Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability

Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing
firstname.lastname@data61.csiro.au
CSIRO's Data61, Australia

ABSTRACT

Artificial Intelligence (AI), particularly through the advent of large-scale generative AI (GenAI) models such as Large Language Models (LLMs), has become a transformative element in contemporary technology. While these models have unlocked new possibilities, they simultaneously present significant challenges, such as concerns over data privacy and the propensity to generate misleading or fabricated content. Current frameworks for Responsible AI (RAI) often fall short in providing the granular guidance necessary for tangible application, especially for *Accountability*—a principle that is pivotal for ensuring transparent and auditable decision-making, bolstering public trust, and meeting increasing regulatory expectations. This study bridges the *Accountability gap* by introducing our effort towards a comprehensive metrics catalogue, formulated through a systematic multivocal literature review (MLR) that integrates findings from both academic and grey literature. Our catalogue delineates process metrics that underpin procedural integrity, resource metrics that provide necessary tools and frameworks, and product metrics that reflect the outputs of AI systems. This tripartite framework is designed to operationalize Accountability in AI, with a special emphasis on addressing the intricacies of GenAI.

CCS CONCEPTS

• **Software and its engineering**;

KEYWORDS

Responsible AI, Accountable AI, Risk assessment, Generative AI

ACM Reference Format:
Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing. 2024. Towards a Responsible AI Metrics Catalogue: A Collection of Metrics for AI Accountability. In *Conference on AI Engineering Software Engineering for AI (CAIN 2024), April 14–15, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3644815.3644959

1 INTRODUCTION

Artificial Intelligence (AI) serves as the linchpin of contemporary technological innovation, with large-scale generative AI (GenAI) models, exemplified by Large Language Models (LLMs), at the forefront of this paradigm shift [9]. These models have demonstrated remarkable proficiency in generating content across modalities, including text, imagery, and code. However, they also precipitate complex legal, ethical and operational challenges, such as data privacy breaches [45, 99], lack of transparency [12], and the erosion of informational integrity, epitomized by the generation of "hallucinations"—erroneous or misleading information produced by the models [77]. These challenges gain further gravitas in critical sectors, where the misuse of AI could lead to biased decision-making and the contentious deployment of dual-use biotechnologies, as underpinned by a recent *US Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence* [89].

The imperative for the practical implementation of responsible AI (RAI) is accentuated by the proliferation of high-level principles published by organizations or governments worldwide [67]. Despite providing guidance, these frameworks often lack the granularity necessary for practical implementation, falling short in providing concrete and actionable solutions [96]. In this context, accountability emerges as a pivotal element within RAI [4]. It serves as a fundamental principle that for establishing transparent, fair, and ethical AI systems. Through responsibility allocation, accountability enhances governance practices and contributes to the auditability and trustworthiness of AI systems' decision-making processes. The significance of accountability in AI is further underscored by evolving global standards and legislation, such as the proposed EU AI Act [18] and the US Blueprint for an AI Bill of Rights [90]. These legislative developments highlight the necessity for actionable and concrete guidelines that operationalize accountability in AI.

In light of these challenges, our research underscores the pivotal role of a process-oriented approach that encompasses both technical and socio-technical dimensions. As elucidated by Raji et al. [82], procedural justice is predicated on the legitimacy of outcomes derived from equitable and comprehensive processes. This principle is foundational to the development of robust frameworks that enable independent audits, adherence to established standards, and enhanced compliance [54]. Similarly, the capAI framework [27], which operationalizes the proposed EU AI Act's directives, emphasizes a process-oriented perspective of AI systems throughout their lifecycle. Our study aligns with this paradigm, concentrating on the development of process metrics as the cornerstone procedural guidelines for the operationalization of AI accountability.

