Responsible Generative AI: A Review of Technical and Regulatory Frontiers

Anonymous Author(s)
Affiliation
Address
email

Abstract

Generative AI (GenAI) has rapidly expanded into domains such as healthcare, finance, education, and media, raising acute concerns around fairness, transparency, accountability, and governance. While prior Responsible AI (RAI) surveys have addressed bias mitigation, privacy, and ethical design, they largely focus on traditional AI and overlook the distinctive risks of GenAI, including hallucinations, stochastic outputs, intellectual property disputes, and large-scale synthetic content generation. This survey addresses that gap by systematically reviewing more than 80 studies published between 2022 and 2024 to examine Responsible Generative AI through both technical and regulatory perspectives. We identify five core problem areas: **data**-related risks, **model**-related risks, challenges with **regulation**, the limited scope of existing **benchmarks**, and poor **explainability**. In response, we highlight emerging solutions across five domains: establishing clear **principles**, adopting **governance frameworks**, defining measurable **metrics**, validating through AI-ready **testbeds**, and enabling adaptive oversight via **regulatory sandboxes**. By mapping these problem and solution spaces, this study contributes an integrated framework for Responsible Generative AI, providing actionable insights for researchers, practitioners, and policymakers seeking to align innovation with ethical, societal, and legal expectations.

1 Introduction

Artificial Intelligence (AI) is emerging as a pervasive technology influencing institutions and daily life across society. AI systems are increasingly integrated into decision-making across healthcare [81], finance [22], transportation [2], and education [194], raising critical concerns about fairness, transparency, accountability, and related issues [64, 164, 197].

To address these challenges, a substantial body of work on Responsible AI (RAI) has developed, encompassing measures such as bias mitigation, privacy protection, security enhancement, and the safeguarding of human rights [122, 143]. Contributions include frameworks from industry [7, 61, 100, 159] and academia [38, 93, 148], offering both theoretical principles [93, 97, 103, 144] and practical implementations [25, 40, 71]. Yet existing surveys remain limited in two important ways. First, most focus on traditional AI rather than the distinctive risks of Generative AI (GenAI), such as hallucination, intellectual property conflicts, and large-scale generation of synthetic content [18, 99, 185]. Second, prior studies typically analyze RAI through either a technical [40, 71] or a regulatory lens [70, 148], overlooking how these perspectives intersect. Recent analyses reveal that existing AI safety frameworks cover only a fraction of identified risks, underscoring persistent gaps between regulation and technical implementation [154].

Submitted to Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

This survey addresses these gaps by focusing on the current challenges and state-of-the-art solutions that enable Responsible Generative AI across both technical and regulatory domains. Our guiding research question is: *Our guiding research question is: What are the most pressing challenges and promising solutions towards Responsible Generative AI from technical-regulatory perspectives?*

To answer this question, we systematically reviewed more than 80 studies published between 2022 and 2024 across leading journals, conferences, and governance reports [23, 37, 71]. As part of this review, we traced the evolution of the Responsible AI landscape (Section 2), prioritized the most important current challenges, highlighted in Section 3, and examined promising solution strategies (Section 4). A structured quality assessment, conducted by a team of twelve academic and industry experts, evaluated rigor, validity, and relevance [47, 80]. This dual perspective strengthens our analysis by grounding it in both scholarly research and real-world practice.

As with similar reviews of Responsible AI frameworks [18, 47, 80, 99], this study has several limitations. First, it is framed primarily through technical and regulatory perspectives, which may underrepresent insights from social sciences and civil society. Second, it synthesizes existing literature and frameworks rather than providing new empirical validation. Third, the rapid evolution of generative AI means some findings may become time-sensitive and require future updates.

Despite these limitations, this survey offers a consolidated view of frontier challenges and solutions across both technical and regulatory domains, providing a foundation for organizations and policymakers to shape effective responsible AI strategies.

2 Evolution of the responsible AI landscape

To address our research question, we first trace the evolution of Responsible AI, highlighting the trends of challenges, and solution strategies that have emerged over time. Over the past decade, multiple perspectives have shaped the evolving landscape of Responsible AI. **Fairness, Accountability, and Transparency (FAT/FAccT)** established bias mitigation, accountability mechanisms, and transparency as early foundations for ethical AI [11, 104]. **Trustworthy AI** was subsequently advanced through ethical frameworks such as AI4People [48] and comparative analyses of global guidelines [80], highlighting privacy, robustness, and security as prerequisites for public trust. In parallel, research on **Explainable AI (XAI)** developed methods for interpretability and transparency, aiming to make opaque systems more understandable to users and regulators [8, 62]. **Human-Centered AI** further emphasized usability, human agency, and participatory design [153], while **AI Safety and Alignment** addressed risks posed by advanced AI systems, focusing on robustness, controllability, and alignment with human values [6, 142]. Alongside these, **Sustainability Perspectives** such as AI for Social Good and Green AI stressed the role of AI in advancing societal well-being while reducing the environmental costs of large-scale training and deployment [151, 178]. More recently, **Responsible AI** has emerged as a unifying framework integrating fairness, safety, accountability, and governance [38, 93], while **Responsible Generative AI** extends these practices to tackle new risks such as hallucination, intellectual property conflicts, and disinformation [18, 99]. Finally, the concept of **Regulatable AI** emphasizes designing systems with features that facilitate compliance and oversight, enabling more effective regulatory intervention [55, 107].

3 Current challenges in responsible GenAI

3.1 Technical challenges

Technical challenges in Responsible Generative AI are empirically testable issues in data, models, and systems. They can be grouped into three main categories: data-related challenges, model-related challenges, and misinformation and media manipulation [82].

Data-related challenges Ensuring integrity, fairness, diversity, and balance in training data is critical. Biased or unrepresentative datasets reinforce inequities, while responsible curation practices (e.g., datasheets) improve accountability. Techniques such as augmentation, bias correction, and audits are used to enhance quality [105, 132, 156, 172].

Model-related challenges Advanced AI models face persistent issues of explainability, transparency, and accountability, as deep architectures often act as black boxes with declining interpretability at scale [58, 63, 98, 190]. They also propagate biases from training data, creating fairness–privacy tradeoffs that require safeguards such as differential privacy, federated learning, and encryption [12, 13, 46, 67]. Further, robustness and generalization remain limited, with models vulnerable to adversarial inputs and misuse risks, including deepfakes and disinformation [27, 102].

Misinformation and media manipulation Generative models exacerbate the spread of misinformation through hallucinated outputs, fake citations, fabricated media, and numerical errors, undermining trust in information ecosystems [53, 130, 145, 189, 192]. Addressing these risks requires robust verification mechanisms as well as content moderation strategies, including filtering algorithms, adversarial defenses, and detection of fake personas and bot networks [17, 83, 131, 133]. At the same time, increasingly realistic visual and audio manipulation—enabled by face swapping, voice cloning, synthetic identities, and diffusion models like Stable Diffusion and DiffVoice—raises serious risks of fraud, deception, and security breaches, necessitating stronger detection methods and regulatory oversight [14, 26, 28, 54, 136, 137, 191].

3.2 Regulatory challenges

Regulatory challenges are the normative and institutional tasks of creating and enforcing frameworks to ensure the ethical, accountable, and privacy-compliant use of AI [37]. They fall into six main categories:

Copyright and intellectual property disputes. Training GenAI on copyrighted data raises unresolved questions of authorship and ownership, with models reproducing protected works and exposing gaps in existing law [52, 82, 120, 196].

