of human authorship.[212] The threshold for copyrightability of AI-assisted works depends on the significance of human creative contribution.[212]
- *Infringement by Training Data:* Models trained on vast datasets scraped from the internet often include copyrighted materials without permission, leading to lawsuits from creators (writers, artists, musicians) alleging infringement.[26] The legality hinges on interpretations of fair use or copyright exceptions.
- *Ownership Ambiguity:* Besides copyright law, AI tool terms of service often dictate ownership or licensing rights for generated outputs, adding complexity.[176]

Economic & Labor Impacts: The creative industries face significant disruption. There are widespread concerns among artists, writers, musicians, and other creators that AI could devalue their skills, reduce demand for human labor, and displace jobs.[26] The ability of AI to mimic styles raises fears of unauthorized use and "theft".[235] Calls are growing for mechanisms to ensure fair compensation for creators whose works are used to train AI models.[213]

Cultural Impacts: AI is not just a tool but also a cultural force.

- *Cultural Diffusion & Transformation:* AI systems, trained on and generating cultural content (text, images, music), act as powerful agents of cultural transmission and change.[237]
- *Homogenization vs. Diversity:* If AI models are primarily trained on data reflecting dominant cultures (e.g., Western, English-language), they risk perpetuating those norms and marginalizing minority languages, values, and aesthetics.[24]
- *Preservation & Revitalization:* Conversely, AI could potentially be used to document, analyze, translate, and even help revitalize endangered languages and cultural heritage.[238]
- *AI as Cultural Artifact:* The outputs of AI systems (e.g., AI art, AI-generated narratives) become cultural artifacts themselves, requiring critical analysis from perspectives like cultural studies and anthropology.[239]

Evaluation Methods: Assessing these impacts requires diverse approaches:

- *Legal Analysis:* Examining copyright statutes, case law, and regulatory developments related to AI authorship and infringement.[25]
- *Economic Analysis:* Studying the impact on employment, wages, and market structures within creative industries.[26]
- *Cultural Studies:* Analyzing how AI influences cultural production, representation,

identity, and meaning-making.[236]

- *Anthropology & Ethnography:* Studying how creative communities adopt, adapt, or resist AI tools, and the impact on their practices and values.[200] Analyzing AI systems and outputs as cultural phenomena.[239]
- *Surveys & Qualitative Research:* Gathering perspectives from creators and the public on AI's role and impact.[26]

Impact on Democracy and Civic Engagement

Risks: AI poses significant threats to democratic processes and healthy civic discourse:

- *Disinformation & Manipulation:* Generative AI makes it cheaper, faster, and easier to create and disseminate convincing fake or misleading content (text, images, audio, video - "deepfakes") at scale.[241] This content can be used in influence operations by domestic or foreign actors to manipulate public opinion, sow discord, suppress votes, damage candidates' reputations, and erode trust in information sources and institutions.[242]
- *Political Polarization:* AI-powered microtargeting can deliver highly personalized (and potentially manipulative or polarizing) messages to specific voter segments.[241] Algorithmic content curation on social media platforms can create echo chambers and amplify extreme or divisive content, further entrenching polarization.[245] AI-generated content often reinforces pre-existing beliefs.[246]
- *Election Integrity:* AI can be used to automate cyberattacks against election infrastructure, overwhelm election officials with frivolous requests, or introduce bias into automated election administration processes like voter roll maintenance or signature verification.[241] The mere *perception* of widespread AI manipulation can undermine trust in election outcomes.[245]
- *Bias & Discrimination:* AI systems used in political contexts (e.g., content moderation, ad targeting, election administration) can inherit and amplify societal biases, particularly racial biases, potentially leading to differential treatment or disenfranchisement of minority groups.[241]
- *Surveillance & Chilling Effects:* AI-powered surveillance technologies can be used to monitor political dissent and activism, potentially chilling free speech and civic participation, especially among marginalized communities.[244]

Potential Benefits: While risks are prominent, some potential positive applications exist:

- *Efficiency & Transparency:* AI could potentially improve the efficiency, accuracy, and transparency of certain election administration tasks.[247]

- *Increased Participation (Potential):* Tools like AI chatbots could provide accessible information about voting and candidates, potentially enfranchising some citizens, although biased information delivery is a risk.[243]
- *Policy Analysis:* AI might assist policymakers or researchers in analyzing complex issues, such as identifying and mitigating restrictive zoning laws.[167]

Metrics & Analysis: Evaluating AI's democratic impact is challenging but crucial:

- *Disinformation Tracking:* Measuring the prevalence, reach, and engagement with AI-generated disinformation across platforms.[246] Assessing the effectiveness of detection tools.[243]
- *Impact on Voters:* Studying the effects of exposure to AI-generated content on voter knowledge, attitudes, polarization levels, trust in media/institutions, and voting behavior (causality is very difficult to establish).[245]
- *Election System Audits:* Auditing AI tools used in election administration for bias, security vulnerabilities, and accuracy.[241]
- *Platform Policy Analysis:* Evaluating content moderation policies of social media platforms regarding AI-generated political content.[246]
- *Polarization Measures:* Using survey data or social network analysis to track levels of political polarization over time.[245]

Frameworks & Studies: Research organizations like the Carnegie Endowment [242] and The Alan Turing Institute's Centre for Emerging Technology and Security (CETaS) [246] monitor AI's impact on elections. Academic research focuses on specific threats like racial harms [241] or the mechanisms of disinformation spread.[243] Legal and policy analysis examines potential regulatory responses.[244]

Adaptation Insights: Utilizing methodologies from Sociology, Economics, Political Science, Anthropology, and Social Impact Assessment (SIA)

Assessing the broad societal impacts of AI necessitates drawing upon the theoretical frameworks and methodological toolkits of various social sciences:

- **Sociology:** Offers theories of social change, inequality, network analysis, and methods like surveys and qualitative analysis to study impacts on social structures, group dynamics, work, and culture.[187]
- **Economics:** Provides models and metrics for analyzing labor market shifts, productivity changes, economic growth, cost-benefit analysis, and inequality.[173]
- **Political Science:** Contributes theories of democracy, political behavior, polarization, international relations, and methods for analyzing elections, public opinion, and the role of information in politics.[244]
- **Anthropology:** Offers ethnographic methods for deep, contextual understanding

of how AI is experienced, adopted, and resisted within specific cultural settings, revealing impacts on values, practices, and lived experience.[200] Frameworks like Actor-Network Theory (ANT) can analyze human-technology interactions.[237]

- **Social Impact Assessment (SIA):** Provides a specific, structured methodology, adapted from Environmental Impact Assessment (EIA), designed to systematically analyze, monitor, and manage the intended and unintended social consequences of planned interventions (like deploying an AI system).[151] SIA typically involves scoping potential impacts, baseline studies, prediction of effects, evaluation of alternatives, stakeholder engagement, mitigation planning, and monitoring.[248] It often employs mixed methods [240] and emphasizes participation.[223] SIA frameworks can be explicitly adapted for AI impact assessment [199], potentially incorporating techniques like system dynamics simulation for predictive modeling.[250]

A significant challenge in evaluating AI's societal impacts is the difficulty of establishing clear causality and finding reliable measurements. Societal systems are inherently complex, influenced by numerous interacting factors.[250] Isolating the specific contribution of AI to macro-level trends like employment rates, election outcomes, or cultural shifts is methodologically demanding, as AI is introduced alongside many other technological, economic, and political changes.[229] Many impacts are indirect, cumulative, emergent, or manifest only over the long term, making them hard to capture with standard metrics.[250] Available quantitative metrics often focus on immediate, measurable effects (e.g., task automation rates [228]) but may miss deeper, systemic transformations. Qualitative methods like SIA and ethnography provide crucial context and understanding of mechanisms but can be resource-intensive and may face challenges with generalizability or direct comparability.[240] Therefore, assessing societal impact requires methodological pluralism, combining quantitative and qualitative approaches, employing longitudinal study designs where possible, and maintaining a degree of humility about the certainty of causal claims. Evaluation frameworks should focus on identifying plausible impact pathways and correlations, supported by triangulated evidence, rather than seeking definitive proof of impact in all cases. Transparency regarding methodological limitations is essential.

Across various societal domains, AI appears not as a monolithic force for either good or ill, but as a technology with a pronounced dual potential. It threatens job displacement through automation but simultaneously offers tools for augmenting human work and creates new employment opportunities.[175] It risks deepening the digital divide due to cost and data biases but also presents potential solutions for personalized education and improved access to information for underserved populations.[24] AI enables the creation of sophisticated disinformation that can harm democracy but might also be leveraged for detecting such content or facilitating civic engagement.[242] In the creative sphere, AI challenges copyright and threatens livelihoods while also providing powerful new tools for artistic expression.[26] Similarly, AI's resource demands can hinder environmental goals, yet AI can also optimize processes like recycling and energy management.[142] This duality implies that evaluating AI's societal impact requires a balanced assessment of both potential harms and potential benefits within each domain. Methodologies should be designed to capture this complexity, moving beyond simple positive or negative scores. The net impact in any given context is highly contingent on specific design choices, implementation strategies, governance frameworks, and the surrounding social, economic, and political environment. Comparative evaluations or rankings must reflect this nuance, potentially focusing on the effectiveness of mitigation strategies alongside risk assessment, rather than assigning a single impact score.

VII. Synthesizing Methodologies for Holistic AI Assessment and Ranking

Bringing together the diverse evaluation approaches across ethical, environmental, accessibility, governance, and societal dimensions requires identifying promising methodologies, considering how systems might be compared, and addressing the inherent challenges involved.

Promising Adaptable Methodologies from Adjacent Fields

Several methodologies developed in other disciplines show significant potential for adaptation to the holistic evaluation of AI systems:

- **Life Cycle Assessment (LCA):** Originating from environmental science and engineering, LCA offers a rigorous, standardized framework for quantifying environmental impacts (energy, carbon, water, materials, waste) across the entire lifecycle of a product or system, from raw material extraction to end-of-life.[118] Its application to AI hardware [128] and potentially software models provides a comprehensive view beyond just operational impacts.
    - *Strengths for AI Eval:* Holistic environmental scope, standardized approach (ISO 14040/44), quantifies embodied impacts.
    - *Challenges/Adaptation Needs:* Requires extensive supply chain data (often proprietary), complex to perform, needs adaptation for rapidly evolving AI software/hardware lifecycles.
- **Social Impact Assessment (SIA):** Developed in urban planning and development studies (often linked to EIA), SIA provides a systematic process to analyze, monitor, and manage the social consequences (positive and negative) of projects

or policies on communities.[151] It emphasizes stakeholder participation, context-specific analysis, and consideration of diverse impacts (economic, cultural, health, community cohesion). It is directly adaptable for assessing the societal impacts of AI deployment.[199]

  - *Strengths for AI Eval:* Focus on real-world social consequences, participatory approach, context sensitivity.
  - *Challenges/Adaptation Needs:* Can be heavily qualitative, lacks universal methodological consensus [249], requires social science expertise, predicting long-term AI impacts is difficult.

- **Capability Approach:** Stemming from philosophy and economics, this framework evaluates well-being and justice based on individuals' substantive freedoms and opportunities ('capabilities') to achieve valued states of being and doing.[108] It assesses how AI systems, considering necessary 'conversion factors', enable or inhibit these capabilities.
  - *Strengths for AI Eval:* Strong normative grounding in human well-being and equity, focuses on real outcomes rather than just resource distribution.
  - *Challenges/Adaptation Needs:* Highly conceptual, requires significant effort to operationalize into measurable indicators for specific AI contexts, identifying relevant capabilities and conversion factors is complex.
- **Web Content Accessibility Guidelines (WCAG):** The established standard from the W3C for digital accessibility, based on POUR principles and providing testable success criteria.[152] These can be directly adapted for evaluating the accessibility of AI user interfaces.[161]
  - *Strengths for AI Eval:* Widely accepted standard, testable criteria, clear principles, addresses needs of users with disabilities.
  - *Challenges/Adaptation Needs:* Primarily designed for web content, needs careful adaptation for dynamic, conversational, or multimodal AI interfaces; WCAG 3.0 aims to address some flexibility needs.[162]
- **Ethnography and Anthropological Methods:** Qualitative methods focused on deep, contextual understanding of human behavior, culture, and social practices through observation and interaction within specific communities.[200] Highly relevant for understanding how AI is actually used, perceived, and impacts diverse cultural contexts and lived experiences.
  - *Strengths for AI Eval:* Rich contextual insights, uncovers tacit knowledge and unintended consequences, centers human experience.
  - *Challenges/Adaptation Needs:* Findings may not be easily generalizable or quantifiable for ranking, resource-intensive, requires anthropological expertise.
- **Human Rights / Fundamental Rights Impact Assessment (HRIA/FRIA):** Adapts

legal and ethical human rights frameworks to systematically identify, assess, and mitigate risks that AI systems pose to fundamental rights.[109]

- *Strengths for AI Eval:* Strong normative and legal grounding, systematic risk identification process, aligns with regulatory trends (e.g., EU AI Act).
- *Challenges/Adaptation Needs:* Requires legal and ethical expertise, defining impact thresholds can be complex.