Anchored in a systematic multivocal literature review (MLR) that combines Systematic Literature Review (SLR) and Grey Literature Review (GLR), we aim to transition from aspirational RAI principles and answer the following research question (RQ): **What are the definitive, actionable metrics for AI risk management with respect to Accountability?** Drawing inspiration from an established software metrics framework [24], we introduce a dedicated AI-centric **system-level** metrics framework specifically tailored

CAIN 2024, April 14–15, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0591-5/24/04…$15.00
https://doi.org/10.1145/3644815.3644959

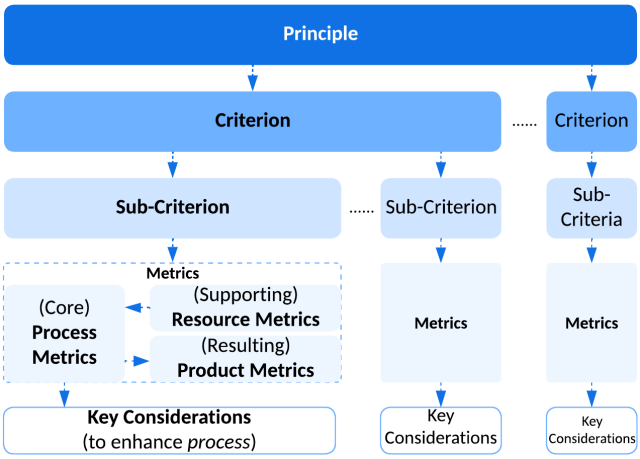

Figure 1: Structure Overview

to the principle of Accountability and sensitive to the intricacies of GenAI. This framework delineates metrics into three interrelated categories: **process metrics**, which establish foundational procedural guidelines; **resource metrics**, which encompass the necessary tools and frameworks; and **product metrics**, which include the resultant artifacts. While our current metrics primarily offer a binary (yes/no) evaluation, this work represents an initial step towards developing a more sophisticated metrics catalog. Our approach also extends beyond these binary evaluations to include qualitative considerations of these processes, as depicted in Fig. 1. This comprehensive approach lays the groundwork for an integrated RAI framework that covers all principal RAI principles. The primary contributions of this paper are as follows:

- **Development of a Process-centric Metrics Catalogue for AI Accountability**: We introduce a metrics catalogue that operationalizes accountability in AI systems, with a particular focus on GenAI. This catalogue provides essential procedural guidelines for embedding accountability in AI.
- **Categorization of AI Accountability Metrics**: We categorize AI accountability metrics into process, resource, and product metrics. This tripartite categorization provides a structured approach to operationalize AI accountability.
- **Foundation for a Comprehensive RAI Framework**: This research lays the groundwork for a more expansive framework that integrates all key RAI principle, marking an important step forward in the discourse on RAI practices.

2 BACKGROUND AND CHALLENGES

2.1 Accountability

2.1.1 The Three Facets of Accountability. Accountability, a concept integral to various domains from governance and law to technology and science, is foundational to ethical practice [37, 84]. In the legal realm, accountability is a well-defined term, emphasizing the necessity for individuals or entities to be responsible for their actions and consequences [48]. Under certain data protection legislation such as the EU's General Data Protection Regulation (GDPR), accountability is articulated as a requirement wherein data controllers need to assume responsibility for compliance and possess demonstrable evidence of their adherence to the regulatory standards and provisions. Expanding beyond its legal roots, accountability in a broader sense revolves around the answerability [13, 69]. This principle demands that actors justify their actions to an overseeing authority, which then has the power to impose consequences based on these justifications and the actor's performance.

The broader interpretation of accountability manifests in three interconnected facets: Responsibility, Auditability, and Redressability (see Fig. 2). **Responsibility** (Section 4) pertains to the attribution of ownership for actions to individuals or entities by establishing who is accountable to whom within a (cross-)organizational context. **Auditability** (Section 5) serves as the cornerstone for "accountable for what." It facilitates the systematic assessment of decisions and their outcomes against established criteria. This facet ensures that actions of responsible entities are *demonstrable* with supporting evidence–not only traceable but also defensible. **Redressability** (Section 6) addresses the aspect of "how to be accountable and rectify" after ascertaining "who is accountable for what." It involves the provision of mechanisms for remediation or compensation when actions lead to adverse outcomes. This facet completes the accountability cycle by providing avenues for remedy and reparation.