Ethical AI alignment. Generative models can produce biased or toxic content, undermining trust [185]. Regulatory measures include bias-free data, detection algorithms [161], ethical audits [84], and fairness and accountability frameworks [119], complemented by user feedback.

Accountability and transparency gaps. GenAI often generates biased, hallucinated, or privacy-sensitive content, leaving responsibility unclear across developers, data providers, and users [42]. The EU AI Act addresses these gaps through risk classification and transparency mandates [117], but disputes over unauthorized data use remain.

Ethical dilemmas in automated decision-Making. Applications in hiring, lending, and healthcare risk perpetuating discrimination when trained on biased data, requiring stronger regulatory oversight to ensure fairness.

Algorithmic accountability and explainability. Mandates such as those in the EU AI Act require explainability for high-risk systems, but complexity in models like GPT-4 limits interpretability [154]. This creates compliance tensions, while evolving risks such as deepfakes in elections demand updated standards.

Lack of standardized codes of practice. The absence of uniform standards for AI integration heightens risks of bias, privacy violations, and unreliable systems, underscoring the need for industry-wide codes [154].

Together, these categories highlight the persistent gaps in aligning generative AI with robust regulatory, ethical, and institutional safeguards.

3.3 Gaps in AI safety benchmarks

We evaluated 17 AI safety benchmarks against the technical risks identified in Section 3.1. While most benchmarks address bias, discrimination, and toxicity, coverage of security, misinformation, and privacy remains limited. Emerging threats such as deepfakes and AI system failures are notably underrepresented, underscoring gaps in risk mitigation and the need for stronger regulatory alignment [16, 33, 51, 57, 72, 73, 88, 89, 91, 96, 134, 139, 147, 152, 176, 182, 195]. Table 1 maps existing benchmarks to regulatory areas, illustrating the uneven coverage of critical risks.

3.4 Gaps in explainability of GenAI

Advanced GenAI models remain opaque black boxes, creating fundamental barriers to transparency and accountability [8, 98]. Key obstacles arise from intrinsic opacity, as highly complex neural architectures defy direct human interpretation; stochastic outputs, where identical prompts yield

Table 1: Mapping AI Safety Benchmarks to Regulatory Areas. A checkmark (‘x‘) indicates that the benchmark addresses the regulatory concern.

Benchmark	**Bias/ Discrimination**	**Toxicity**	**Security**	**Mis-/ Disinformation**	**Deepfakes**	**Privacy**	**System Failure**	**Malicious Actors**
HarmBench [96]	x	x	x	x				
SALAD-Bench [88]		x	x	x		x		x
Risk Taxonomy [33]	x	x	x			x		x
TrustGPT [72]	x	x						
RealToxicityPrompts [57]		x						
DecodingTrust [182]	x	x				x		x
SafetyBench [195]	x	x				x		x
MM-SafetyBench [91]		x	x	x		x		x
Xstest [139]						x		x
Rainbow Teaming [147]	x	x	x				x	x
MFG for GenAI [51]	x		x	x				
AI Safety Benchmark [176]	x	x						x
HELM [89]	x		x	x				
SHIELD [152]			x		x			
BIG-Bench [16]	x	x	x				x	
BEADs [134]	x	x		x				
TrustLLM [73]	x	x		x		x	x	

different results depending on sampling parameters such as temperature or top-k decoding [75]; and multimodal complexity, where models integrate text, images, and audio in ways that complicate causal reasoning [123]. These issues are compounded by hallucinations and fabricated outputs, which undermine trust, complicate accountability, and increase risks in sensitive domains such as healthcare and finance [185].

The lack of explainability in these systems directly **erodes confidence of users**, making AI adoption increasingly difficult [39]. At the same time, explainability is foundational to governance objectives such as transparency, auditability, and accountability. Without it, these objectives become unattainable, which in turn **makes regulatory compliance difficult** [8, 80, 181, 186]. These gaps highlight explainability not as a secondary consideration but as a central challenge for Responsible Generative AI. Section 3.4 therefore examines frontier solutions for explainability, focusing on approaches to mitigate these deficiencies. Despite these challenges there are signifcant efforts towards explainable AI.

Explainability is not merely another principle within Responsible AI but functions simultaneously as a **regulatory compliance** and a **trust mechanism** [181]. It is increasingly mandated in high-risk AI contexts to ensure oversight and accountability, while also enabling users and stakeholders to evaluate outputs, thereby fostering confidence and facilitating adoption [8, 39, 186]. This dual role situates explainability at the core of Responsible Generative AI, requiring continuous technical innovation alongside regulatory support.

In principle, one might aim for *intrinsic* interpretability, where models are designed to be transparent by construction, such as decision trees or linear models. This remains feasible for smaller-scale machine learning models but not for large, complex foundation models due to their multi-layered architectures and scale [140]. Consequently, researchers rely on *post-hoc* explainability methods that generate explanations after predictions are produced, without modifying the underlying model. Techniques such as LIME and SHAP exemplify this approach, providing local, retrospective insights by approximating model behavior and highlighting influential features [94, 138].

Figure 1 presents a conceptual mapping of these frontier post-hoc approaches. As the figure illustrates, interpretability pathways clarify decision rationales at the feature level [193], model simplification enhances transparency by approximating complex systems with more tractable forms [32], and visualization tools such as attention maps and heatmaps make hidden processes more accessible to stakeholders [68, 174]. In parallel, fairness and bias mitigation methods operationalize ethical constraints through quantifiable metrics [109], while trust enhancement mechanisms reinforce user confidence by validating outputs in human-understandable terms [186]. Taken together, these approaches contribute to transparency, accountability, and trust in systems such as LLMs and VLMs. By situating explainability in this structured manner, researchers can bridge the gap between advanced AI capabilities and societal expectations for ethical and accountable deployment [125, 128, 136].

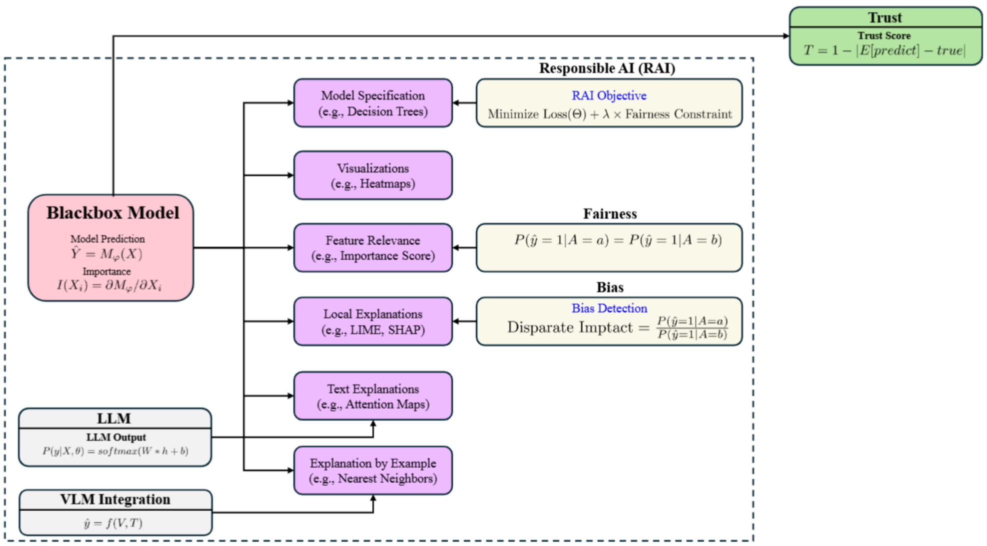

Figure 1: A conceptual mapping of post-hoc explainability approaches in AI.