Table 4: Adaptable Methodologies from Non-AI Fields

Methodology	**Originating Field(s)**	**Core Concept/Purpose**	**Potential AI Application Area(s)**	**Key Strengths for AI Eval**	**Key Challenges/ Adaptation Needs**
Life Cycle Assessment (LCA) [149]	Environmental Science, Engineering	Assess environmental impacts across entire product/system lifecycle (cradle-to-grave)	Environmental Impact (Hardware, Models)	Holistic scope, standardized (ISO 14040), quantifies embodied impacts	Data-intensive (supply chain), complex, needs adaptation for software/fast cycles
Social Impact Assessment (SIA) [151]	Social Sciences, Planning, Development Studies	Analyze, monitor, manage social consequences of interventions	Societal Impact (Employment, Democracy, Culture), Equity	Focus on real-world consequences, participatory, context-sensitive	Often qualitative, methodological variety [249], prediction difficulty, expertise needed
Capability Approach [108]	Philosophy, Economics	Evaluate impact on individuals' substantive freedoms & opportunities	Equity, Societal Impact, Ethics	Focus on human well-being & equity, outcome-oriented	Conceptual, requires operationalization into metrics, complex analysis
Web Content Accessibility Guidelines (WCAG) [152]	Web Development, Accessibility Studies	Ensure digital content is Perceivable, Operable, Understandable, Robust (POUR) for	Accessibility & Equity (Interface Usability)	Established standard, testable criteria, addresses disability access	Needs adaptation for dynamic/conversational AI, primarily web-focused

		PWD			
Ethnography / Anthropological Methods [239]	Anthropology, Sociology	Deep contextual understanding of culture, practices, lived experience via observation/interaction	Societal Impact, Cultural Impact, Equity, User Experience	Rich qualitative insights, reveals tacit knowledge & context	Not easily scalable/quantifiable, resource-intensive, expertise needed
HRIA / FRIA [109]	Law, Ethics, Human Rights	Systematically assess risks to fundamental human rights	Ethical Considerations, Responsibility & Governance	Rights-based, aligns with legal norms, systematic risk focus	Requires legal/ethical expertise, defining impact thresholds
Multi-Attribute Decision Making (MADM) [251]	Operations Research, Decision Science	Systematically evaluate options based on multiple criteria	Comparative Analysis & Ranking (across all dimensions)	Structured comparison, handles multiple criteria & trade-offs, incorporates preferences	Requires defining criteria/weights/scores, sensitive to choices

Approaches to Comparative Analysis and Ranking

While challenging, various approaches exist or could be adapted for comparing AI systems based on non-performance factors:

- **AI Indices and Benchmarks:** Existing indices like the Stanford AI Index [203] track trends, including some related to responsible AI, while the OECD AI Index aims for a comprehensive measure of trustworthy AI.[206] Performance benchmarks like MLPerf are incorporating efficiency.[15] However, current benchmarks often lack comprehensive coverage of non-performance dimensions or face issues with standardization, validity, and potential for "gaming".[4] Specialized ethical benchmarks are emerging but are still nascent and face scalability and validity challenges.[252] A significant gap exists between the need for holistic non-performance evaluation and the capabilities of current standard AI benchmarking practices.
- **Multi-Attribute Decision Making (MADM):** A family of techniques from decision science designed explicitly for comparing alternatives based on multiple, often conflicting, criteria.[251] Methods like the Analytic Hierarchy Process (AHP), weighted scoring methods, or more advanced fuzzy logic approaches (like complex intuitionistic fuzzy rough sets - CIFRS [251]) allow for structured evaluation. They can incorporate both quantitative metrics and qualitative assessments (converted to scores), and allow for weighting criteria based on context or stakeholder preferences.[113] MADM frameworks can explicitly model and analyze trade-offs.[113]
- **Comparative Frameworks:** Utilizing frameworks designed for comparing policies or principles can be adapted. For example, comparing AI systems against the requirements of the EU AI Act [256] or the principles of the NIST RMF [92] provides a structured basis for comparison, particularly within specific risk categories or domains. Frameworks comparing ethical principles across models or domains also exist.[42]
- **Scorecards:** Developing tailored scorecards that combine relevant quantitative and qualitative metrics for specific non-performance dimensions (like human impact [258] or specific ethical principles) allows for a summarized assessment that can facilitate comparison.
- **Tiered/Risk-Based Categorization:** Instead of a single numerical rank, systems could be categorized into tiers based on their non-performance risk profile or maturity level, similar to the EU AI Act's risk classification.[65] This acknowledges complexity while still providing comparative information.

Addressing Challenges in Quantification, Adaptation, and Trade-offs

Any attempt at holistic assessment and ranking must confront significant challenges:

- **Quantification:** Translating qualitative concepts like fairness nuances, ethical alignment, cultural sensitivity, or the effectiveness of human oversight into meaningful quantitative metrics remains a major hurdle.[5] While progress has been made (e.g., fairness metrics [33]), many dimensions lack reliable, agreed-upon measures.[6] There is a need for robust methodologies to map qualitative insights (e.g., from SIA or ethnography) onto quantitative scales without losing critical nuance.[8] Oversimplification through single scores is a significant risk.[15]
- **Adaptation:** Effectively applying methodologies from fields like LCA, SIA, or ethics requires careful adaptation to the specific context of AI systems, considering their unique characteristics (e.g., data-driven nature, potential for emergent behavior, rapid evolution).[161] This necessitates interdisciplinary collaboration and expertise.[75]
- **Context-Dependency:** Many crucial non-performance aspects—including fairness definitions, risk tolerance, societal impacts, and the relevance of specific ethical principles—are highly dependent on the specific application domain, cultural context, user population, and deployment environment.[5] Universal metrics or rankings may therefore lack validity or practical relevance.[15] Context-specific assessments, potentially using tailored profiles (like NIST RMF Use-Case Profiles [92]) or methodologies like SIA, are often more appropriate.
- **Trade-offs:** Perhaps the most fundamental challenge is managing the inherent tensions and trade-offs between different desirable properties of AI systems. Optimizing for accuracy might compromise fairness or explainability [113]; enhancing transparency might conflict with privacy or security [30]; reducing environmental impact might affect performance [119]; maximizing privacy might hinder fairness assessments.[88] These trade-offs are unavoidable.[88] Therefore, any meaningful comparative evaluation cannot simply aggregate scores across dimensions but must explicitly address how these trade-offs are identified, evaluated, prioritized, justified, and documented within a specific context.[113] The management of trade-offs is itself a critical aspect of responsible AI governance. A simple additive ranking score across all non-performance dimensions is likely flawed because it ignores these inherent conflicts. Comparative methodologies need to incorporate ways to represent and assess these trade-offs, perhaps using multi-objective optimization perspectives or MADM approaches with context-dependent weighting.[251]

The Role of Socio-Technical Evaluation Approaches

Addressing the complexity and context-dependency of non-performance evaluation necessitates adopting socio-technical perspectives and methods. Recognizing that AI systems are not purely technical artifacts but exist within and interact with complex social systems is crucial.[5]

- **Necessity:** Purely technical metrics (e.g., algorithm speed, prediction accuracy) are insufficient to capture the real-world impacts and ethical implications of AI.[239] Evaluation must consider the interplay between the technology, the people who develop and use it, the organizational context, and the broader societal environment.[8]
- **Methods:** Socio-technical evaluation involves integrating qualitative methods (like ethnography, interviews, case studies, participatory workshops) with quantitative data collection and analysis.[5] It requires involving a diverse range of stakeholders—including domain experts, social scientists, ethicists, legal experts, end-users, and members of potentially impacted communities—throughout the AI lifecycle, from design to post-deployment monitoring.[5] Frameworks like the NIST AI RMF explicitly encourage this by incorporating context mapping (MAP function) and allowing for mixed-method measurement (MEASURE function).[5] Assessing human factors, user experience, and organizational processes is integral.[5]
- **Examples:** Practical examples include evaluating AI in hotel housekeeping by considering housekeeper needs and safety [9], assessing AI for city planning or infrastructure monitoring with community input [9], or using participatory workshops to evaluate health AI impacts.[197] The development of the NIST RMF itself involved extensive multi-stakeholder input, reflecting a socio-technical approach to framework creation.[205]

Table 5: Comparison of Major AI Governance Frameworks

Framework	Key Focus	Legal Status	Key Requirements/Principles	Enforcement/Assessment	Strengths	Weaknesses/Limitations
NIST AI RMF [6]	Risk Management, Trustworthiness	Voluntary (US Gov framework)	Govern, Map, Measure, Manage functions; Trustworthiness characteristics (Fairness, Safety, Transparency, etc.)	Self-assessment, Use-Case Profiles [92]	Flexible, adaptable, comprehensive risk focus, promotes socio-technical view	Voluntary (non-binding), less prescriptive on specific metrics/controls
EU AI Act [65]	Legal Regulation, Risk-Based	Mandatory (EU Law)	Risk tiers (Prohibited, High, Limited, Minimal); Strict requirements for High-Risk (data, docs, transparency, oversight, robustness); Penalties	Conformity Assessment (Internal/Notified Body), Market Surveillance, Fines [77]	Legally binding, comprehensive, strong enforcement potential, sets global precedent	Complex, potentially burdensome for innovation, definitions/scope may evolve
OECD AI Principles [10]	Ethical Principles, Policy Guidance	Intergovernmental Standard (Adherence is	5 Values (Incl. Growth, Human-Centric/Fair	Peer review, National strategy alignment,	Widely adopted, influential globally, balances	High-level principles, lacks specific metrics/en

		voluntary commitment)	ness, Transparency, Robustness, Accountability); 5 Policy Recs	OECD.AI monitoring [206]	innovation & values	forcement mechanisms
ISO/IEC 42001 [80]	AI Management System (AIMS)	International Standard (Voluntary Certification)	Requires AIMS establishment (policies, objectives, processes); Risk/Impact Assessment; Lifecycle Mgmt; Specific Controls (Annex A)	Third-party certification audits [207]	Structured, auditable, integrates with other ISO standards (e.g., 27001), practical controls	Voluntary, cost/effort of implementation & certification, may need supplementing with domain-specific standards
IEEE P7000 Series / EAD [3]	Specific Ethical Design Aspects	Standards (Voluntary adoption)	Detailed standards on Ethics Process (7000), Transparency (7001), Privacy (P7002), Bias (7003), Fail-Safe (P7008), etc.	Conformance testing, IEEE CertifAIEd™ [67]	Deep focus on specific technical/ ethical areas, practical guidance	Fragmented across multiple standards, adoption level varies
UNESCO	Global	Internatio	4 Values	Readiness	Global	High-level

Rec. on Ethics of AI [111]	Ethical Principles	nal Recommendation (Member State commitment)	(Human Rights, Peace, Diversity, Environment); 10 Principles (Proportionality, Safety, Privacy, Fairness, Transparency, etc.)	Assessment Methodology (RAM), Ethical Impact Assessment (EIA) tool [111]	scope, strong human rights focus, emphasizes inclusivity & environment	principles, relies on member state implementation, lacks direct enforcement

VIII. Conclusion and Future Directions

Recap of the State-of-the-Art

The evaluation of Artificial Intelligence systems is rapidly evolving beyond traditional performance metrics to encompass a critical range of non-performance dimensions: ethical considerations, environmental impact, accessibility and equity, responsibility and governance, and broader societal impacts. This report has synthesized a diverse landscape of methodologies, frameworks, metrics, and tools aimed at assessing these crucial aspects.

Significant progress is evident in certain areas. Robust toolkits and metrics exist for evaluating algorithmic fairness (e.g., AIF360, Fairlearn) [17] and model explainability (e.g., LIME, SHAP).[45] Standardization efforts are yielding concrete documentation practices like Model Cards [20] and Datasheets for Datasets.[59] Tools for tracking operational carbon footprints (e.g., CodeCarbon [126]) are becoming available, and Life Cycle Assessment (LCA) provides a comprehensive methodology for environmental impact.[149] Accessibility evaluation can leverage established standards like WCAG [152], adapted for AI. Major governance frameworks like the NIST AI RMF [6], the EU AI Act [65], and ISO/IEC 42001 [80] provide structured approaches to risk management and responsible practices. Methodologies like Social Impact Assessment (SIA) [151] and rights-based assessments [109] offer pathways for evaluating broader societal effects.

However, the landscape remains fragmented. There is a lack of universally accepted metrics for many dimensions, particularly for complex societal impacts, cultural effects, and the effectiveness of human oversight. Methodologies are often siloed within disciplines, and adapting them for the unique challenges of AI requires significant interdisciplinary effort. A critical gap persists between the proliferation of high-level ethical principles and their practical, verifiable implementation.[16] Furthermore, the inherently socio-technical nature of AI necessitates evaluation approaches that integrate technical assessment with deep contextual understanding and stakeholder engagement.[8]

Assessment of Feasibility and Limitations of a Comprehensive Ranking System

The desire for a comparative ranking system for AI based on non-performance factors is understandable, as it could potentially incentivize more responsible development and guide decision-making. Based on the synthesized methodologies, a limited form of comparison appears feasible under specific conditions. Using Multi-Attribute Decision Making (MADM) frameworks [251] or tailored scorecards [258], it might be possible to compare AI systems within a specific domain (e.g., healthcare chatbots) or against a defined set of criteria (e.g., compliance with EU AI Act requirements for high-risk systems). Such comparisons would require careful selection of relevant dimensions, metrics (both quantitative and qualitative, appropriately scaled), and context-specific weighting of criteria based on stakeholder input or regulatory priorities.

However, creating a single, universal, comprehensive ranking system for AI across all non-performance dimensions faces profound limitations and is likely infeasible and potentially misleading. The key challenges identified throughout this report—the difficulty of quantifying qualitative aspects [5], the critical importance of context-dependency for validity [30], the lack of standardized and reliable metrics for many crucial dimensions [4], the difficulty in meaningfully aggregating diverse metrics, and the fundamental problem of navigating inherent trade-offs between desirable principles [113]—collectively argue against the utility and validity of a simplistic overall score or rank. Such a score would inevitably obscure crucial nuances, context, and the value judgments embedded in weighting different factors.