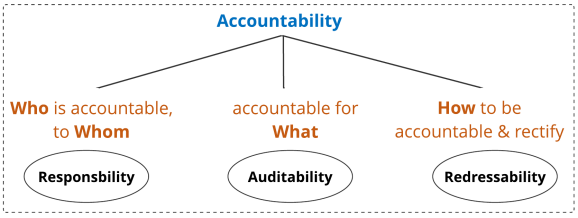

Figure 2: Three Facets of Accountability

These elements ensure that individuals and organizations are not only expected to justify their actions but also that their justifications are backed by evidence and liable to appropriate consequences. This tripartite framework is essential for the ethical operation and maintenance of equitable systems within complex societal structures [8], highlighting the multifaceted nature of accountability.

2.1.2 Accountability and Transparency. The relationship between accountability and transparency is symbiotic. Transparency is not an adjunct but a foundational element that empowers accountability. Effective accountability is contingent upon the disclosure of relevant information, enabling stakeholders to scrutinize, understand, and, if necessary, contest decisions. This necessitates the provision of comprehensive provenance documentation, elucidating the design, development and decision-making pathways [47]. Moreover, transparency mandates the explicit delineation of responsibilities, which is vital to prevent the dilution of accountability and to ensure that responsible parties can be identified and held to account, especially from a cross-organizational supply chain perspective [17] where there could be shared accountability.

2.2 AI Accountability

Accountability within AI systems is not merely an extension of traditional accountability concepts but a distinct challenge due to

the autonomous and often opaque nature of AI decision-making processes. As AI becomes more integrated into societal functions, the imperative for robust accountability frameworks is underscored by a range of RAI guidelines and principles [43, 83]. Examples include both organizational (e.g., Microsoft [62], OECD [71]) and governmental (e.g., Australia [7], US [74], EU [75]) efforts.

2.2.1 AI Accountability Challenge. Built on Nissenbaum [68], Cooper et al. [20] revisited the barriers to algorithmic accountability:

Many Hands: The involvement of multiple parties in AI development and use, including reliance on third-party tools, disperses responsibility and complicates accountability.

"Bugs": AI "bugs" extend beyond typical software errors, encompassing data biases, modeling errors, and design flaws, often systemic and unpredictable due to ML's statistical nature.

The "computer" as scapegoat: Misattributing moral agency to AI systems obscures human responsibility, complicating meaningful accountability.

Ownership without liability: Firms often claim rights over AI assets while avoiding responsibility for their impacts, impeding transparency and independent auditing. This calls for stronger legal and ethical frameworks to ensure liability, such as [23].

2.2.2 Accountability Challenge in the GenAI Era. GenAI models, characterized by their extensive scale, complexity, and adaptability, introduces new dimensions to these established challenges:

"Many *More* Hands": The collaborative nature of GenAI model development, involving a broad spectrum of contributors from individuals to multinational corporations, significantly disperses responsibility. This dispersion is intensified when third parties modify or tailor these models for specific applications. The distribution of GenAI models via APIs or cloud platforms adds another layer of complexity to the accountability equation.

"Bugs" at Scale: The scale and complexity of GenAI models exacerbate the consequences of such "bugs", transforming them from mere technical faults into systemic challenges. This means that biases, errors, and unpredictable behaviors can have far-reaching consequences. These issues complicate AI accountability, as the probabilistic nature of GenAI models often precludes clear-cut explanations for their outputs.

The "Big Black Box" as Scapegoat: The complex nature of large GenAI models often leads to anthropomorphization, where these systems are perceived as more human-like. This perception, combined with their "black box" nature, can be misleadingly used to deflect accountability from human actors to the AI systems.

Ownership with Disavowed Liability: The proprietary nature of many GenAI models often leads to a dichotomy where organizations claim credit for successes but not liability for failures. This issue is compounded by the wide application of GenAI models and the lack of transparency that hampers independent auditing efforts.

2.3 Related Work

RAI Operationalization. Operationalizing RAI has garnered attention in both industry and academia, with a notable shift towards process-centric approaches. Singapore's AI Verify [3] stands out for its evaluation of AI systems against 12 RAI principles, including accountability, through a blend of process checks and technical tests.