4 Frontier solutions of responsible GenAI

4.1 Principles for responsible GenAI

Establishing clear principles is indispensable to implement Responsible Generative AI [49, 80]. They provide the normative foundation that guides the entire life cycle of AI design, development, deployment, and use [106]. Our analysis highlights four overarching categories of principles, including technical, legal, sustainable, and innovation management, as illustrated in Figure 2. Together, they provide a framework that guides the selection of principles for Responsible Generative AI.

The **technical dimension** encompasses accountability, transparency, fairness, privacy, safety, and autonomy. Accountability clarifies responsibility and enables redress when harms occur [180]. Transparency facilitates auditability and oversight [8, 39]. Fairness reduces discriminatory outcomes and promotes equitable treatment [11, 41]. Privacy safeguards legitimacy and public trust by protecting data boundaries [110]. Safety addresses unintended failures and risks of harm in deployment [6]. Autonomy preserves human agency and ensures that decision-making remains under meaningful human control [20].

The **legal dimension** anchors AI within binding regulatory frameworks and engages with international standards to promote coherence across jurisdictions. Compliance with national laws provides enforceable safeguards, while voluntary adoption of international best practices strengthens legitimacy and trust. Examples include the OECD AI Principles and the EU AI Act, which emphasize accountability, fairness, and proportionality as legal standards for responsible AI development [59, 115].

The **sustainable dimension** addresses the broader social, environmental, and economic impacts of AI systems [24, 158, 178]. This includes reducing the carbon footprint of model training and deployment, ensuring inclusion and accessibility in applications, and fostering equitable benefit sharing across communities [106]. Sustainability must also extend to *procurement*, with organizations prioritizing suppliers and AI systems that meet environmental and social responsibility criteria [56].

The **responsible innovation management dimension** emphasizes anticipation, reflexivity, responsiveness, and inclusion as guiding commitments [157, 179]. Anticipation identifies risks and opportunities early in the design process. Reflexivity ensures continuous reassessment of assumptions, methods, and impacts. Responsiveness enables adaptation to emerging evidence and societal needs.

Inclusion embeds diverse stakeholders in development and oversight. Together, these commitments integrate ethical responsibility into organizational processes and knowledge management, ensuring responsibility is systematically practiced rather than applied ad hoc [118, 149].

In short, these categories provide an integrated structure for selecting and tailoring principles that enable the responsible design and deployment of generative AI.

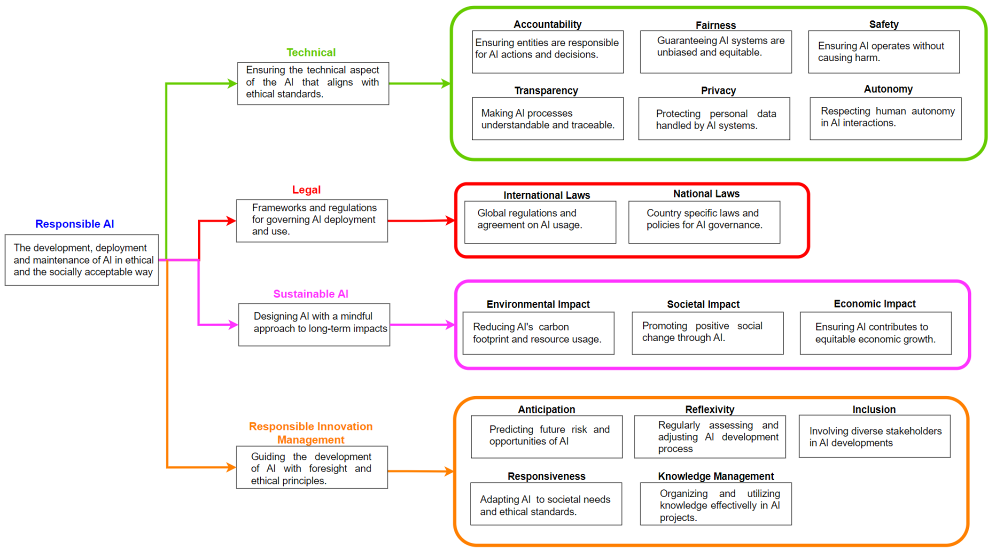

Figure 2: Principles for Responsible Generative AI grouped into technical, legal, sustainable, and innovation management dimensions.

4.2 Governance frameworks and tools

Governance frameworks provide actionable mechanisms to translate principles and high-level commitments into operational practice [49, 80]. They establish structured processes for documentation, risk management, monitoring, auditability, and compliance, among others [106]. Effective adoption requires organizations to examine established frameworks, assess their relevance to regulatory and industrial contexts, and tailor them holistically to reflect the principles set for Responsible Generative AI [90, 126].

In this study, we examined a range of governance frameworks and identified representative examples, summarized in Table 2. The table also includes complementary technical tools that support practitioners in implementing these frameworks or specific components of them. On the governance side, standards such as ISO/IEC TR 24027:2021 and NIST SP-1270 provide systematic structures for bias mitigation, while the OECD AI principles and U.S. GAO accountability framework establish mechanisms for transparency and oversight [115, 171]. On the technical side, toolkits like AI Fairness 360 [15], Fairlearn [184], and AIX360 [9] operationalize these governance principles by enabling bias detection, interpretability, and explainability in practice. Similarly, robustness toolkits (e.g., ART) and privacy solutions (e.g., TensorFlow Privacy) extend governance mandates into concrete assurance techniques [158, 184]. Taken together, these frameworks and tools provide practitioners with a robust set of instruments to operationalize governance and align Responsible Generative AI with real-world deployment.

4.3 Key performance indicators for responsible AI

The objectives and tasks embedded within governance frameworks for Responsible AI must be operationalized through clear and measurable indicators; otherwise, commitments to responsibility remain aspirational rather than actionable [48, 80]. Selecting appropriate Key Performance Indicators (KPIs) across the full AI lifecycle—from design and data preparation to development, deployment, and monitoring—is therefore critical. These indicators translate the principles and objectives

Table 2: Representative governance frameworks and technical AI tools for Responsible AI.

Category	Governance Frameworks	Technical Tools
Fairness & Bias Mitigation	ISO/IEC TR 24027:2021 [1] (AI and data quality); NIST SP-1270 [150] (bias identification and management); AI4ALL [4] (diversity and inclusion in AI)	Google What-if Tool [187] (visual fairness evaluation); IBM AI Fairness 360 [15] (bias detection and mitigation); Microsoft Fairlearn [184] (fairness metrics and debiasing)
Interpretability & Explainability	IEEE P2976 [124] (standards for XAI); IEEE P2894 [30] (guidelines for XAI implementation); UK ICO–Turing Guidance [87] (regulatory guidance for explainability)	LIT [162] (interactive model interpretability); InterpretML [146] (explainability with LIME/SHAP); ELI5 [35] (simple explanations); IBM AIX360 [9] (toolkit for multiple explanation techniques)
Transparency & Accountability	US GAO AI Accountability Framework [171] (accountability practices for AI); OECD AI Principles [115] (global policy recommendations)	Evidently [44] (monitoring and evaluation); MLBench [108] (benchmarking transparency and robustness)
Privacy, Safety & Security	NASA Hazard Modes [155] (safety in ML for space); IEEE P7009 [45] (fail-safe autonomous systems); ISO/IEC TS 27022:2021 [160] (security guidelines); NIST AI 100-2 [114] (cybersecurity for AI); UK AI Cybersecurity Code [36]; MIT Risk Repo [154] (AI risk documentation); Google SAIF [76] (secure AI framework)	Unitary [170] (toxicity detection); ART Robustness Toolbox (adversarial robustness); Privacy Meter (privacy auditing); Google DP [60] (differential privacy); SecretFlow (privacy-preserving ML); TensorFlow Privacy [163]; Betterdata PET [77] (synthetic data privacy)
Ethical Guidelines & Compliance	GDPR [169] (EU data protection law); EU AI Act [85] (risk-based AI regulation); Canadian Voluntary Code of Conduct [112]; US AI Executive Order [69]; WHO Ethics in AI for Health [116]	Google RAI Practices [3]; Australia AI Ethics Principles [10]; Microsoft RAI Toolbox [101]; OneTrust [113] (compliance and risk management)
Monitoring & Auditing	AI Incident Database [34] (documented harms from AI); UK ICO AI Auditing Framework [167] (auditing standards)	FairVis [21] (bias visualization); Aequitas [50] (fairness audit toolkit); Audit-AI [78] (bias auditing in ML)