Recommendations for Advancing Research, Standardization, and Practice

To advance the holistic evaluation of AI systems, concerted effort is needed across research, standardization, and practical implementation:

Research:

- **Interdisciplinary Collaboration:** Foster deeper collaboration between computer scientists, engineers, social scientists, environmental scientists, ethicists, legal scholars, and domain experts to develop and validate evaluation methodologies.[16]
- **Metric Development:** Prioritize research on developing robust, reliable, and valid metrics for currently under-measured dimensions, including long-term societal impacts, cultural impacts, human oversight effectiveness, and the nuances of fairness beyond group statistics. Explore methods for rigorously mapping qualitative data to quantitative scales.[8]
- **Trade-off Analysis:** Develop and refine methodologies for systematically identifying, analyzing, representing, and justifying trade-offs between different non-performance objectives.[113]
- **Empirical Validation:** Conduct more empirical studies to assess the real-world effectiveness and impact of different AI governance mechanisms, ethical design practices, and evaluation techniques.[75] Longitudinal studies are needed to understand long-term impacts.
- **Benchmark Evolution:** Move beyond performance-only benchmarks to develop integrated benchmarks that incorporate key non-performance dimensions, while addressing validity and scalability concerns.[15]

Standardization:

- **Consensus Building:** Support multi-stakeholder efforts (involving industry, academia, civil society, government) to develop consensus standards for key metrics, documentation formats (e.g., extending Model Cards/Datasheets), and evaluation protocols, particularly for high-risk or high-impact AI applications.[57]
- **Lifecycle Assessment Standards:** Promote the development and adoption of standardized LCA methodologies specifically tailored for AI hardware and software systems to ensure comprehensive environmental assessment.[129]
- **Transparency Mandates:** Encourage or mandate greater transparency from AI developers and cloud providers regarding model training data, energy consumption, water usage, hardware specifications, and known limitations to facilitate independent evaluation.[57]

Practice:

- **Adopt Governance Frameworks:** Encourage organizations to adopt comprehensive AI governance frameworks, such as NIST AI RMF or ISO/IEC 42001, as a foundational structure for managing risks and embedding responsible practices.[6]
- **Embrace Socio-Technical Evaluation:** Integrate socio-technical perspectives and participatory methods into AI evaluation processes, ensuring diverse stakeholder engagement and contextual understanding.[9]
- **Document and Justify:** Implement rigorous documentation practices (Model Cards, Datasheets, audit logs) and transparently justify design choices, particularly those involving trade-offs between competing values.[20]
- **Foster AI Literacy:** Promote education and awareness about AI's capabilities, limitations, risks, and ethical implications among developers, deployers, policymakers, and the general public.[24]

By advancing on these fronts, the field can move towards more robust, meaningful, and holistic evaluations of AI systems, fostering the development and deployment of technologies that are not only powerful but also trustworthy and beneficial for all.

Appendices

List of Tools & Benchmarks Mentioned

Ethical AI - Fairness:

- ***AIF360 (AI Fairness 360):*** *IBM open-source toolkit with fairness metrics and bias mitigation algorithms.*
- ***Fairlearn:*** *Microsoft open-source toolkit for assessing and mitigating unfairness. Offers mitigation techniques like Exponentiated Gradient, Grid Search, and Threshold Optimizer.*
- ***What-If Tool:*** *Google tool for interactive visualization and analysis, including fairness evaluation.*
- ***TensorFlow Fairness Indicators:*** *Google tool providing capabilities for fairness evaluation.*
- ***OECD.AI Metrics Catalogue:*** *Includes metrics like 'Equal performance'.*

Ethical AI - Transparency & Explainability:

- ***LIME (Local Interpretable Model-agnostic Explanations):*** *Technique and tool for local model explanation.*
- ***SHAP (SHapley Additive exPlanations):*** *Technique and tool for local and global explanation based on Shapley values. Includes specialized versions like TreeSHAP, DeepExplainer/DeepSHAP, Expected Gradients.*
- ***InterpretML:*** *Part of Microsoft's Responsible AI Toolbox for interpretability.*
- ***Error Analysis:*** *Part of Microsoft's Responsible AI Toolbox.*
- ***Holistic AI Library:*** *Open-source library offering explainability tools.*
- ***Azure AI:*** *Microsoft platform offering explainability tools.*
- ***ELI5:*** *Framework for explanation.*
- ***Model Cards:*** *Standardized documentation format for model transparency.*
- ***Datasheets for Datasets:*** *Standardized documentation format for dataset transparency.*

Ethical AI - Accountability:

- ***Valohai:*** *MLOps platform mentioned for providing audit logs.*
- *(General Tools): Version control systems, tools for data/model/decision lineage tracking.*

Ethical AI - Privacy & Data Security:

- ***Diffprivlib:*** *IBM library for differential privacy.*

- ***Microsoft SEAL:*** *Library implementing Homomorphic Encryption (HE).*
- ***Google's DP libraries:*** *Libraries for differential privacy.*
- ***OpenFL:*** *Library for federated learning.*
- ***Microsoft Presidio:*** *Tool to aid data de-identification.*
- ***Microsoft Defender for Cloud:*** *Security tool mentioned.*
- ***Microsoft Counterfit:*** *Tool for security testing of AI systems.*

Ethical AI - Safety & Reliability:

- ***AutoAdvExBench:*** *Benchmark focusing on automated exploit generation against adversarial example defenses.*
- ***SAMURAI:*** *Proposed hardware-level monitoring technique using AI Performance Counters to detect adversarial inputs.*
- *(General Methods): Formal methods tools.*

Ethical AI - Potential for Misuse / Security:

- ***AI Red Teaming Tools/Methodologies:*** *General practice mentioned.*
- ***MITRE ATLAS (Adversarial Threat Landscape for AI Systems):*** *Framework/taxonomy of AI attacks.*
- ***OWASP AI Security Top 10:*** *Taxonomy of AI vulnerabilities.*
- ***garak:*** *Open-source tool for adversarial testing (LLM vulnerability scanner).*
- ***PyRIT (Python Risk Identification Toolkit):*** *Microsoft tool for AI red teaming.*
- ***ART (Adversarial Robustness Toolbox):*** *IBM toolbox for adversarial testing.*
- ***OpenAI Preparedness Framework (Beta):*** *Mentioned as an AI safety framework focusing on catastrophic risks.*

Environmental Impact:

- ***CodeCarbon:*** *Python package to estimate CO2 emissions from code execution.*
- ***ML CO2 Impact:*** *Web-based tool to estimate carbon footprint based on user inputs.*
- ***Eco2AI:*** *Python library for tracking CO2 emissions.*
- ***Cloud Provider Tools (AWS, Azure, GCP):*** *Dashboards/reports providing energy/carbon data.*
- ***NVIDIA SMI:*** *Hardware monitoring tool for GPU power usage.*
- ***Intel Power Gadget:*** *Hardware monitoring tool for CPU power usage.*
- ***AIWattch:*** *Browser extension mentioned for estimating LLM carbon emissions.*
- ***ElectricityMap:*** *API/source for location-specific grid carbon intensity data.*

- ***Life Cycle Assessment (LCA):*** *Comprehensive methodology (ISO 14040/14044).*

Accessibility & Equity:

- ***WCAG (Web Content Accessibility Guidelines):*** *Foundational standards (POUR principles, success criteria) adaptable for AI interfaces.*
- *(General Methods/Tools): Automated accessibility testing tools, usability testing with People with Disabilities (PWD).*

Language Support (NLP Benchmarks):

- ***GLUE (General Language Understanding Evaluation):*** *Benchmark suite.*
- ***XTREME:*** *Multilingual benchmark suite.*
- ***XNLI:*** *Cross-lingual Natural Language Inference benchmark.*
- ***TyDi QA:*** *Typologically Diverse Question Answering benchmark.*

General AI Benchmarks & Indices:

- ***MLPerf:*** *Industry benchmark suite for ML performance and energy efficiency.*
- ***Stanford AI Index:*** *Tracks trends in AI, including some related to responsible AI.*
- ***OECD AI Index:*** *Aims for a comprehensive measure of trustworthy AI.*

Summaries of Key Frameworks

1. NIST AI Risk Management Framework (AI RMF)

- ***Focus:*** *A voluntary framework developed in the US to help organizations manage AI risks and promote trustworthy AI.*
- ***Structure:*** *Organized around four core functions: Govern, Map, Measure, and Manage.*
- ***Key Aspects:*** *Emphasizes characteristics of trustworthy AI (e.g., fairness, safety, transparency). It promotes a socio-technical view and requires documentation, explanation, risk tolerance definition, and ensuring systems can fail safely. The 'Govern' function focuses on establishing organizational structures and accountability.*
- ***Nature:*** *Voluntary US government framework. Assessment is typically via self-assessment, potentially using Use-Case Profiles.*
- ***Strengths (as noted in doc):*** *Flexible, adaptable, comprehensive risk focus, promotes socio-technical view.*
- ***Limitations (as noted in doc):*** *Voluntary (non-binding), less prescriptive on specific metrics/controls compared to regulations.*

2. EU AI Act

- ***Focus:*** *Landmark comprehensive regulation establishing a risk-based legal framework for AI within the European Union.*
- ***Structure:*** *Categorizes AI systems into risk tiers: Prohibited, High-Risk, Limited Risk, Minimal Risk.*
- ***Key Aspects:*** *Imposes strict mandatory requirements for high-risk systems concerning data quality, documentation, transparency, human oversight, accuracy, robustness, and cybersecurity. Mandates conformity assessments and establishes significant penalties for non-compliance. Requires human oversight capabilities for high-risk systems. Mandates transparency for specific systems like chatbots and deepfakes.*
- ***Nature:*** *Mandatory EU Law. Enforcement involves conformity assessments (internal or third-party Notified Body), market surveillance, and fines.*
- ***Strengths (as noted in doc):*** *Legally binding, comprehensive, strong enforcement potential, sets a global precedent.*
- ***Limitations (as noted in doc):*** *Complex, potentially burdensome for innovation, definitions/scope may evolve.*

3. OECD AI Principles

- ***Focus:*** *An intergovernmental standard promoting AI that is innovative, trustworthy, respects human rights, and supports inclusive growth.*
- ***Structure:*** *Based on five value-based principles (inclusive growth, human-centered values/fairness, transparency/explainability, robustness/security/safety, accountability) and five policy recommendations.*
- ***Key Aspects:*** *Emphasizes human rights, transparency, robustness, safety, and accountability throughout the AI lifecycle.*
- ***Nature:*** *Intergovernmental standard; adherence is a voluntary commitment by signatory countries. Monitored via peer review, national strategy alignment, and the OECD.AI platform.*
- ***Strengths (as noted in doc):*** *Widely adopted, influential globally, balances innovation and values.*
- ***Limitations (as noted in doc):*** *High-level principles, lacks specific metrics or direct enforcement mechanisms.*

4. ISO/IEC 42001:2023

- ***Focus:*** *The first international standard specifying requirements for establishing, implementing, maintaining, and continually improving an AI Management System (AIMS) within an organization.*
- ***Structure:*** *Provides a framework following the Plan-Do-Check-Act cycle, similar to other ISO management systems (e.g., ISO 27001). Includes specific controls in Annex A.*
- ***Key Aspects:*** *Mandates processes for accountability, risk management, AI impact assessment, system lifecycle management, data management, and transparency. Requires documentation as part of the AIMS.*
- ***Nature:*** *International standard; compliance is demonstrated through voluntary third-party certification audits.*
- ***Strengths (as noted in doc):*** *Structured, auditable, integrates with other ISO standards, provides practical controls.*
- ***Limitations (as noted in doc):*** *Voluntary, involves cost/effort for implementation and certification, may need supplementing with domain-specific standards.*

5. IEEE P7000™ Series / Ethically Aligned Design (EAD)

- ***Focus:*** *A family of standards addressing specific technical and ethical considerations in system design.*
- ***Structure:*** *Comprises multiple individual standards (P7000-P7010+) covering areas like overall ethical process (7000), transparency (7001), data privacy*

(P7002), algorithmic bias (7003), child/student data (P7004), fail-safe design (P7008), ethical nudging (P7009), and wellbeing metrics (P7010). EAD provides the overarching ethical framework.