of Responsible AI into measurable outcomes and allow organizations to track progress systematically [95, 129].

Based on our examination, we identify a set of representative KPIs that measure the "responsibleness" of AI solutions across the lifecycle. Table 3 presents these indicators. In the design and data preparation stage, **data quality and integrity** measure the accuracy, consistency, and reliability of datasets, ensuring trustworthy inputs [121]. **Data privacy compliance** evaluates adherence to regulatory standards such as GDPR or CCPA, quantifying the proportion of compliant data records [86]. **Bias detection** quantifies disparities across protected groups using fairness metrics such as Statistical Parity Difference or Disparate Impact [121, 135].

In model development, **fairness scores** assess the equity of outcomes across demographic groups, **explainability indices** capture the interpretability of model decisions to stakeholders, and **robustness assessments** test system reliability under perturbed or adversarial conditions [65, 177]. During deployment, **equal performance** ensures comparable predictive accuracy across groups [127], while **sustainability metrics** track energy consumption and environmental impact of large-scale models [158, 173]. Finally, monitoring and governance rely on **high-stakes error rates**, which quantify harmful or unsafe outputs, and **audit frequency and resolution times**, which reinforce accountability by measuring the pace of issue detection and remediation [37, 66].

In sum, these KPIs provide a systematic foundation for aligning AI systems with ethical and societal expectations, reinforcing trust, accountability, and sustainable innovation.

4.4 AI-ready testbeds

Testing in controlled environments prior to deployment is essential for translating Responsible AI principles into practice and for validating systems against both technical and governance requirements [141, 166, 175]. In this landscape, *AI-ready testbeds* provide experimental settings for algorithmic validation and repeatable evaluation (e.g., robustness, bias, privacy), while *regulatory sandboxes* offer supervised, real or simulated conditions where innovators and regulators jointly probe compliance, accountability, and oversight [141, 175].

Testbeds help examine GenAI's distinctive risks (stochastic outputs, multimodality, and hallucinations) using structured experiments and repeat runs; they also enable targeted probes for fairness,

Table 3: Key Performance Indicators (KPIs) for Evaluating Responsible AI

KPI	Description and Formula
Design & Data Preparation	
Data Quality and Integrity	Accuracy, consistency, and reliability of datasets. $$Q = \frac{\text{Valid Entries}}{\text{Total Entries}} \times 100$$
Data Privacy Compliance	Proportion of data compliant with privacy standards (e.g., GDPR, CCPA). $$P = \frac{\text{Compliant Data Points}}{\text{Total Data Points}} \times 100$$
Bias Detection	Measures disparities across groups, e.g., Statistical Parity Difference (SPD). $$\text{SPD} = P(\hat{y} = 1 \mid A = a) - P(\hat{y} = 1 \mid A = b)$$
Model Development	
Fairness Score	Disparate Impact (DI) quantifies equity across groups. $$\text{DI} = \frac{P(\hat{y} = 1 \mid A = a)}{P(\hat{y} = 1 \mid A = b)}$$
Explainability Index	Aggregate contribution of features to interpretability. $$E = \sum_{i=1}^{n} \lvert \text{Feature Contribution}_i \rvert$$
Robustness Assessment	Reliability under perturbations or adversarial conditions. $$R = \frac{\text{Errors under Perturbation}}{\text{Total Perturbed Inputs}} \times 100$$
Deployment	
Equal Performance	Consistency in predictive accuracy across groups. $$\Delta_{\text{accuracy}} = \lvert \text{Accuracy}_{A=a} - \text{Accuracy}_{A=b} \rvert$$
Sustainability Metrics	Environmental cost of computation. $$E_c = P \times T$$
Monitoring & Governance	
High-Stakes Error Rate	Proportion of harmful or critical errors. $$E_h = \frac{\text{Critical Errors}}{\text{Total Predictions}} \times 100$$
Audit Frequency and Resolution Time	Frequency of audits and average resolution speed. $$F_a = \frac{\text{Number of Audits}}{\text{Time Period}}, \quad T_r = \text{Average Time to Resolve Issues}$$

safety, and privacy before release [111, 122, 183]. Table 4 summarizes representative testbeds used to study RAI properties in practice, including platforms oriented to explainability, accountability, and human-centered evaluation [31, 92, 188].

Regulatory Sandboxes Unlike technical testbeds, sandboxes are policy instruments designed to balance innovation with compliance by permitting supervised experimentation under legal and procedural safeguards [141, 166, 175]. They have become a central feature of the EU AI Act's implementation discourse and several national strategies, functioning as a bridge between technical risk assessment and institutional oversight [19].

In practice, organizations can use *AI-ready testbeds* to gather technical evidence (e.g., fairness, robustness, privacy leakage) and then leverage *regulatory sandboxes* to validate procedures, docu-

Table 4: Representative AI-ready testbeds for evaluating Responsible AI practices.

Testbed Name	Domain	Key Features
AI4EU AI-on-Demand [31]	General AI Development	EU-funded platform promoting responsible AI development with ethical principles, transparency, and explainability.
IEEE Ethical AI Testbed [188]	Ethical AI Development	Evaluation with ethical frameworks, human-centered AI, and fairness.
AI Testbed for Trustworthy AI (TNO) [165]	Trustworthy AI	Assesses robustness, transparency, and fairness.
ETH Zurich Safe AI Lab [43]	Safe and Fair AI	Safety-critical validation with focus on robustness, fairness, and reliability.
HUMANE AI [74]	Human-Centric AI	Ensures alignment with human values, societal impact, and fairness.
AI for Good Testbed (ITU) [79]	Social Good	Ethical AI aligned with UN SDGs, emphasizing ethics and societal benefit.
UKRI TAS Hub [168]	Autonomous Systems	Trustworthy autonomy with accountability, transparency, and compliance.
Algorithmic Justice League [5]	Algorithmic Fairness	Fairness testing, bias mitigation, and ethical AI development.
ClarityNLP Healthcare [29]	Healthcare AI	Fairness, transparency, and ethical data usage in clinical NLP.
ToolSandbox [92]	LLM Tool Use	Evaluates LLMs for privacy, fairness, transparency, and accountability in stateful tasks.

Table 5: Representative sample of regulatory sandboxes for AI.