- ***Key Aspects:*** *Emphasizes accountability, transparency, safety, and privacy through specific design guidance.*
- ***Nature:*** *Voluntary standards. Conformance can be assessed through testing and IEEE's CertifAIEd™ program.*
- ***Strengths (as noted in doc):*** *Deep focus on specific technical/ethical areas, provides practical guidance.*
- ***Limitations (as noted in doc):*** *Fragmented across multiple standards, adoption level varies.*

6. UNESCO Recommendation on the Ethics of AI

- ***Focus:*** *A global standard providing ethical principles and policy recommendations for AI, adopted by UNESCO member states.*
- ***Structure:*** *Based on four core values (Human Rights & Dignity, Environmental Flourishing, Diversity & Inclusiveness, Peaceful Societies) and ten principles (including Proportionality, Safety/Security, Fairness, Transparency/Explainability, Responsibility/Accountability, Sustainability, Privacy, Human Oversight).*
- ***Key Aspects:*** *Strong emphasis on human rights, diversity, and environmental considerations.*
- ***Nature:*** *International Recommendation; relies on member state commitment for implementation. Supported by assessment tools like the Readiness Assessment Methodology (RAM) and an Ethical Impact Assessment (EIA) tool.*
- ***Strengths (as noted in doc):*** *Global scope, strong human rights focus, emphasizes inclusivity and environment.*
- ***Limitations (as noted in doc):*** *High-level principles, relies on member state implementation, lacks direct enforcement.*

7. Blueprint for an AI Bill of Rights (US White House)

- ***Focus:*** *Outlines five principles intended to guide the design, use, and deployment of automated systems to protect American rights.*
- ***Structure:*** *Built around five core principles: Safe and Effective Systems; Algorithmic Discrimination Protections; Data Privacy; Notice and Explanation; Human Alternatives, Consideration, and Fallback.*
- ***Key Aspects:*** *Emphasizes safety, non-discrimination, data privacy, transparency (notice and explanation), and the availability of human recourse.*

- ***Nature:*** *Non-binding principles and practices issued by the White House Office of Science and Technology Policy; not legislation. It serves as guidance for government agencies and a call to action for the private sector.*

Glossary

A

- ***Accountability:*** *The obligation of AI actors (developers, deployers, operators) to take responsibility for the proper functioning and outcomes of AI systems, based on their roles and context. It involves being answerable for AI decisions and impacts, particularly when harm occurs.*
- ***Accessibility:*** *Ensuring ease of use for diverse populations, including people with disabilities. In AI, this involves considerations like content adaptability, multimodal interaction, contextual awareness, assistive technology compatibility, and time sufficiency.*
- ***Adversarial Robustness Testing:*** *Evaluating an AI system's resilience against inputs intentionally crafted to deceive or cause failure (e.g., perturbed images, malicious prompts).*
- ***AIF360 (AI Fairness 360):*** *An IBM open-source toolkit providing a comprehensive library of fairness metrics and bias mitigation algorithms.*
- ***AI Indices/Benchmarks:*** *Tools like the Stanford AI Index or OECD AI Index that track trends or aim to measure aspects of AI, including trustworthiness or performance efficiency. However, comprehensive non-performance benchmarking faces significant challenges.*
- ***AI Literacy:*** *Understanding how AI works, its capabilities and limitations, and the ability to effectively use AI tools.*
- ***AI Management System (AIMS):*** *A framework for establishing, implementing, maintaining, and improving AI governance within an organization, as defined by standards like ISO/IEC 42001.*
- ***AI Red Teaming:*** *A practice where dedicated teams simulate attacks by malicious actors to identify vulnerabilities, test limits, and uncover potential misuse scenarios in AI systems.*
- ***Algorithmic Fairness:*** *Ensuring equitable treatment and outcomes for individuals and groups from AI systems, irrespective of protected attributes, and actively mitigating harmful biases.*
- ***Anonymization/Pseudonymization:*** *Privacy-enhancing techniques involving removing or replacing personally identifiable information (PII) from data.*
- ***Audit Trails/Logging:*** *Maintaining detailed, immutable records of system operations, data usage, model changes, user interactions, and decisions made by an AI system, crucial for traceability and accountability.*

B

- ***Bias (in AI):*** *Can manifest in various forms: systemic (reflecting societal inequalities), computational/statistical (arising from data or algorithms), and human-cognitive (introduced by developers or users). Mitigation aims to prevent unjustified adverse effects.*
- ***BLEU (Bilingual Evaluation Understudy):*** *A metric commonly used to evaluate machine translation quality by comparing n-gram overlap between machine output and reference translations.*

C

- ***Capability Approach:*** *A theoretical framework evaluating well-being and justice based on individuals' substantive freedoms ('capabilities') to achieve valued functionings, assessing how AI enables or hinders these.*
- ***Carbon Emissions:*** *Greenhouse gas (GHG) emissions associated with AI, measured in CO2 equivalent (CO2e), primarily resulting from energy consumption during training, inference, and hardware manufacturing. Calculated by multiplying energy use by carbon intensity.*
- ***Carbon Footprint:*** *The total amount of greenhouse gases generated by an AI system or its components, often assessed across its lifecycle using LCA.*
- ***Carbon Intensity:*** *The amount of GHG emissions per unit of electricity generated (e.g., kgCO2e/kWh), varying significantly by region and energy source.*
- ***CCI (Compute Carbon Intensity):*** *A proposed metric measuring the lifetime carbon emissions (including manufacturing) of AI hardware per unit of computation (e.g., gCO2e/Exa-FLOP).*
- ***CodeCarbon:*** *A Python package that estimates CO2 emissions from code execution by tracking hardware power consumption and using location-specific carbon intensity data.*
- ***Community Engagement:*** *Involving diverse stakeholders (users, communities, experts) in AI development and deployment to identify harms, align with values, build trust, and promote equity.*
- ***Compliance:*** *Adhering to relevant laws, regulations (e.g., EU AI Act, GDPR), and standards (e.g., ISO/IEC 42001, IEEE P7000) governing AI development and deployment.*
- ***Conformity Assessment:*** *The process of verifying whether an AI system meets specified requirements, particularly mandated for high-risk systems*

under the EU AI Act, involving self-assessment or third-party evaluation.

- ***Context-Dependency:*** *The characteristic where the appropriateness or definition of non-performance aspects like fairness, risk, or ethics depends heavily on the specific application, cultural setting, user group, and deployment environment.*
- ***Copyright (in AI):*** *Legal rights concerning the creation and use of works. Key issues include the copyrightability of AI-generated output (often requiring human authorship) and potential infringement from using copyrighted data for training models.*
- ***Cost of Ownership (Total Cost of Ownership - TCO):*** *A comprehensive assessment of all costs associated with an AI system, including development, data, hardware, integration, maintenance, compute, and compliance.*
- ***Cross-Entropy Loss:*** *A metric used in language model training that measures the difference between the model's predicted probability distribution and the actual distribution of tokens.*

D

- ***Data Governance:*** *The overall management of data availability, usability, integrity, and security in an enterprise, critical for responsible AI. Includes policies for the data lifecycle, quality management, security, and privacy.*
- ***Data Lineage:*** *Tracking the origin, transformations, and usage of data throughout the AI lifecycle, essential for transparency and accountability.*
- ***Data Minimization:*** *A privacy principle dictating that only data strictly necessary for a specific purpose should be collected.*
- ***Datasheets for Datasets:*** *Standardized documentation detailing a dataset's motivation, composition, collection, preprocessing, uses, distribution, and maintenance to increase transparency.*
- ***Deepfakes:*** *Convincing synthetic media (images, audio, video) generated by AI, often used to spread disinformation or for malicious purposes.*
- ***Demographic Parity (Statistical Parity):*** *A group fairness metric requiring the likelihood of a positive outcome to be equal across different demographic groups.*
- ***Differential Privacy (DP):*** *A mathematical framework providing quantifiable privacy guarantees by adding calibrated noise to data or outputs, ensuring statistical similarity regardless of individual data inclusion. Key parameters are epsilon (ε) and delta (δ).*
- ***Digital Divide:*** *The gap between individuals, communities, or regions regarding access to, use of, and knowledge of ICTs, including AI. Encompasses*

infrastructure, skills, affordability, and meaningful use.

- ***Disinformation:** False or misleading information spread deliberately, often amplified by AI's ability to generate convincing fake content at scale.*

E

- ***E-waste (Electronic Waste):** Discarded electronic devices, a growing problem exacerbated by the rapid hardware cycles in AI, containing toxic substances and representing lost resources.*
- ***Ecological Footprint:** The overall environmental impact of AI, encompassing energy, carbon, water, material consumption, and waste generation.*
- ***Embodied Impacts/Emissions:** Environmental impacts (e.g., carbon emissions, water use) associated with the manufacturing, transportation, and disposal phases of a product's lifecycle, as opposed to its operational use phase.*
- ***Energy Consumption:** The amount of electricity used by AI systems, particularly during training and inference, measured in kWh or MWh.*
- ***Environmental Impact Assessment (EIA):** A formal process for evaluating the likely environmental impacts of a proposed project or development. Related to SIA.*
- ***Equal Opportunity:** A group fairness metric requiring the true positive rate (sensitivity) of a classifier to be equal across different demographic groups.*
-
- ***Equalized Odds:** A group fairness metric requiring both the true positive rate and the false positive rate to be equal across different demographic groups.*
- ***Equity (in AI):** Fairness and justice in how AI systems are developed, deployed, and impact different individuals and groups, considering factors like accessibility, bias, and the digital divide.*
- ***Ethical Considerations:** Examining the alignment of AI systems with moral values and human rights, encompassing fairness, transparency, accountability, privacy, safety, and potential for misuse.*
- ***EU AI Act:** Landmark EU regulation establishing a risk-based legal framework for AI, imposing strict requirements on high-risk systems and prohibiting certain uses.*
- ***Explainability (XAI):** The ability to explain the internal mechanisms or logic driving an AI's decisions or predictions in understandable terms.*
- ***Explainable AI (XAI):** Field focused on developing methods and techniques to make AI decisions understandable to humans.*

F

- ***F1 Score:*** *A standard metric in classification tasks (like NER) that combines precision and recall into a single score.*
- ***Fail-Safe Design:*** *Incorporating mechanisms into an AI system to ensure it can be safely controlled, overridden, or shut down if it behaves undesirably or risks harm.*
- ***Fairlearn:*** *A Microsoft open-source toolkit for assessing and mitigating unfairness in machine learning models.*
- ***Fairness (in AI):*** *See Algorithmic Fairness. Definitions are often context-specific and can involve trade-offs between different mathematical formulations.*
- ***Federated Learning (FL):*** *A machine learning approach where models are trained locally on decentralized devices without centralizing raw data, sharing only model updates to preserve privacy.*
- ***Formal Methods:*** *Mathematical techniques used to specify and verify system properties, aiming to provide guarantees of behavior within certain bounds.*
- ***Fundamental Rights Impact Assessment (FRIA):*** *See Human Rights Impact Assessment (HRIA).*

G

- ***GDPR (General Data Protection Regulation):*** *EU regulation governing the processing of personal data, applicable to AI systems handling data of EU residents.*
- ***Governance (AI):*** *The structures, policies, processes, and controls established to direct, manage, and monitor the development and deployment of AI responsibly and ethically.*
- ***Green AI:*** *A field of research and practice focused on measuring, understanding, and mitigating the environmental impact of artificial intelligence.*

H

- ***Hardware Sustainability:*** *Addressing the environmental impact of AI hardware (GPUs, TPUs, etc.) through principles like circularity, focusing on material use, recycling, reuse, and lifespan extension.*
- ***Homomorphic Encryption (HE):*** *A privacy-enhancing technology that allows computation to be performed directly on encrypted data without needing to decrypt it first.*

- ***Human Oversight:*** *The capability for humans to monitor, understand, intervene in, and ultimately override AI system decisions, particularly crucial for high-risk applications. Mandated by regulations like the EU AI Act.*
- ***Human Rights Impact Assessment (HRIA) / Fundamental Rights Impact Assessment (FRIA):*** *Methodologies adapting legal and ethical human rights frameworks to systematically identify, assess, and mitigate risks posed by AI systems to fundamental rights.*

I

- ***IEEE P7000™ Series:*** *A family of standards from IEEE addressing specific ethical considerations in system design, including transparency, bias, privacy, and fail-safe design.*
- ***Individual Fairness:*** *A fairness concept focusing on treating similar individuals similarly, as opposed to comparing outcomes across groups.*
- ***Intellectual Property (IP):*** *Legal rights protecting creations of the mind, such as inventions (patents) and literary/artistic works (copyright). AI challenges traditional IP notions of authorship and inventorship.*
- ***Interpretability:*** *The degree to which a human can understand the cause of an AI's decision or prediction, or consistently predict its output. Closely related to Explainability.*
- ***ISO/IEC 42001:*** *The first international standard specifying requirements for establishing, implementing, maintaining, and continually improving an AI Management System (AIMS) within an organization.*

L

- ***LCA (Life Cycle Assessment):*** *A systematic framework for evaluating the environmental impacts of a product, process, or service throughout its entire lifespan, from raw material extraction to disposal.*
- ***LIME (Local Interpretable Model-agnostic Explanations):*** *An XAI technique that explains individual predictions of complex models by approximating them locally with simpler, interpretable models.*
- ***LLM (Large Language Model):*** *A type of AI model trained on vast amounts of text data, capable of understanding and generating human-like language[cite: 288].*

M

- ***MADM (Multi-Attribute Decision Making):*** *A family of techniques used to* *evaluate and compare alternatives based on multiple, often conflicting, criteria. Can incorporate quantitative and qualitative data and weighting.*

- ***Material Flow Analysis (MFA):*** *A method used to track the flows of materials through systems, relevant for assessing resource efficiency and circularity, particularly for AI hardware.*
- ***METEOR (Metric for Evaluation of Translation with Explicit ORdering):*** *A machine translation evaluation metric that considers precision, recall, stemming, synonymy, and word order.*
- ***Misuse Potential:*** *The possibility that an AI system could be intentionally used for harmful or unethical purposes. Evaluated through methods like AI Red Teaming.*
- ***MLPerf:*** *An industry benchmark suite for measuring the performance (and increasingly, energy efficiency) of machine learning hardware and software.*
- ***Model Card:*** *Standardized documentation providing a concise summary of an AI model's characteristics, including intended use, performance, limitations, training data, and ethical considerations.*
- ***Model Lineage:*** *Documenting the versions, training procedures, parameters, and updates of AI models, contributing to traceability.*

N

- ***NIST AI RMF (Risk Management Framework):*** *A voluntary framework from the US National Institute of Standards and Technology to help organizations manage AI risks and promote trustworthy AI, organized around Govern, Map, Measure, and Manage functions.*

O

- ***OECD AI Principles:*** *Influential intergovernmental principles promoting innovative and trustworthy AI based on values like human-centeredness, fairness, transparency, robustness, and accountability.*
- ***Operational Impacts/Emissions:*** *Environmental impacts occurring during the use phase of an AI system (e.g., energy consumed during training or inference), distinct from embodied impacts.*

P

- ***Participatory Design:*** *An approach to design that actively involves relevant stakeholders (e.g., users, community members) in the design process.*
- ***Perplexity (PPL):*** *A metric used in language modeling to measure how well a*

model predicts a sequence of text; lower perplexity indicates better performance.