Program / Instrument	Jurisdiction	Focus in the literature
EU AI Act Regulatory Sandboxes [141, 175]	European Union	Supervised experimentation to align innovation with risk-based requirements; dialogue with regulators; openness and oversight design.
Spain AI Regulatory Sandbox (pilot) [19]	Spain	Early pilot implementation; allocation of oversight, procedural clarity, and cross-border consistency.
Cross-sector EU experimentation facilities [141]	EU (comparative)	Relationship between AI sandboxes, living labs, and experimentation facilities; governance challenges and standardization needs.
Conceptual models for high-risk AI sandboxes [166]	Comparative	Legal design rationales; criteria for eligibility, supervision, and exit; risks of fragmentation.

mentation, and controls under supervisory conditions, thereby aligning technical performance with regulatory expectations [141, 175].

5 Conclusion

This survey has mapped the problem and solution spaces for Responsible Generative AI through technical and regulatory perspectives. Taken together, these dimensions highlight the frontier for advancing Responsible Generative AI.

On the problem side, what stands out as particularly concerning are challenges in **data** (bias, quality, diversity, and privacy at scale) and **model** (opacity, stochastic outputs, robustness limits, and misuse risks). Alongside these technical issues, challenges with **regulation** add further complexity through accountability gaps, intellectual property disputes, and compliance burdens across jurisdictions. Limited **benchmarks** continue to lag in covering critical risks, while poor **explainability** undermines transparency, accountability, and trust in generative systems.

On the solution side, promising directions include establishing clear **principles** that anchor responsible practices, adopting **governance frameworks** that translate commitments into operational processes, and defining measurable **metrics** to track progress systematically. In addition, technical **testbeds** enable pre-deployment validation across domains, while supervised **regulatory sandboxes** provide dynamic policy environments for experimentation under oversight.

References

[1] ISO/IEC JTC 1/SC 42/WG 3. SO/IEC TR 24027:2021. Technical Report ISO/IEC TR 24027:2021, ISO - International Organization for Standardization, 11 2021. URL `https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:24027:ed-1:v1:en`.

[2] Rusul Abduljabbar, Hussein Dia, Sohani Liyanage, and Saeed Asadi Bagloee. Applications of artificial intelligence in transport: An overview. *Sustainability*, 11(1):189, 2019.

[3] Google AI. Responsible ai practices, 2023. URL `https://ai.google/responsibility/responsible-ai-practices/`. [Accessed 01-10-2024].

[4] AI4ALL. Our vision for ai, 2024. URL `https://ai-4-all.org/about/our-story/`. [Accessed 30-09-2024].

[5] Algorithmic Justice League. Algorithmic justice league. `https://www.ajl.org/`, 2024. [Accessed: 2024-09-25].

[6] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. URL `https://arxiv.org/abs/1606.06565`.

[7] Anthropic. Core views on AI safety: When, why, what, and how, November 2024. URL `https://www.anthropic.com/news/core-views-on-ai-safety`. Accessed: 2024-11-18.

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[9] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, september 2019. URL `https://arxiv.org/abs/1909.03012`.

[10] Australian Government Department of Industry, Science and Resources. Australia's Artificial Intelligence Ethics Framework, 2019. URL `https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework`. [Accessed 01-10-2024].

[11] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. In *Fairness, accountability, and transparency in machine learning*. NIPS Tutorial, 2017. URL `https://fairmlbook.org/`.

[12] Susanne Barth and Menno DT De Jong. The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. *Telematics and informatics*, 34(7):1038–1058, 2017.

[13] Syed Raza Bashir, Shaina Raza, and Vojislav Misic. A narrative review of identity, data and location privacy techniques in edge computing and mobile crowdsourcing. *Electronics*, 13 (21):4228, 2024.

[14] BC Campus. Unlocking the power of ai face swap technology. `https://pressbooks.bccampus.ca/nexus/part/unlocking-the-power-of-ai-face-swap-technology/`, 2024. Accessed on November 12, 2024.

[15] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, and et al. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018. URL `http://arxiv.org/abs/1810.01943`.

[16] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023(5):1–95, 2023. ISSN 2835-8856.

[17] Besedo. Ai and content moderation: How the regulatory landscape is shaping up. *Besedo Blog*, May 2024. URL https://besedo.com/blog/ai-and-content-moderation-how-the-regulatory-landscape-is-shaping-up/. Accessed on November 12, 2024.

[18] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Percy Liang, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2022. doi: 10.48550/arXiv.2108.07258. Version v3 submitted July 12, 2022.

[19] Aikaterini Boura. Risk-based governance and ai sandboxes in the eu: opportunities and limits. In Martin Ebers and Orla Lynskey, editors, *Research Handbook on EU Artificial Intelligence Law*, pages 233–250. Edward Elgar Publishing, 2024. doi: 10.4337/9781800379222.00023.

[20] Joanna J Bryson. Patiency is not a virtue: Ai and the design of ethical systems. In *Ethics of Artificial Intelligence*, pages 197–205. Oxford University Press, 2018. doi: 10.1093/oso/9780190905033.003.0013.

[21] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56, Vancouver, BC, Canada, 2019. IEEE. doi: 10.1109/VAST47406.2019.8986948.

[22] Longbing Cao. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55(3):1–38, 2022.

[23] Capgemini Research Institute. Generative ai in organizations. Research report, Capgemini, 2024. URL https://www.capgemini.com/wp-content/uploads/2024/07/Generative-AI-in-Organizations-Refresh-1.pdf. [Accessed 21-10-2024.

[24] Corinne Cath. Governing artificial intelligence: Upholding human rights & dignity. *London School of Economics Law Review*, 2018(3):1–26, 2018. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3265913.

[25] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7), apr 2024. ISSN 0360-0300. doi: 10.1145/3616865. URL https://doi.org/10.1145/3616865.

[26] Manish Chadha. The dangers of voice cloning and how to combat it. *The Conversation*, June 2024. URL https://theconversation.com/the-dangers-of-voice-cloning-and-how-to-combat-it-239926. Accessed on November 12, 2024.

[27] Bhanu Chander, Chinju John, Lekha Warrier, and Kumaravelan Gopalakrishnan. Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness. *ACM Computing Surveys*, 2024.

[28] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4024, 2023.

[29] ClarityNLP. Claritynlp overview. https://claritynlp.readthedocs.io/en/latest/user_guide/intro/overview.html, 2024. [Accessed: 2024-09-25.

[30] C/AISC Artificial Intelligence Standards Committee. IEEE Standards Association — standards.ieee.org. https://standards.ieee.org/ieee/2894/11296/, 2024. [Accessed 05-09-2024].

[31] Ulises Cortés, Atia Cortés, and Cristian Barrué. Trustworthy ai. the ai4eu approach. *Proceedings of Science*, 372:1–14, 2020. doi: 10.22323/1.372.0014.

[32] Vinícius G Costa and Carlos E Pedreira. Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, 56(5):4765–4800, 2023.

[33] Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems, 2024.

[34] AI Incident Database. AI Incident Database, 2023. URL `https://incidentdatabase.ai/`. [Accessed 01-10-2024].

[35] Deepgram. Explain like i'm five, september 2024. URL `https://deepgram.com/ai-apps/explain-like-i'm-five`. [Accessed 30-09-2024].

[36] Department of Science, Innovation & Technology. A call for views on the cyber security of AI. `https://www.gov.uk/government/calls-for-evidence/cyber-security-of-ai-a-call-for-views/a-call-for-views-on-the-cyber-security-of-ai`, 2024. [Accessed 05-09-2024].

[37] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896, 2023.

[38] Virginia Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer, 2019. doi: 10.1007/978-3-030-30371-6.

[39] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[40] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, 55(9), jan 2023. ISSN 0360-0300. doi: 10.1145/3561048. URL `https://doi.org/10.1145/3561048`.

[41] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012. doi: 10.1145/2090236.2090255.

[42] Marc TJ Elliott, Deepak P, and Muiris Maccarthaigh. Evolving generative ai: Entangling the accountability relationship. *Digital Government: Research and Practice*, 5(4):1–15, 2024.

[43] ETH Zurich. Safe ai - laboratory for trustworthy ai systems, 2024. URL `https://safeai.ethz.ch/`. [Accessed: 2024-10-16].

[44] EvidentlyAI. Evidently, 2024. URL `https://github.com/evidentlyai/evidently`. [Accessed 30-09-2024].

[45] Marie Farrell, Matt Luckcuck, Laura Pullum, Michael Fisher, Ali Hessami, Danit Gal, Zvikomborero Murahwi, and Ken Wallace. Evolution of the IEEE P7009 Standard: Towards Fail-Safe Design of Autonomous Systems. In *2021 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, pages 401–406, IEEE, 2021. Wuhan, China. doi: 10.1109/ISSREW53611.2021.00109.

[46] Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. Ted-spad: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13598–13609, Paris, France, 2023. IEEE.

[47] Jessica Fjeld, Hannah Achten, Ewan Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Nature Machine Intelligence*, 2(6):365–373, 2020. doi: 10.1038/s42256-020-0210-6.

[48] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018.

[49] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Pierre Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4):689–707, 2018. doi: 10.1007/s11023-018-9482-5.

[50] Center for Data Science and Public Policy University of Chicago. Bias and fairness audit toolkit, 2018. URL `http://aequitas.dssg.io/`. [Accessed 01-10-2024].

[51] AI Verify Foundation. Model ai governance framework for generative ai, 2024. URL `https://aiverifyfoundation.sg/resources/mgf-gen-ai/`. [Accessed 01-10-2024].

[52] Joshua Freeman, Chloe Rippe, Edoardo Debenedetti, and Maksym Andriushchenko. Exploring memorization and copyright violation in frontier llms: A study of the new york times v. openai 2023 lawsuit. In *Neurips Safe Generative AI Workshop*, pages 1–8, Vancouver, Canada, 2024. NeurIPS.

[53] Boris A Galitsky. Truth-o-meter: Collaborating with llm in fighting its hallucinations, 2023.

[54] Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*, 2024.

[55] Urs Gasser and Virgilio AF Almeida. A layered model for ai governance. *IEEE Internet Policy Review*, 6(3):1–16, 2017.

[56] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[57] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL `https://aclanthology.org/2020.findings-emnlp.301`.

[58] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89, Turin, Italy, 2018. IEEE, IEEE.

[59] Ellen Goodman and Julia Chen. The european union's proposed ai regulation: A framework for accountability. *AI and Ethics*, 1(1):5–16, 2020. doi: 10.1007/s43681-020-00005-w.

[60] Google. Google differential privacy library. `https://github.com/google/differential-privacy`, 2021. Accessed: 2025-08-29.

[61] Google AI. Responsible AI practices, 2024. URL `https://ai.google/responsibility/responsible-ai-practices/`. Accessed: 2024-11-18.

[62] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42, 2018.

[63] Muhammad Usman Hadi, Rizwan Qureshi, Ayesha Ahmed, and Nadeem Iftikhar. A lightweight corona-net for covid-19 detection in x-ray images. *Expert Systems with Applications*, 225:120023, 2023.

[64] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage, 2023.

[65] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, et al. Robustness and explainability of artificial intelligence. *Publications Office of the European Union*, 207:2020, 2020.

[66] Ahmed Rizvan Hasan. Artificial intelligence (ai) in accounting & auditing: A literature review. *Open Journal of Business and Management*, 10(1):440–465, 2021.

[67] Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. Privacy preservation in blockchain based iot systems: Integration issues, prospects, challenges, and future research directions. *Future Generation Computer Systems*, 97:512–529, 2019.

[68] Yuchen He et al. Harnessing attention maps for explainability in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[69] The White House. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence, 2023. [Accessed 01-10-2024].

[70] Changwu Huang, Zeqi Zhang, Bifei Mao, and Xin Yao. An Overview of Artificial Intelligence Ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819, 2023. doi: 10.1109/TAI.2022.3194503.

[71] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation. *Artificial Intelligence Review*, 57(7):175, 2024.

[72] Yue Huang, Qihui Zhang, Philip S. Y, and Lichao Sun. Trustgpt: A benchmark for trustworthy and responsible large language models, 2023.

[73] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Position: Trustllm: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20166–20270, Vienna, Austria, 2024. PMLR.

[74] Humane Inc. Humane - technology built for people, 2024. URL `https://humane.com/`. [Accessed 2024-10-16].

[75] Tsuyoshi Idé and Naoki Abe. Generative perturbation analysis for probabilistic black-box anomaly attribution. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 845–856, 2023.

[76] Google Inc. Google's secure ai framework, 2024. URL `https://safety.google/cybersecurity-advancements/saif/`. [Accessed 01-10-2024].

[77] Pixel inc. Betterdata.ai : Programmatic synthetic data that is faster, safer and better, 2024. [Accessed 30-10-2024].

[78] Pymetrics Inc. Audit ai, 2020. URL `https://github.com/pymetrics/audit-ai/tree/master`. [Accessed 01-10-2024].

[79] International Telecommunication Union (ITU). Ai for good. `https://aiforgood.itu.int/`, 2024. [Accessed: 2024-09-25.

[80] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019. doi: 10.1038/s42256-019-0088-2.

[81] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2):1–38, 2022.

[82] Krishnaram Kenthapadi, Himabindu Lakkaraju, and Nazneen Rajani. Generative ai meets responsible ai: Practical challenges and opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5805–5806, NY, USA, 2023. ACM.

[83] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying llm safety against adversarial prompting, 2023.

[84] Joakim Laine, Matti Minkkinen, and Matti Mäntymäki. Ethics-based ai auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Information & Management*, 61(5):103969, 2024.

[85] Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.

[86] Anna Leschanowsky, Silas Rech, Birgit Popp, and Tom Bäckström. Evaluating Privacy, Security, and Trust Perceptions in Conversational AI: A Systematic Review. *Computers in Human Behavior*, 159:108344, 2024. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2024.108344. URL `https://www.sciencedirect.com/science/article/pii/S0747563224002127`.

[87] David Leslie. Explaining decisions made with ai, 2022.

[88] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. SALAD-bench: A hierarchical and comprehensive safety benchmark for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.235. URL `https://aclanthology.org/2024.findings-acl.235`.

[89] Percy Liang, Rishi Bommasani, and Tony Lee et. al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 1(1):1–162, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=iO4LZibEqW`. Featured Certification, Expert Certification.

[90] Q Vera Liao, Daniel Gruen, and Shion Guha Miller. Closing the gap: Towards effective human-ai teams in iterative decision-making tasks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[91] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LVI*, page 386–403, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72991-1. doi: 10.1007/978-3-031-72992-8_22. URL `https://doi.org/10.1007/978-3-031-72992-8_22`.

[92] Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, et al. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities, 2024.

[93] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering. *ACM Computing Surveys*, 56(7):1–35, 2024.