- ***PETs (Privacy-Enhancing Technologies):*** *Technologies designed to protect personal data, such as Differential Privacy, Homomorphic Encryption, and Secure Multi-Party Computation.*
- ***POUR Principles:*** *The four core principles of web accessibility defined by WCAG: Perceivable, Operable, Understandable, Robust.*
- ***Power Usage Effectiveness (PUE):*** *A standard data center efficiency metric calculated as Total Facility Energy divided by IT Equipment Energy; a value closer to 1 indicates higher efficiency.*
- ***Privacy:*** *Protecting personal information from unauthorized access, use, or disclosure. In AI, involves data security, PETs, compliance with regulations like GDPR, and principles like data minimization.*

Q

- ***Quantification:*** *The process of converting qualitative concepts or assessments into numerical values or metrics, a key challenge in evaluating many non-performance aspects of AI.*

R

- ***Reliability:*** *The ability of an AI system to consistently function as intended under specified conditions over a given period, without failure.*
- ***Resilience:*** *The ability of an AI system to withstand attacks or disruptions and maintain function or recover quickly.*
- ***Responsibility (in AI Governance):*** *Evaluating the practices of AI developers and deployers, including organizational transparency, community engagement, regulatory compliance, and handling of IP/data.*
- ***Return on Investment (ROI):*** *A financial metric comparing the benefits derived from an investment (like an AI system) against its cost.*
- ***Robustness:*** *The ability of an AI system to maintain its level of performance even under adverse conditions, such as noisy inputs or adversarial attacks.*
- ***ROUGE (Recall-Oriented Understudy for Gisting Evaluation):*** *A metric primarily used for evaluating text summarization (and sometimes translation) based on overlapping units like n-grams or sequences.*

S

- ***Safety:*** *The absence of conditions in an AI system that could endanger human* *life, health, property, or the environment.*

- ***Secure Development Lifecycle (SDLC):** Integrating security practices throughout the entire process of software (including AI) development.*
- ***SHAP (SHapley Additive exPlanations):** An XAI technique based on game theory (Shapley values) that explains model predictions by attributing the contribution of each feature.*
- ***SIA (Social Impact Assessment):** A methodology used to analyze, monitor, and manage the intended and unintended social consequences of planned interventions, adaptable for assessing AI impacts.*
- ***Socio-Technical System:** A system viewed as comprising both social components (people, organizations, norms) and technical components (hardware, software), where these interact and influence each other. AI systems are inherently socio-technical.*
- ***Stakeholder:** Any individual, group, or organization that can affect or is affected by an AI system.*

T

- ***TCO (Total Cost of Ownership):** See Cost of Ownership.*
- ***Threat Modeling:** A systematic process to identify potential threats, vulnerabilities, and attack vectors relevant to an AI system.*
- ***Traceability:** The ability to reconstruct the lifecycle and decision-making process of an AI system, including data origins, model versions, and operational logs. Crucial for accountability.*
- ***Trade-offs:** Inherent conflicts between different desirable properties in AI systems, such as accuracy vs. fairness, privacy vs. transparency, or performance vs. environmental impact. Managing these is a core challenge.*
- ***Transparency:** The availability of appropriate information about an AI system, its capabilities, limitations, data, and outputs, tailored to the stakeholder and context. A key principle in AI governance, often enabled by documentation (Model Cards, Datasheets) and explainability techniques.*
- ***Trustworthy AI:** AI systems that are lawful, ethical, and technically robust, encompassing characteristics like validity, reliability, safety, fairness, transparency, accountability, and privacy.*

U

- ***Usability Testing:** Evaluating how easy and effective a system is to use by observing representative users performing tasks. Essential for assessing* *accessibility, often involving people with disabilities.*

V

- ***Validation & Verification (V&V):*** *Processes to confirm that an AI system meets its specified requirements (validation) and is built correctly according to its design (verification).*

W

- ***Water Footprint:*** *The total volume of freshwater used directly or indirectly by an AI system, encompassing operational use (e.g., data center cooling) and embodied use (e.g., manufacturing, electricity generation).*
- ***Water Usage Effectiveness (WUE):*** *A data center efficiency metric calculated as annual water consumption (liters) divided by IT equipment energy consumption (kWh); lower is better.*
- ***WCAG (Web Content Accessibility Guidelines):*** *Widely recognized international standards for making web content accessible to people with disabilities, adaptable for evaluating AI interfaces.*
- ***What-If Tool:*** *A Google tool for interactive visualization and analysis of ML models, including fairness evaluation.*

X

- ***XAI (Explainable AI):*** *See Explainable AI.*

Works cited

1. Artificial intelligence - OECD, accessed April 29, 2025, https://www.oecd.org/en/topics/artificial-intelligence.html
2. White Papers 2024 Understanding the EU AI Act - ISACA, accessed April 29, 2025, https://www.isaca.org/resources/white-papers/2024/understanding-the-eu-ai-act
3. IEEE Standards Association Statement of Intention Our Role in Addressing Ethical Considerations of Autonomous and Intelligent Systems (A/IS), accessed April 29, 2025, https://standards.ieee.org/wp-content/uploads/import/documents/other/ethical-considerations-ai-as-29mar2018.pdf
4. Evaluating the Societal Impact of AI: A Comparative Analysis of Human and AI Platforms Using the Analytic Hierarchy Process - MDPI, accessed April 29, 2025, https://www.mdpi.com/2673-2688/6/4/86
5. Artificial Intelligence Risk Management Framework (AI RMF 1.0) - NIST Technical Series Publications, accessed April 29, 2025, https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
6. NIST AI Risk Management Framework: The Ultimate Guide - Hyperproof, accessed April 29, 2025, https://hyperproof.io/navigating-the-nist-ai-risk-management-framework/
7. What is Trustworthy AI? - IBM, accessed April 29, 2025, https://www.ibm.com/think/topics/trustworthy-ai
8. Operationalizing the Measure Function of the NIST AI Risk ..., accessed April 29, 2025, https://trails.gwu.edu/operationalizing-measure-function-nist-ai-risk-management-framework
9. Operationalizing the NIST AI RMF Framework - Carnegie Mellon University, accessed April 29, 2025, https://www.cmu.edu/block-center/responsible-ai/cmu_blockcenter_operationalizing-the-nist-ai-rmf-framework-fin.pdf
10. AI principles - OECD, accessed April 29, 2025, https://www.oecd.org/en/topics/ai-principles.html
11. How to Assess Trustworthy AI in Practice. - arXiv, accessed April 29, 2025, https://arxiv.org/pdf/2206.09887
12. AI Risk Management Framework - Palo Alto Networks, accessed April 29, 2025, https://www.paloaltonetworks.com/cyberpedia/ai-risk-management-framework
13. Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach - | sinbad2, accessed April 29, 2025, https://sinbad2.ujaen.es/sites/default/files/publications/published_6.pdf
14. A Taxonomy of Trustworthiness for Artificial Intelligence - CLTC Berkeley, accessed April 29, 2025, https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf
15. The Benchmark Trap: Why AI's Favorite Metrics Might Be Misleading Us -

VKTR.com, accessed April 29, 2025, https://www.vktr.com/ai-market/the-benchmark-trap-why-ais-favorite-metrics-might-be-misleading-us/

16. Ethical AI in Social Sciences Research: Are We Gatekeepers or Revolutionaries? - MDPI, accessed April 29, 2025, https://www.mdpi.com/2075-4698/15/3/62
17. Fairness Metrics in Machine Learning - Coralogix, accessed April 29, 2025, https://coralogix.com/ai-blog/fairness-metrics-in-machine-learning/
18. Best Practices in Data Governance for Organizations - Innovapte, accessed April 29, 2025, https://innovapte.com/blog/best-practices-in-data-governance-for-organizations/
19. What is AI traceability? Benefits, tools & best practices - Data.world, accessed April 29, 2025, https://data.world/blog/what-is-ai-traceability-benefits-tools-best-practices/
20. Model Cards for AI Model Transparency - Salesforce, accessed April 29, 2025, https://www.salesforce.com/blog/model-cards-for-ai-model-transparency/
21. Measuring AI's Environmental Impact: Technical Bootcamp by Pascal Joly on Maven, accessed April 29, 2025, https://maven.com/it-climate-ed/measuring-the-carbon-footprint-of-ai
22. What Is Water Usage Effectiveness (WUE) in Data Centers? - The Equinix Blog, accessed April 29, 2025, https://blog.equinix.com/blog/2024/11/13/what-is-water-usage-effectiveness-wue-in-data-centers/
23. Climate And Resource Awareness is Imperative to Achieving Sustainable AI (and Preventing a Global AI Arms Race) - arXiv, accessed April 29, 2025, https://arxiv.org/html/2502.20016v1
24. AI for the Global Majority: The Digital Divide No One's Talking About!, accessed April 29, 2025, https://globaldigitalinclusion.org/2025/02/21/ai-for-the-global-majority-the-digital-divide-no-ones-talking-about/
25. AI-generated content and IP rights: Challenges and policy considerations - Diplo, accessed April 29, 2025, https://www.diplomacy.edu/blog/ai-generated-content-and-ip-rights-challenges-and-policy-considerations/
26. Artificial Intelligence and the Creative Double Bind - Harvard Law Review, accessed April 29, 2025, https://harvardlawreview.org/print/vol-138/artificial-intelligence-and-the-creative-double-bind/
27. Building AI Responsibly - AWS, accessed April 29, 2025, https://aws.amazon.com/ai/responsible-ai/
28. Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review, accessed April 29, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11024755/
29. AI Ethics Part Two: AI Framework Best Practices - Alvarez & Marsal, accessed April 29, 2025,

https://www.alvarezandmarsal.com/insights/ai-ethics-part-two-ai-framework-best-practices

30. Privacy and responsible AI - IAPP, accessed April 29, 2025, https://iapp.org/news/a/privacy-and-responsible-ai
31. Quantitative and Qualitative AI Ethics Lab (QQAEL) - Morgan State University, accessed April 29, 2025, https://www.morgan.edu/ceamls/research/labs/quantitative-and-qualitative-ai-ethics-lab-(qqael)
32. Getting Started — aif360 0.6.1 documentation, accessed April 29, 2025, https://aif360.readthedocs.io/en/stable/Getting%20Started.html
33. aif360.metrics.ClassificationMetric - Read the Docs, accessed April 29, 2025, https://aif360.readthedocs.io/en/latest/modules/generated/aif360.metrics.ClassificationMetric.html
34. Common fairness metrics — Fairlearn 0.13.0.dev0 documentation, accessed April 29, 2025, https://fairlearn.org/main/user_guide/assessment/common_fairness_metrics.html
35. Metrics for Trustworthy AI - OECD.AI, accessed April 29, 2025, https://oecd.ai/en/catalogue/metrics
36. [1810.01943] 1 Introduction - ar5iv - arXiv, accessed April 29, 2025, https://ar5iv.labs.arxiv.org/html/1810.01943
37. Essential Open-Source Tools for Bias Detection and Mitigation - Turing Post, accessed April 29, 2025, https://www.turingpost.com/p/ai-fairness-tools
38. Mitigation — Fairlearn 0.8.0 documentation, accessed April 29, 2025, https://fairlearn.org/v0.8/user_guide/mitigation.html
39. fairlearn/fairlearn: A Python package to assess and improve fairness of machine learning models. - GitHub, accessed April 29, 2025, https://github.com/fairlearn/fairlearn
40. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence - NIST Technical Series Publications, accessed April 29, 2025, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf
41. Article 43: Conformity Assessment | EU AI Act - Securiti, accessed April 29, 2025, https://securiti.ai/eu-ai-act/article-43/
42. Full article: AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development, accessed April 29, 2025, https://www.tandfonline.com/doi/full/10.1080/08839514.2025.2463722
43. Safeguard the Future of AI: The Core Functions of the NIST AI RMF | AuditBoard, accessed April 29, 2025, https://www.auditboard.com/blog/nist-ai-rmf/
44. Trustworthy XAI and Its Applications - arXiv, accessed April 29, 2025, https://arxiv.org/html/2410.17139v2
45. LIME: Local Interpretable Model-Agnostic Explanations - C3 AI, accessed April 29, 2025, https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/
46. Enhancing Transparency in AI: Explainability Metrics via LIME with Holistic AI Library, accessed April 29, 2025,