[94] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[95] Juliette Mattioli, Henri Sohier, Agnès Delaborde, Kahina Amokrane-Ferka, Afef Awadid, Zakaria Chihani, Souhaiel Khalfaoui, and Gabriel Pedroza. An overview of key trustworthiness attributes and kpis for trusted ml-based systems engineering. *AI and Ethics*, 4(1):15–25, 2024.

[96] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024.

[97] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

[98] Christian Meske, Babak Abedin, Mathias Klier, and Fethi Rabhi. Explainable and responsible artificial intelligence. *Electronic Markets*, 32(4):2103–2106, 2022.

[99] David Mhlanga. Open ai in education, the responsible and ethical use of chatgpt towards lifelong learning. In *FinTech and artificial intelligence for sustainable development: The role of smart technologies in achieving development goals*, pages 387–409. Springer, 2023.

[100] Microsoft. Responsible AI, 2024. URL https://www.microsoft.com/en-ca/ai/responsible-ai. Accessed: 2024-11-18.

[101] Microsoft. Responsible ai principles and approach, 2024. URL https://www.microsoft.com/en-us/ai/principles-and-approach. [Accessed 01-10-2024].

[102] Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Uncertainty in Artificial Intelligence*, pages 129–139. PMLR, 2021.

[103] Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. Explainable Artificial Intelligence: A Comprehensive Review. *Artificial Intelligence Review*, 55(5):3503–3568, Jun 2022. ISSN 1573-7462. doi: 10.1007/s10462-021-10088-y. URL https://doi.org/10.1007/s10462-021-10088-y.

[104] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 220–229, 2019.

[105] Surbhi Mittal, Kartik Thakral, Richa Singh, Mayank Vatsa, Tamar Glaser, Cristian Canton Ferrer, and Tal Hassner. On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare. *Nature Machine Intelligence*, 6(8):936–949, 2024.

[106] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11):501–507, 2019. doi: 10.1038/s42256-019-0114-4.

[107] Brent Mittelstadt. Principles for responsible ai: Towards a regulatory framework. *AI and Ethics*, 1:1–16, 2021.

[108] MLBench. Mlbench: Distributed machine learning benchmark, 2020. URL https://mlbench.github.io/. [Accessed 30-09-2024].

[109] Hung Truong Thanh Nguyen, Hung Quoc Cao, Khang Vo Thanh Nguyen, and Nguyen Dinh Khoi Pham. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, pages 1–6, 2021.

[110] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009. ISBN 9780804752374.

[111] Stany Nzobonimpa and Jean-François Savard. Ready but irresponsible? analysis of the government artificial intelligence readiness index. *Policy & Internet*, 15(3):397–414, 2023.

[112] Government of Canada. Voluntary code of conduct on the responsible development and management of advanced generative ai systems, 2023. [Accessed 01-10-2024].

[113] LLC. OneTrust. Trust intelligence platform, 2023. URL `https://www.onetrust.com/`.

[114] Alina Oprea and Apostol Vassilev. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2023.

[115] Organisation for Economic Co-operation and Development (OECD. OECD AI Principles: Accountability (Principle 1.5), 2024. URL `https://oecd.ai/en/dashboards/ai-principles/P9`. [Accessed 01-10-2024].

[116] World Health Organization et al. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models, 2024.

[117] Celso Cancela Outeda. The eu's ai act: a framework for collaborative governance. *Internet of Things*, 27(1):101291, 2024.

[118] Richard Owen, Phil Macnaghten, and Jack Stilgoe. Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy*, 39(6): 751–760, 2012.

[119] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 2023.

[120] Frank Pasquale and Haochen Sun. Consent and compensation: Resolving generative ai's copyright crisis. *Cornell Legal Studies Research Paper Forthcoming*, 110(7):207–247, 2024.

[121] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[122] Dorian Peters, Karina Vold, Diana Robinson, and Rafael A Calvo. Responsible ai—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, 1(1): 34–47, 2020.

[123] Yulu Pi. Beyond xai: Obstacles towards responsible ai. *arXiv preprint arXiv:2309.03638*, 2023.

[124] Nineta Polemi, Isabel Praça, Kitty Kioskli, and Adrien Bécue. Challenges and Efforts in Managing AI Trustworthiness Risks: A State of Knowledge. *Frontiers in Big Data*, 7:1381163, 2024.

[125] Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141, 2020.

[126] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020.

[127] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[128] NL Rane and M Paramesha. Explainable artificial intelligence (xai) as a foundation for trustworthy artificial intelligence. *Trustworthy Artificial Intelligence in Industry and Society*, pages 1–27, 2024.

[129] Anand S. Rao. Responsible ai: Operationalizing principles through key performance indicators. *AI and Ethics*, 2(3):455–467, 2022. doi: 10.1007/s43681-021-00068-2.

[130] Shaina Raza and Chen Ding. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4): 335–362, 2022.

[131] Shaina Raza, Oluwanifemi Bamgbose, Shardul Ghuge, Fatemeh Tavakoli, and Deepak John Reji. Developing safe and responsible large language models–a comprehensive framework, 2024.

[132] Shaina Raza, Muskan Garg, Deepak John Reji, Syed Raza Bashir, and Chen Ding. Nbias: A natural language processing framework for bias identification in text. *Expert Systems with Applications*, 237:121542, 2024.

[133] Shaina Raza, Mizanur Rahman, Safiullah Kamawal, Armin Toroghi, Ananya Raval, Farshad Navah, and Amirmohammad Kazemeini. A comprehensive review of recommender systems: Transitioning from theory to practice. *arXiv preprint arXiv:2407.13699*, 2024.

[134] Shaina Raza, Mizanur Rahman, and Michael R Zhang. Beads: Bias evaluation across domains. *arXiv preprint arXiv:2406.04220*, 2024.

[135] Shaina Raza, Deepak John Reji, and Chen Ding. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*, 17(1):39–59, 2024.

[136] Shaina Raza, Caesar Saleh, Emrul Hasan, Franklin Ogidi, Maximus Powers, Veronica Chatrath, Marcelo Lotif, Roya Javadi, Anam Zahid, and Vahid Reza Khazaie. Vilbias: A framework for bias detection using linguistic and visual cues. *arXiv preprint arXiv:2412.17052*, 2024.

[137] Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, and Mubarak Shah. Vldbench: Vision language models disinformation detection benchmark, 2025. URL `https://arxiv.org/abs/2502.11361`.

[138] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[139] Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL `https://aclanthology.org/2024.naacl-long.301`.

[140] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[141] Hannah Ruschemeier. Ai regulatory sandboxes in the european union: experimental governance in practice. *Law, Innovation and Technology*, 15(2):214–240, 2023. doi: 10.1080/17579961.2023.2251761.

[142] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[143] Malak Sadek, Emma Kallina, Thomas Bohné, Céline Mougenot, Rafael A Calvo, and Stephen Cave. Challenges of responsible ai in practice: scoping review and recommended actions. *AI & SOCIETY*, 0(0):1–17, 2024.

[144] Waddah Saeed and Christian Omlin. Explainable AI (XAI): A Systematic Meta-survey of Current Challenges and Future Opportunities. *Knowledge-Based Systems*, 263:110273, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110273. URL `https://www.sciencedirect.com/science/article/pii/S0950705123000230`.

[145] Sonia Salman, Jawwad Ahmed Shamsi, and Rizwan Qureshi. Deep fake generation and detection: Issues, challenges, and solutions. *IT Professional*, 25(1):52–59, 2023. doi: 10.1109/MITP.2022.3230353.