https://www.holisticai.com/blog/transparency-in-ai-explainability-metrics

47. Techniques for Explainable AI: LIME and SHAP - Unnat Bak (Founder @ Revscale, TABS Suite) Growth Hacking and Venture Advisory, accessed April 29, 2025, https://www.unnatbak.com/blog/techniques-for-explainable-ai-lime-and-shap
48. Exploring Explainable AI with LIME Technology - Steadforce, accessed April 29, 2025, https://www.steadforce.com/blog/explainable-ai-with-lime
49. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME - arXiv, accessed April 29, 2025, https://arxiv.org/html/2305.02012v3
50. Explainability In Machine Learning: Top Techniques - Arize AI, accessed April 29, 2025, https://arize.com/blog-course/explainability-techniques-shap/
51. Enhancing Transparency in AI: Explainability Metrics via SHAP Feature Importance with Holistic AI Open-Source Library, accessed April 29, 2025, https://www.holisticai.com/blog/enhancing-ai-transparency-shap
52. How to interpret machine learning (ML) models with SHAP values – Mage AI Blog, accessed April 29, 2025, https://www.mage.ai/blog/how-to-interpret-explain-machine-learning-models-using-shap-values
53. Responsible AI Tools and Practices | Microsoft AI, accessed April 29, 2025, https://www.microsoft.com/en-us/ai/tools-practices
54. Evaluation and monitoring metrics for generative AI - Azure AI Foundry | Microsoft Learn, accessed April 29, 2025, https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/evaluation-metrics-built-in
55. Google Model Cards, accessed April 29, 2025, https://modelcards.withgoogle.com/
56. Responsible AI: The Role of Data and Model Cards | Datatonic, accessed April 29, 2025, https://datatonic.com/insights/responsible-ai-data-model-cards/
57. PROCEEDINGS: Towards Standards for Data Transparency for AI Models - The White House, accessed April 29, 2025, https://ai.gov/wp-content/uploads/2024/06/PROCEEDINGS_Towards-Standards-for-Data-Transparency-for-AI-Models.pdf
58. Model cards: elevating AI transparency in recruitment by adapting a Google invention, accessed April 29, 2025, https://sapia.ai/resources/blog/how-sapia-labs-lifted-the-bar-on-ai-transparency-in-recruitment/
59. Datasheets for Earth Science Datasets - AMS Journals, accessed April 29, 2025, https://journals.ametsoc.org/view/journals/bams/106/4/BAMS-D-24-0203.1.pdf
60. Datasheet for dataset template - Overleaf, Online LaTeX Editor, accessed April 29, 2025, https://www.overleaf.com/latex/templates/datasheet-for-dataset-template/jgqyyzyprxth
61. Data title, accessed April 29, 2025, https://cdn.serc.carleton.edu/files/usingdata/accessdata/documents/datasheet_template_03_07.doc
62. AudreyBeard/Datasheets-for-Datasets-Template - GitHub, accessed April 29,

2025, https://github.com/AudreyBeard/Datasheets-for-Datasets-Template
63. [Discussion]:"Datasheets for Datasets" paper and its relevance in the real world - Reddit, accessed April 29, 2025, https://www.reddit.com/r/MachineLearning/comments/1bdbuef/discussiondatasheets_for_datasets_paper_and_its/
64. Datasheets for Datasets - Communications of the ACM, accessed April 29, 2025, https://cacm.acm.org/research/datasheets-for-datasets/
65. EU AI Act Compliance Checker | EU Artificial Intelligence Act, accessed April 29, 2025, https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/
66. Standardization for Compliance in the European Union's AI Act - WilmerHale, accessed April 29, 2025, https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20241204-standardization-for-compliance-in-the-european-unions-ai-act
67. Autonomous and Intelligent Systems (AIS) - IEEE Standards Association, accessed April 29, 2025, https://standards.ieee.org/initiatives/autonomous-intelligence-systems/
68. Standards and Certifications for Ethical AI and Autonomous Systems - Capella Alliance, accessed April 29, 2025, https://capella.nisum.com/standards-and-certifications-for-ethical-ai-and-autonomous-systems/
69. IEEE P7000™ Projects - OCEANIS, accessed April 29, 2025, https://ethicsstandards.org/p7000/
70. OECD AI Principles overview, accessed April 29, 2025, https://oecd.ai/en/ai-principles
71. Empowering responsible AI practices | Microsoft AI, accessed April 29, 2025, https://www.microsoft.com/en-us/ai/responsible-ai
72. Responsible AI - AMD, accessed April 29, 2025, https://www.amd.com/en/solutions/ai/responsible-ai.html
73. From audit trails to accountability: how traceability transforms compliance - FinTech Global, accessed April 29, 2025, https://fintech.global/2025/04/16/from-audit-trails-to-accountability-how-traceability-transforms-compliance/
74. Valohai's Audit Log: Traceability built for AI governance, accessed April 29, 2025, https://valohai.com/blog/valohai-audit-log-traceability-built-for-ai-governance/
75. How to Test for Compliance with Human Oversight Requirements in AI Regulation? - arXiv, accessed April 29, 2025, https://arxiv.org/html/2504.03300v1
76. Top 9 AI Data Governance Best Practices for Security, Compliance, and Quality, accessed April 29, 2025, https://www.pmi.org/blog/ai-data-governance-best-practices
77. Article 99: Penalties | EU Artificial Intelligence Act, accessed April 29, 2025, https://artificialintelligenceact.eu/article/99/
78. Conformity Assessments in the EU AI Act: What You Need to Know - Holistic AI, accessed April 29, 2025,

https://www.holisticai.com/blog/conformity-assessments-in-the-eu-ai-act
79. EU AI Act Compliance: Complete Guide for AI in Sales & Support 2024 - Qualimero, accessed April 29, 2025, https://www.qualimero.com/en/blog/ai-act-compliance-guide
80. ISO/IEC 42001: The latest AI management system standard - KPMG International, accessed April 29, 2025, https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html
81. Understanding ISO 42001 and Demonstrating Compliance - ISMS.online, accessed April 29, 2025, https://www.isms.online/iso-42001/
82. IEEE Ethically Aligned Design - Palo Alto Networks, accessed April 29, 2025, https://www.paloaltonetworks.com/cyberpedia/ieee-ethically-aligned-design
83. Traceability - future ai, accessed April 29, 2025, https://future-ai.eu/principle/traceability/
84. Differential Privacy in AI: What it is and Why it Matters? - ClanX, accessed April 29, 2025, https://clanx.ai/glossary/differential-privacy-in-ai
85. Differential Privacy in Responsible AI - FlippingBook, accessed April 29, 2025, https://online.flippingbook.com/view/913296844
86. Differential Privacy and Risk Metrics, accessed April 29, 2025, https://privacy-analytics.com/resources/articles/differential-privacy-and-risk-metrics/
87. Differential privacy in machine learning (preview), accessed April 29, 2025, https://docs.azure.cn/en-us/machine-learning/concept-differential-privacy?view=azureml-api-2
88. Empirical Analysis of Privacy-Fairness-Accuracy Trade-offs in Federated Learning: A Step Towards Responsible AI - arXiv, accessed April 29, 2025, https://arxiv.org/html/2503.16233v1
89. NIST AI Risk Management Framework (AI RMF) - Palo Alto Networks, accessed April 29, 2025, https://www.paloaltonetworks.com/cyberpedia/nist-ai-risk-management-framework
90. Understanding the NIST AI RMF: What It Is and How to Put It Into Practice - Secureframe, accessed April 29, 2025, https://secureframe.com/blog/nist-ai-rmf
91. A concise EU AI Act summary for in-scope organizations - Vanta, accessed April 29, 2025, https://www.vanta.com/resources/eu-ai-act-guide
92. NIST AI Risk Management Framework 1.0: Meaning, challenges, implementation, accessed April 29, 2025, https://www.scrut.io/post/nist-ai-risk-management-framework
93. [2504.07887] Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge - arXiv, accessed April 29, 2025, https://arxiv.org/abs/2504.07887
94. Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge - arXiv, accessed April 29, 2025, https://arxiv.org/html/2504.07887v1
95. AutoAdvExBench: Benchmarking autonomous exploitation of adversarial example defenses, accessed April 29, 2025, https://arxiv.org/html/2503.01811v1

96. Runtime Detection of Adversarial Attacks in AI Accelerators Using Performance Counters, accessed April 29, 2025, https://arxiv.org/html/2503.07568v1
97. ISO/IEC 42001: AI management systems - Regulatory knowledge for medical devices, accessed April 29, 2025, https://blog.johner-institute.com/quality-management-iso-13485/iso-iec-42001/
98. Ultimate Guide to Human Oversight in AI Workflows - Magai, accessed April 29, 2025, https://magai.co/guide-to-human-oversight-in-ai-workflows/
99. The crucial role of humans in AI oversight - Cornerstone OnDemand, accessed April 29, 2025, https://www.cornerstoneondemand.com/resources/article/the-crucial-role-of-humans-in-ai-oversight/
100. The Strategic Necessity of Human Oversight in AI Systems - Lumenova AI, accessed April 29, 2025, https://www.lumenova.ai/blog/strategic-necessity-human-oversight-ai-systems/
101. AI Red-Teaming Methodology - SECNORA, accessed April 29, 2025, https://secnora.com/blog/ai-red-teaming-methodology/
102. What is AI Red Teaming? The Complete Guide - Mindgard, accessed April 29, 2025, https://mindgard.ai/blog/what-is-ai-red-teaming
103. What is AI Red Teaming? - Wiz, accessed April 29, 2025, https://www.wiz.io/academy/ai-red-teaming
104. AI Framework Tracker - Fairly AI, accessed April 29, 2025, https://www.fairly.ai/blog/policies-platform-and-choosing-a-framework
105. How to Improve AI Red-Teaming: Challenges and Recommendations | Center for Security and Emerging Technology - CSET, accessed April 29, 2025, https://cset.georgetown.edu/article/how-to-improve-ai-red-teaming-challenges-and-recommendations/
106. AI Ethics beyond Principles: Strengthening the Life-world Perspective - PMC, accessed April 29, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11811459/
107. A Unified Framework of Five Principles for AI in Society - Harvard Data Science Review, accessed April 29, 2025, https://hdsr.mitpress.mit.edu/pub/l0jsh9d1
108. A Capability Approach to AI Ethics - Scholarly Publishing Collective, accessed April 29, 2025, https://scholarlypublishingcollective.org/uip/apq/article/62/1/1/394622/A-Capability-Approach-to-AI-Ethics
109. arXiv:2503.18994v1 [cs.CY] 23 Mar 2025, accessed April 29, 2025, https://arxiv.org/pdf/2503.18994
110. (PDF) Qualitative and quantitative analyses of artificial intelligence ethics in education using VOSviewer and CitNetExplorer - ResearchGate, accessed April 29, 2025, https://www.researchgate.net/publication/369125815_Qualitative_and_quantitative_analyses_of_artificial_intelligence_ethics_in_education_using_VOSviewer_and_CitNetExplorer
111. Ethics of Artificial Intelligence | UNESCO, accessed April 29, 2025, https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
112. Exploring the landscape of trustworthy artificial intelligence: Status and

challenges, accessed April 29, 2025, https://content.iospress.com/articles/intelligent-decision-technologies/idt240366
113. Resolving Ethics Trade-offs in Implementing Responsible AI - ResearchGate, accessed April 29, 2025, https://www.researchgate.net/publication/377400381_Resolving_Ethics_Trade-offs_in_Implementing_Responsible_AI
114. Understanding artificial intelligence ethics and safety - The Alan Turing Institute, accessed April 29, 2025, https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf
115. Solving the AI energy dilemma - WTW, accessed April 29, 2025, https://www.wtwco.com/en-cm/insights/2025/03/solving-the-ai-energy-dilemma
116. Towards Green AI: Current Status and Future Research - arXiv, accessed April 29, 2025, https://arxiv.org/pdf/2407.10237
117. Tracking Carbon Emission During MLOps - dataroots, accessed April 29, 2025, https://dataroots.io/blog/tracking-carbon-emission-during-mlops
118. Lifecycle Assessment of AI Models: Measuring and Mitigating Energy Consumption, accessed April 29, 2025, https://www.researchgate.net/publication/389100406_Lifecycle_Assessment_of_AI_Models_Measuring_and_Mitigating_Energy_Consumption
119. AI Environmental Impact: Understanding the Energy and Water Footprints of AI Models, accessed April 29, 2025, https://tilburg.ai/2024/09/ai-environmental-impact/
120. How to estimate and reduce the carbon footprint of machine learning models, accessed April 29, 2025, https://towardsdatascience.com/how-to-estimate-and-reduce-the-carbon-footprint-of-machine-learning-models-49f24510880/
121. What Is the PUE Data Center Definition? - Vertiv, accessed April 29, 2025, https://www.vertiv.com/en-in/about/news-and-insights/articles/educational-articles/what-is-the-pue-data-center-definition/
122. Understanding Water Usage Effectiveness (WUE) in Data Center - Komprise, accessed April 29, 2025, https://www.komprise.com/glossary_terms/water-usage-effectiveness-wue/
123. How AI Consumes Water: The unspoken environmental footprint - Deepgram, accessed April 29, 2025, https://deepgram.com/learn/how-ai-consumes-water
124. Measuring energy and water efficiency for Microsoft datacenters, accessed April 29, 2025, https://datacenters.microsoft.com/sustainability/efficiency/
125. How Microsoft measures datacenter water and energy use to improve Azure Cloud sustainability, accessed April 29, 2025, https://azure.microsoft.com/en-us/blog/how-microsoft-measures-datacenter-water-and-energy-use-to-improve-azure-cloud-sustainability/
126. CodeCarbon.io, accessed April 29, 2025, https://codecarbon.io/
127. Geography for AI sustainability and sustainability for GeoAI - Taylor & Francis Online, accessed April 29, 2025, https://www.tandfonline.com/doi/full/10.1080/15230406.2025.2479796?src=