[146] Mehrnoosh Sameki, Sarah Bird, and Kathleen Walker. Interpretml: A toolkit for understanding machine learning models, 2020. URL `https://interpret.ml/`. [Accessed 30-09-2024].

[147] Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts, 2024.

[148] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. Principles to practices for responsible ai: closing the gap. *arXiv preprint arXiv:2006.04707*, 2020.

[149] Johan Schot and W Edward Steinmueller. New orientations towards responsible innovation: Collaborative and reflexive governance. *Research Policy*, 47(9):1581–1589, 2018.

[150] Reva Schwartz, Apostol Vassilev, Kristen K. Greene, Lori Perine, Andrew Burt, and Patrick Hall. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, 2022-03-15 2022. URL `https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464`.

[151] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[152] Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models, 2024.

[153] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2020.

[154] Peter Slattery, Alexander K Saeri, Emily AC Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence, 2024.

[155] Colin Smith, Ewen Denney, and Ganesh Pai. Hazard contribution modes of machine learning components. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2020.

[156] Thinking Stack. Responsible ai, 2024. URL `https://www.thinkingstack.ai/responsible-ai`. [Accessed: 2024-08-28.

[157] Jack Stilgoe, Richard Owen, and Phil Macnaghten. Developing a framework for responsible innovation. *Research Policy*, 42(9):1568–1580, 2013. doi: 10.1016/j.respol.2013.05.008.

[158] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.342.

[159] Cole Stryker. What is responsible AI?, February 2024. URL `https://www.ibm.com/topics/responsible-ai`. Accessed: 2024-11-18.

[160] IST/33/1 Information Security Management Systems. ISO/IEC TS 27022:2021. Technical Report ISO/IEC TS 27022:2021, ISO - International Organization for Standardization, 3 2021. URL `https://www.iso.org/obp/ui/en/#iso:std:61004:en`.

[161] Fabio Tango and Marco Botta. Real-time detection system of driver distraction using machine learning. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):894–905, 2013.

[162] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models, 2020. URL https://www.aclweb.org/anthology/2020.emnlp-demos.15.

[163] TensorFlow. Tensorflow privacy, 2023. URL https://github.com/tensorflow/privacy. [Accessed 01-10-2024].

[164] Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakal, Rao M Anwer, Michael Felsberg, Tim Baldwin, Eric P Xing, and Fahad Shahbaz Khan. Mobillama: Towards accurate and lightweight fully transparent gpt. *arXiv preprint arXiv:2402.16840*, 2024.

[165] TNO. Ai research - tno, 2024. URL https://www.tno.nl/en/digital/artificial-intelligence/ai-research/. [Accessed: 2024-10-16.

[166] Jon Truby. Ai regulatory sandboxes: a bridge between innovation and regulation. *Computer Law & Security Review*, 46:105710, 2022. doi: 10.1016/j.clsr.2022.105710.

[167] UK Government Information Commissioner's Office. Guidance on the AI auditing framework, 2020. URL https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf. [Accessed 01-10-2024].

[168] UKRI Trustworthy Autonomous Systems (TAS) Hub. Ukri trustworthy autonomous systems (tas) hub. https://tas.ac.uk/, 2024. [Accessed: 2024-09-25.

[169] European Union. General data protection regulation, 2018. URL https://gdpr-info.eu/. [Accessed 01-10-2024].

[170] Unitary. Our context aware ai understands the nuances of each piece of content, 2023. URL https://www.unitary.ai/product. [Accessed 01-10-2024].

[171] U.S. Government Accountability Office. Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities, 2021. URL https://www.gao.gov/products/gao-21-519sp. [Accessed 01-10-2024].

[172] Benjamin Van Giffen, Dennis Herhausen, and Tobias Fahse. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, 144:93–106, 2022.

[173] Aimee Van Wynsberghe. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218, 2021.

[174] David Vazquez et al. Heatmaps for interpretability in large language models. In *Proceedings of the International Conference on Learning Representations*, 2024.

[175] Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act. *Computer Law Review International*, 22(4):97–112, 2021. doi: 10.2139/ssrn.3896853.

[176] Bertie Vidgen, Adarsh Agrawal, and Ahmed M. Ahmed et. al. Introducing v0.5 of the ai safety benchmark from mlcommons, 2024.

[177] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[178] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone D. Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):233, 2020. doi: 10.1038/s41467-019-14108-y.

[179] René Von Schomberg. Responsible innovation: Its role in an era of technological change. *Journal of Responsible Innovation*, 9(2):185–203, 2022. doi: 10.1080/23299460.2022.2050456.

[180] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017. doi: 10.1093/idpl/ipx005.

[181] Fabian Walke, Lars Bennek, and Till J. Winkler. Artificial intelligence explainability requirements of the ai act and metrics for measuring compliance. In *Proceedings of the 18th International Conference on Wirtschaftsinformatik (WI 2023)*, Paderborn, Germany, September 2023. AIS Electronic Library (AISeL). URL `https://aisel.aisnet.org/wi2023/77`.

[182] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pages 1–110, New Orleans, Louisiana, United States of America, 2023. NeurIPS. URL `https://openreview.net/forum?id=kaHpo8OZw2`.

[183] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. Farsight: Fostering responsible ai awareness during ai application prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–40, Honolulu, HI, USA, 2024. ACM.

[184] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems, 2023.

[185] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models, 2021.

[186] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. " do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 7–9, Paris, France, 2019. ACM.

[187] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(01):56–65, 2020. ISSN 1941-0506.

[188] Alan Winfield. Ethical standards in robotics and ai. *Nature Electronics*, 2(2):46–48, 2019.

[189] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9291–9298, Ottawa, ON, Canada, 2023. ACM.

[190] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8*, pages 563S–574, Germany, 2019. Springer, Springer.

[191] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[192] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples, 2023.

[193] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.

[194] Xuesong Zhai, Xiaoyan Chu, Ching Sing Chai, Morris Siu Yung Jong, Andreja Istenic, Michael Spector, Jia-Bao Liu, Jing Yuan, and Yan Li. A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity*, 2021(1):8812542, 2021.

[195] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. doi: 10.48550/arXiv.2309.07045. URL `https://arxiv.org/abs/2309.07045`.

[196] Christopher T Zirpoli. Generative artificial intelligence and copyright law, 2023.

[197] James Zou and Londa Schiebinger. Ai can be sexist and racist—it's time to make it fair, 2018.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly present the scope of our survey.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We acknowledge that, as a survey, the work does not provide new empirical experiments but instead synthesizes and evaluates existing literature.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper is a survey and does not introduce new theoretical results, theorems, or formal proofs. Instead, it synthesizes and critiques existing technical, governance, and regulatory literature.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This paper does not introduce new experiments or models.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The paper is a literature survey and does not involve human or animal subjects, sensitive data collection, or potentially harmful interventions. All referenced works are properly cited, and the study fully adheres to the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive and negative societal impacts of Responsible Generative AI.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release new datasets or models. Instead, it surveys existing work and highlights safeguard mechanisms proposed in the literature.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All referenced works,models, and governance frameworks are cited to their original sources with appropriate bibliographic references (see References section). Since no external code or datasets are directly reused, there are no additional licensing concerns beyond proper scholarly citation.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve human-subject research or crowdsourcing studies.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve experiments with human subjects and therefore does not require IRB approval.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

Justification: This survey does not use LLMs as part of its methodology or results. Any AI tools used were limited to standard writing/editing assistance and did not affect the scientific content or originality of the work.