128. TPUs improved carbon-efficiency of AI workloads by 3x | Google Cloud Blog, accessed April 29, 2025, https://cloud.google.com/blog/topics/sustainability/tpus-improved-carbon-efficiency-of-ai-workloads-by-3x
129. Google Cloud measures its climate impact through LCA, accessed April 29, 2025, https://cloud.google.com/blog/topics/sustainability/google-cloud-measures-its-climate-impact-through-life-cycle-assessment
130. Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends - arXiv, accessed April 29, 2025, https://arxiv.org/html/2502.01671v1
131. How Google's AI Chip Upgrades Set Sustainability Standards - AI Magazine, accessed April 29, 2025, https://aimagazine.com/articles/how-google-tripled-ai-chip-carbon-efficiency-lca
132. [2502.01671] Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends - arXiv, accessed April 29, 2025, https://arxiv.org/abs/2502.01671
133. Measuring and Modeling CO2 Emissions in Machine Learning Processes - IJS, accessed April 29, 2025, https://aile3.ijs.si/dunja/SiKDD2024/Papers/IS2024_-_SIKDD_2024_paper_23.pdf
134. Measuring your ML impact with CodeCarbon | Talk Python To Me Podcast, accessed April 29, 2025, https://talkpython.fm/episodes/show/318/measuring-your-ml-impact-with-codecarbon
135. mlco2/codecarbon: Track emissions from Compute and recommend ways to reduce their impact on the environment. - GitHub, accessed April 29, 2025, https://github.com/mlco2/codecarbon
136. [2301.11047] A Systematic Review of Green AI - arXiv, accessed April 29, 2025, https://arxiv.org/abs/2301.11047
137. Carbon Emissions in the Tailpipe of Generative AI - Harvard Data Science Review, accessed April 29, 2025, https://hdsr.mitpress.mit.edu/pub/fscsqwx4
138. Toward Green AI: A Methodological Survey of the Scientific Literature - PubliCatt, accessed April 29, 2025, https://publicatt.unicatt.it/retrieve/05b8bfac-5a38-4757-bb39-82a56e299c4e/2024.%20Toward%20Green%20AI%20A%20Methodological%20Survey%20of%20the%20Scientific%20Literature.pdf
139. A Systematic Review of Green AI - arXiv, accessed April 29, 2025, https://arxiv.org/pdf/2301.11047
140. How much water does AI consume? The public deserves to know - OECD.AI, accessed April 29, 2025, https://oecd.ai/en/wonk/how-much-water-does-ai-consume
141. Towards A Comprehensive Assessment of AI's Environmental Impact - arXiv, accessed April 29, 2025, https://arxiv.org/html/2405.14004v1
142. The disruptive and transformative role of AI in the circular economy transition

- Ramboll, accessed April 29, 2025, https://www.ramboll.com/insights/resource-management-and-circular-economy/the-disruptive-and-transformative-role-of-ai-in-the-circular-economy-transition

143. AI and Circular Economy: Reducing Waste in IT and Computing - ResearchGate, accessed April 29, 2025, https://www.researchgate.net/publication/390005123_AI_and_Circular_Economy_Reducing_Waste_in_IT_and_Computing
144. Simplifying e-waste management with AI innovations - Ultralytics, accessed April 29, 2025, https://www.ultralytics.com/blog/simplifying-e-waste-management-with-ai-innovations
145. Advancing Circular Economy Through AI-Driven E-Waste Management: A Comprehensive Review of Current Research, Challenges, and Future Directions - Preprints.org, accessed April 29, 2025, https://www.preprints.org/manuscript/202503.1989/v1
146. AI-Driven Sustainability Metrics in Electronics Manufacturing → Scenario, accessed April 29, 2025, https://prism.sustainability-directory.com/scenario/ai-driven-sustainability-metrics-in-electronics-manufacturing/
147. Monetizing Sustainability: How AI And Data Drive The Circular Economy - Forbes, accessed April 29, 2025, https://www.forbes.com/councils/forbestechcouncil/2025/02/21/monetizing-sustainability-how-ai-and-data-drive-the-circular-economy/
148. AI Driven Circular Economy for Electronic Waste Recycling. → Scenario, accessed April 29, 2025, https://prism.sustainability-directory.com/scenario/ai-driven-circular-economy-for-electronic-waste-recycling/
149. cloud.google.com, accessed April 29, 2025, https://cloud.google.com/blog/topics/sustainability/google-cloud-measures-its-climate-impact-through-life-cycle-assessment#:~:text=LCA%20is%20a%20process%2Danalysis,%2C%20disposal%2C%20etc.).
150. AI-Powered Life Cycle Assessments (LCAs) - CarbonBright, accessed April 29, 2025, https://carbonbright.co/ai-and-life-cycle-assessments-lcas
151. Social impact assessment (SIA) | EBSCO Research Starters, accessed April 29, 2025, https://www.ebsco.com/research-starters/religion-and-philosophy/social-impact-assessment-sia
152. Understanding WCAG: A guide to accessibility in digital design - Siteimprove, accessed April 29, 2025, https://www.siteimprove.com/blog/-understanding-wcag/
153. Accessible Human-Computer Interaction + Inclusive Design - Bookish.press, accessed April 29, 2025, https://bookish.press/tac/HCI
154. Identification of Challenges and Best Practices for Including Users with Disabilities in User-Based Testing - MDPI, accessed April 29, 2025, https://www.mdpi.com/2076-3417/13/9/5498

155. Accessibility First: How AI is Improving Web Accessibility Standards - BSS Monaco, accessed April 29, 2025, https://bss.mc/accessibility-first-how-ai-is-improving-web-accessibility-standards/
156. How to Use AI to Improve Web Accessibility - Accessibud, accessed April 29, 2025, https://www.accessibud.com/blog/ai-web-accessibility/
157. AI and Accessibility - Communications of the ACM, accessed April 29, 2025, https://cacm.acm.org/opinion/ai-and-accessibility/
158. Toward Fairness in AI for People with Disabilities: A Research Roadmap - Microsoft, accessed April 29, 2025, https://www.microsoft.com/en-us/research/uploads/prod/2019/07/Research_Roadmap_ASSETS_2019_Workshop_final.pdf
159. AI MATTERS, VOLUME 5, ISSUE 3 SEPTEMBER 2019 - Considerations for AI Fairness for People with Disabilities - acm sigai, accessed April 29, 2025, https://sigai.acm.org/static/aimatters/5-3/AIMatters-5-3-09-Trewin-accesible.pdf
160. Web Content Accessibility Guidelines (WCAG) 2.2 - W3C, accessed April 29, 2025, https://www.w3.org/TR/WCAG22/
161. (PDF) Towards Inclusive AI: Developing a W3C-Inspired Accessibility Benchmark for Large Language Models - ResearchGate, accessed April 29, 2025, https://www.researchgate.net/publication/384662444_Towards_Inclusive_AI_Developing_a_W3C-Inspired_Accessibility_Benchmark_for_Large_Language_Models
162. Understanding WCAG 3.0: The Future of Web Accessibility - EqualWeb, accessed April 29, 2025, https://www.equalweb.com/a/44475/11527/understanding_wcag_3.0:_what%E2%80%99s_next_for_web_accessibility
163. Tips for Usability Testing with People with Disabilities | Section508.gov, accessed April 29, 2025, https://www.section508.gov/test/usability-testing-with-people-with-disabilities/
164. The Impact of AI on Web Accessibility | Evok Advertising, accessed April 29, 2025, https://evokad.com/impact-ai-web-accessibility/
165. Disability, Accessibility, and Assistive Technology in User Research - Outwitly, accessed April 29, 2025, https://outwitly.com/blog/disability-accessibility-and-assistive-technology-in-user-research/
166. Assistive XR research for disability at ACM ASSETS: A Scoping Review - arXiv, accessed April 29, 2025, https://arxiv.org/html/2504.13849v1
167. Housing Industry Innovation: 5 Ways AI Can Help Boost Supply and Affordability, accessed April 29, 2025, https://bipartisanpolicy.org/explainer/ai-housing-industry-innovations/
168. Can AI Solve the Housing Affordability Issue? | Maket, accessed April 29, 2025, https://www.maket.ai/post/can-ai-solve-the-housing-affordability-issue
169. Leveraging AI-Driven Predictive Pricing Models to Optimize Affordability, PBM Collaboration, and Patient Adherence Strategies - ResearchGate, accessed April 29, 2025, https://www.researchgate.net/publication/389498893_Leveraging_AI-Driven_Pred

ictive_Pricing_Models_to_Optimize_Affordability_PBM_Collaboration_and_Patient_Adherence_Strategies

170. AI Pricing Review in 2025: LLM Economics in Popular Model - Weam AI, accessed April 29, 2025, https://weam.ai/blog/guide/ai-pricing/

171. The CEO's Guide to Generative AI: Cost of compute - IBM, accessed April 29, 2025, https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/ceo-generative-ai/ceo-ai-cost-of-compute

172. AI Development Cost: Analyzing Expenses and Returns - TechMagic, accessed April 29, 2025, https://www.techmagic.co/blog/ai-development-cost

173. PwC's Global Artificial Intelligence Study: Sizing the prize, accessed April 29, 2025, https://www.pwc.com/gx/en/issues/artificial-intelligence/publications/artificial-intelligence-study.html

174. Superagency in the workplace: Empowering people to unlock AI's full potential - McKinsey & Company, accessed April 29, 2025, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

175. The Impact of AI on Work and Employment, accessed April 29, 2025, https://www.ioe-emp.org/index.php?eID=dumpFile&t=f&f=160463&token=8a7078c15874881a559cd18ae85a0b9283afd5db

176. Who owns the content generated by AI? - Marks & Clerk, accessed April 29, 2025, https://www.marks-clerk.com/insights/latest-insights/102k38x-who-owns-the-content-generated-by-ai/

177. Evaluating the Social Impact of Generative AI Systems in Systems and Society, accessed April 29, 2025, https://montrealethics.ai/evaluating-the-social-impact-of-generative-ai-systems-in-systems-and-society/

178. Large Language Model Evaluation: The Complete Guide - Granica AI, accessed April 29, 2025, https://granica.ai/blog/large-language-model-evaluation-grc

179. Evaluation Metrics for Language Modeling - The Gradient, accessed April 29, 2025, https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

180. Evaluating Large Language Models: A Complete Guide | Build Intelligent Applications on SingleStore, accessed April 29, 2025, https://www.singlestore.com/blog/complete-guide-to-evaluating-large-language-models/

181. Best Practices and Metrics for Evaluating Large Language Models (LLMs) - Frugal Testing, accessed April 29, 2025, https://www.frugaltesting.com/blog/best-practices-and-metrics-for-evaluating-large-language-models-llms

182. NLP Model Evaluation - Metrics, Benchmarks, and Beyond - DeconvoluteAI, accessed April 29, 2025, https://deconvoluteai.com/blog/evaluating-nlp-models

183. LLM Evaluation: Key Metrics, Best Practices and Frameworks - Aisera, accessed April 29, 2025, https://aisera.com/blog/llm-evaluation/
184. AI AS THE REASON AND THE SOLUTION OF DIGITAL DIVIDE | Request PDF, accessed April 29, 2025, https://www.researchgate.net/publication/370156651_AI_AS_THE_REASON_AND_THE_SOLUTION_OF_DIGITAL_DIVIDE
185. Behavioral and Psychosocial Dynamics of Engagement: The Digital Divide in Artificial Intelligence [AI]-Driven Sports Podcasts - MDPI, accessed April 29, 2025, https://www.mdpi.com/2076-328X/14/10/911
186. A New Framework, Measurement, and Determinants of the Digital Divide in China - MDPI, accessed April 29, 2025, https://www.mdpi.com/2227-7390/12/14/2171
187. Sociological Implications of the Digital Divide: Exploring Access to Information and Social Inequality in the Age of Artificial Intelligence and Automation | RESEARCH REVIEW International Journal of Multidisciplinary, accessed April 29, 2025, https://rrjournals.com/index.php/rrijm/article/view/1031
188. AN AI-DRIVEN APPROACH TO MITIGATE THE DIGITAL DIVIDE IN EDUCATIONAL RESOURCES - IRJMETS, accessed April 29, 2025, https://www.irjmets.com/uploadedfiles/paper//issue_6_june_2024/59086/final/fin_irjmets1718356920.pdf
189. The State of Digital Impact Assessment Practice, accessed April 29, 2025, https://iaia.org/downloads/State%20of%20Digital%20IA%20Practice_converted.pdf
190. Full article: Digital inequality beyond the digital divide: conceptualizing adverse digital incorporation in the global South - Taylor & Francis Online, accessed April 29, 2025, https://www.tandfonline.com/doi/full/10.1080/02681102.2022.2068492
191. Empowering Accessibility: Exploring AI-Enhanced Universal Design for Inclusive Web Interfaces - kth .diva, accessed April 29, 2025, https://kth.diva-portal.org/smash/get/diva2:1938955/FULLTEXT01.pdf
192. AI Impact on Education: Its Effect on Teaching and Student Success - Netguru, accessed April 29, 2025, https://www.netguru.com/blog/ai-in-education
193. How AI Will Affect Teaching - Panorama Education, accessed April 29, 2025, https://www.panoramaed.com/blog/how-ai-will-affect-teaching
194. AI in Schools: Pros and Cons - College of Education | Illinois, accessed April 29, 2025, https://education.illinois.edu/about/news-events/news/article/2024/10/24/ai-in-schools--pros-and-cons
195. Playbook - NIST AIRC - National Institute of Standards and Technology, accessed April 29, 2025, https://airc.nist.gov/airmf-resources/playbook/
196. Community engagement for artificial intelligence health research in Africa, accessed April 29, 2025, https://wellcomeopenresearch.org/articles/10-158/pdf
197. Health and AI: Advancing responsible and ethical AI for all communities, accessed April 29, 2025, https://www.brookings.edu/articles/health-and-ai-advancing-responsible-and-ethical-ai-for-all-communities/

198. Developing Ethics and Equity Principles, Terms, and Engagement Tools to Advance Health Equity and Researcher Diversity in AI and Machine Learning: Modified Delphi Approach, accessed April 29, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11041493/
199. Algorithmic impact assessment: user guide - Ada Lovelace Institute, accessed April 29, 2025, https://www.adalovelaceinstitute.org/resource/aia-user-guide/
200. Anthropology and AI: A Framework for Mutually Beneficial Collaboration - Matt Artz, accessed April 29, 2025, https://www.mattartz.me/anthropology-and-ai-a-framework-for-mutually-beneficial-collaboration/
201. Towards an AI Anthropology - Azimuth Labs, accessed April 29, 2025, https://azimuthlabs.io/future-perspectives-and-trends/towards-an-ai-anthropology/
202. The EU AI Act Compliance Checklist - DPO Europe, accessed April 29, 2025, https://data-privacy-office.eu/usefull-materials/the-eu-ai-act-compliance-checklist/
203. The 2025 AI Index Report | Stanford HAI, accessed April 29, 2025, https://hai.stanford.edu/ai-index/2025-ai-index-report
204. ISO/IEC 42001 Certification: AI Management System - DNV, accessed April 29, 2025, https://africa.dnv.com/services/iso-iec-42001-artificial-intelligence-ai--250876/
205. AI Risk Management Framework | NIST, accessed April 29, 2025, https://www.nist.gov/itl/ai-risk-management-framework
206. AI Index - OECD.AI, accessed April 29, 2025, https://oecd.ai/en/site/ai-index
207. Understanding the ISO/IEC 42001 for AI Management Systems - Prompt Security, accessed April 29, 2025, https://www.prompt.security/blog/understanding-the-iso-iec-42001
208. EU AI Act Checklist: Six Steps To Find A Compliant AI Partner - Forbes, accessed April 29, 2025, https://www.forbes.com/councils/forbestechcouncil/2025/01/03/eu-ai-act-checklist-six-steps-to-find-a-compliant-ai-partner/
209. Essential Checklist for Responsible EU AI Act Compliance | Resources - OneTrust, accessed April 29, 2025, https://www.onetrust.com/resources/essential-checklist-for-responsible-eu-ai-act-compliance-checklist/
210. EU AI Act Compliance Checklist - Gartner, accessed April 29, 2025, https://www.gartner.com/en/legal-compliance/trends/eu-ai-act
211. AI-Powered Data Governance: Implementing Best Practices - Coherent Solutions, accessed April 29, 2025, https://www.coherentsolutions.com/insights/ai-powered-data-governance-implementing-best-practices-and-frameworks
212. AI, Copyright, and the Law: The Ongoing Battle Over Intellectual Property Rights, accessed April 29, 2025, https://sites.usc.edu/iptls/2025/02/04/ai-copyright-and-the-law-the-ongoing-battle-over-intellectual-property-rights/

213. Artificial Intelligence and Copyright: Navigating the New Legal Landscape - Senior Executive, accessed April 29, 2025, https://seniorexecutive.com/ai-copyright-law-ownership-intellectual-property-rights/

214. AI and intellectual property rights - Dentons, accessed April 29, 2025, https://www.dentons.com/en/insights/articles/2025/january/28/ai-and-intellectual-property-rights

215. US Copyright Office on AI: Human creativity still matters, legally - WIPO, accessed April 29, 2025, https://www.wipo.int/web/wipo-magazine/articles/us-copyright-office-on-ai-human-creativity-still-matters-legally-73696

216. Recent Developments in AI, Art & Copyright: Copyright Office Report & New Registrations, accessed April 29, 2025, https://itsartlaw.org/2025/03/04/recent-developments-in-ai-art-copyright-copyright-office-report-new-registrations/

217. Why the Obsession with Human Creativity? A Comparative Analysis on Copyright Registration of AI-Generated Works - Harvard University, accessed April 29, 2025, https://journals.law.harvard.edu/ilj/2025/02/why-the-obsession-with-human-creativity-a-comparative-analysis-on-copyright-registration-of-ai-generated-works/

218. Data and AI Governance - Frameworks & Best Practices - Sentra, accessed April 29, 2025, https://www.sentra.io/blog/enhancing-ai-governance-the-crucial-role-of-data-security

219. AI Governance: Best Practices and Importance | Informatica, accessed April 29, 2025, https://www.informatica.com/resources/articles/ai-governance-explained.html.html.html.html.html.html.html.html.html.html.html.html.html.html.html.html.html.html.html

220. Best practices for data and AI governance - Databricks Documentation, accessed April 29, 2025, https://docs.databricks.com/aws/en/lakehouse-architecture/data-governance/best-practices

221. a-comparative-framework-for-ai-regulatory-policy.pdf - Ceimia, accessed April 29, 2025, https://ceimia.org/wp-content/uploads/2023/05/a-comparative-framework-for-ai-regulatory-policy.pdf

222. The ethics of artificial intelligence: Issues and initiatives - European Parliament, accessed April 29, 2025, https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

223. Social Impact Assessment and Sustainable Project Planning Under Bill C-69 - WSP, accessed April 29, 2025, https://www.wsp.com/en-gl/insights/social-impact-assessment-and-sustainable-project-planning-under-bill-c-69

224. Understanding ISO 42001: The World's First AI Management System Standard | A-LIGN, accessed April 29, 2025, https://www.a-lign.com/articles/understanding-iso-42001
225. AI and the Future of Work: Insights from the World Economic Forum's Future of Jobs Report 2025 - Sand Technologies, accessed April 29, 2025, https://www.sandtech.com/insight/ai-and-the-future-of-work/
226. (PDF) THE FUTURE OF WORK - AI'S IMPACT ON EMPLOYMENT AND THE ECONOMY, accessed April 29, 2025, https://www.researchgate.net/publication/387739709_THE_FUTURE_OF_WORK_-_AI'S_IMPACT_ON_EMPLOYMENT_AND_THE_ECONOMY
227. Insights on Generative AI and the Future of Work | NC Commerce, accessed April 29, 2025, https://www.commerce.nc.gov/news/the-lead-feed/generative-ai-and-future-work
228. AI Impact on Jobs: Metrics for Measuring Workforce Transformation - ZINFI Technologies, accessed April 29, 2025, https://www.zinfi.com/blog/ai-impact-on-jobs-workforce-metrics/
229. Future of Jobs Report 2025 | World Economic Forum, accessed April 29, 2025, https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf
230. elearningindustry.com, accessed April 29, 2025, https://elearningindustry.com/revolutionizing-education-with-ai-driven-assessments#:~:text=AI%20enhances%20standardization%20by%20applying,strategies%20and%20personalized%20student%20support.
231. Revolutionizing Education With AI-Driven Assessments - eLearning Industry, accessed April 29, 2025, https://elearningindustry.com/revolutionizing-education-with-ai-driven-assessments
232. Exploring the potential of artificial intelligence tools in educational measurement and assessment, accessed April 29, 2025, https://www.ejmste.com/article/exploring-the-potential-of-artificial-intelligence-tools-in-educational-measurement-and-assessment-13428
233. Assessment and Evaluation in the Age of Artificial Intelligence | CUA, accessed April 29, 2025, https://centerforteaching.catholic.edu/cte-spotlight/2024/11/assessment-and-evaluation-in-the-age-of-ai.html
234. Generative artificial intelligence, human creativity, and art | PNAS Nexus - Oxford Academic, accessed April 29, 2025, https://academic.oup.com/pnasnexus/article/3/3/pgae052/7618478
235. Encouraging Human Creativity in the AI-Powered Future - Stanford Social Innovation Review, accessed April 29, 2025, https://ssir.org/articles/entry/ai-creativity-copyrights-patents
236. Creativity and artificial intelligence: a view from the perspective of copyright - ResearchGate, accessed April 29, 2025, https://www.researchgate.net/publication/352117219_Creativity_and_artificial_intelligence_a_view_from_the_perspective_of_copyright

237. The impact of technological advancement on culture and society - PMC, accessed April 29, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11685575/
238. Applying AI to Anthropological Research - All Things Innovation, accessed April 29, 2025, https://allthingsinnovation.com/content/applying-ai-to-anthropological-research/
239. Generative AI as Cultural Artifact: Applying Anthropological Methods to AI Literacy, accessed April 29, 2025, https://www.researchgate.net/publication/390876703_Generative_AI_as_Cultural_Artifact_Applying_Anthropological_Methods_to_AI_Literacy
240. Full article: AI and SIA: some reflections, accessed April 29, 2025, https://www.tandfonline.com/doi/full/10.1080/14615517.2024.2432730?src=exp-la
241. Overcoming Racial Harms to Democracy from Artificial Intelligence - Iowa Law Review, accessed April 29, 2025, https://ilr.law.uiowa.edu/sites/ilr.law.uiowa.edu/files/2025-01/ILR-110-Overton.pdf
242. Can Democracy Survive the Disruptive Power of AI? | Carnegie Endowment for International Peace, accessed April 29, 2025, https://carnegieendowment.org/research/2024/12/can-democracy-survive-the-disruptive-power-of-ai
243. AI Threats to Politics, Elections, and Democracy: A Blockchain-Based Deepfake Authenticity Verification Framework - MDPI, accessed April 29, 2025, https://www.mdpi.com/2813-5288/2/4/20
244. Analyzing the Benefits of Artificial Intelligence to Racially Inclusive Democracy - Scholarly Commons, accessed April 29, 2025, https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=3040&context=faculty_publications
245. Wreck the Vote: How AI-Driven Misinformation Could Undermine Democracy, accessed April 29, 2025, https://www.gsb.stanford.edu/insights/wreck-vote-how-ai-driven-misinformation-could-undermine-democracy
246. AI-Enabled Influence Operations: Safeguarding Future Elections, accessed April 29, 2025, https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-safeguarding-future-elections
247. Impact of Technology on the Future of Democracy, accessed April 29, 2025, https://community-democracies.org/app/uploads/2024/07/Impact-of-Technology-on-the-Future-of-Democracy-Trend-and-Policy-Brief-July-2024.pdf
248. Social Impact Assessment, accessed April 29, 2025, https://www.iaia.org/wiki-details.php?ID=23
249. Systematic Map of the Social Impact Assessment Field - MDPI, accessed April 29, 2025, https://www.mdpi.com/2071-1050/11/15/4106
250. System dynamic simulation: A new method in social impact assessment (SIA) | Request PDF, accessed April 29, 2025, https://www.researchgate.net/publication/309170146_System_dynamic_simulation_A_new_method_in_social_impact_assessment_SIA
251. Selection of artificial intelligence provider via multi-attribute decision-making

technique under the model of complex intuition - AIMS Press, accessed April 29, 2025, https://www.aimspress.com/aimspress-data/math/2024/11/PDF/math-09-11-1581.pdf

252. Bias in Decision-Making for AI's Ethical Dilemmas: A Comparative Study of ChatGPT and Claude - arXiv, accessed April 29, 2025, https://arxiv.org/html/2501.10484v1
253. A Comparative Analysis on Ethical Benchmarking in Large Language Models, accessed April 29, 2025, https://www.researchgate.net/publication/385316806_A_Comparative_Analysis_on_Ethical_Benchmarking_in_Large_Language_Models
254. Selection of artificial intelligence provider via multi-attribute decision-making technique under the model of complex intuitionistic fuzzy rough sets - AIMS Press, accessed April 29, 2025, https://www.aimspress.com/article/doi/10.3934/math.20241581?viewType=HTML
255. Comparison study of multi-attribute decision analytic software - ResearchGate, accessed April 29, 2025, https://www.researchgate.net/publication/229674721_Comparison_study_of_multi-attribute_decision_analytic_software
256. Implementing the NIST Artificial Intelligence Risk Management Framework – Map, accessed April 29, 2025, https://angle.ankura.com/post/102j3pa/implementing-the-nist-artificial-intelligence-risk-management-framework-map
257. A Comparative Analysis of Model Alignment Regarding AI Ethics Principles - ResearchGate, accessed April 29, 2025, https://www.researchgate.net/publication/383041387_A_Comparative_Analysis_of_Model_Alignment_Regarding_AI_Ethics_Principles
258. Evaluating Metrics for Impact Quantification - Ethics + Emerging Sciences Group, accessed April 29, 2025, https://www.ethics.calpoly.edu/EMIQ_report.pdf
259. AI Metrics 101: Measuring the Effectiveness of Your AI Governance Program - Zendata, accessed April 29, 2025, https://www.zendata.dev/post/ai-metrics-101-measuring-the-effectiveness-of-your-ai-governance-program
260. Establishing and Evaluating Trustworthy AI: Overview and Research Challenges - arXiv, accessed April 29, 2025, https://arxiv.org/html/2411.09973v1
261. Elements of the NIST AI RMF: What you need to know - Holistic AI, accessed April 29, 2025, https://www.holisticai.com/blog/nist-ai-rmf-core-elements
262. Resolving Ethics Trade-offs in Implementing Responsible AI - arXiv, accessed April 29, 2025, https://arxiv.org/html/2401.08103v3
263. Multi-Attribute, Multimodal Bias Mitigation in AI Systems - YouTube, accessed April 29, 2025, https://www.youtube.com/watch?v=izruljNpLwU
264. The Importance of a Socio-technical Approach in AI Development - Regulations.gov, accessed April 29, 2025, https://downloads.regulations.gov/NIST-2023-0009-0146/attachment_1.pdf
265. [2307.05333] Unbiased Pain Assessment through Wearables and EHR Data:

Multi-attribute Fairness Loss-based CNN Approach - arXiv, accessed April 29, 2025, https://arxiv.org/abs/2307.05333