Table 1: Comparison of Metrics Related Work

	NIST [92]	OECD [72]	Stanford [12]	AI Verify [3]	This Work
Focus	Model-level	Model-level	**System-level**	**System-level**	**System-level**
Accountability	No	No	No	**Yes**	**Yes**
GenAI	No	Partly	**Yes**, GenAI only	No	**Yes**

The EU's capAI project [27] is pioneering a conformity assessment procedure aligned with the EU AI Act. Credo AI [2] integrates process checks into its AI governance platform, aligning with policy requirements. UC Berkeley's taxonomy [16] aligns with US National Institute of Standards and Technology (NIST) AI RMF's RAI principles, focusing on organizational processes. Fraunhofer IAIS's work [32] provides concrete criteria for RAI assessment, while Raja et al. [83] propose processes for RAI assessment in banking. More specifically in AI accountability, however, discourse predominantly remains conceptual, with studies (e.g., [14, 22, 81, 85]) focusing on high-level theoretical aspects.

AI Measurement and Metrics The field of RAI measurement and metrics is rapidly evolving. Research on AI explainability [41] and fairness [15, 28] typically targets model-level requirements, highlighting a gap in system-level analysis. Notably, NIST's work [92] on AI metrics emphasizes model-level metrics like accuracy and bias, ignoring accountability. The OECD's RAI Metrics catalogue [72], while comprehensive, remains focused on model-level technical requirements and lacks metrics for accountability. The catalogue has specific metrics related to content generation tasks. Stanford's Foundation Model Transparency Index [12] offers a broader range of system-level metrics for assessing foundation model transparency, spanning from model construction to downstream usage.

Our research diverges by focusing on concrete, process-centric metrics for AI accountability, complemented by resource and product metrics (see Table 1). This approach offers system-level metrics applicable to both traditional AI and GenAI systems, providing concrete guidance for operationalizing AI accountability.

3 METHODOLOGY

Our research methodology unfolds in two sequential phases, as depicted in Fig. 3: a Multivocal Literature Review (MLR) followed by thematic coding. The MLR, adhering to established guidelines [30, 46], merges Systematic Literature Review (SLR) and Grey Literature Review (GLR) to ensure a holistic understanding of AI accountability practices. This blended approach mitigates the limitations inherent in using either SLR or GLR in isolation. While SLR alone may overlook emerging trends and practical insights, GLR exclusively might lack the methodological rigor of peer-reviewed academic literature. Our methodology, therefore, balances these aspects to provide a comprehensive and robust analysis.

3.1 Multivocal Literature Review (MLR)

3.1.1 Systematic Literature Review (SLR). **Search Strategy**: Our keyword choices aimed to ensure comprehensive coverage of the domain. The search string used was: *(AI OR ML OR "artificial intelligence" OR "machine learning" OR "large language model" OR LLM OR "Foundation Model" OR "Frontier Model" OR "Generative AI") AND (accountability OR accountable)*. Pilot tests ensured their suitability. The search string was used to search title only. Searches were

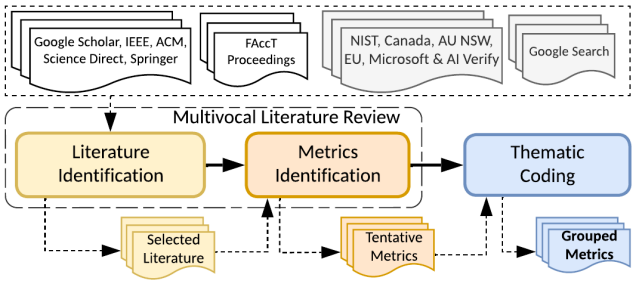

Figure 3: Methodology Overview

conducted on 26 September 2023. Searches spanned five databases: IEEE Xplore, ACM Digital Library, Science Direct, Springer, and Google Scholar, yielding 221 results. Additionally, we reviewed ACM Conference on Fairness, Accountability, and Transparency proceedings given its relevance to our research focus, focusing on titles including "Accountability" or "Accountable," resulting in 33 papers. The search was not constrained by publication date.

Inclusion & Exclusion, and Quality Assessment: We included sources based on: 1) their relevance to AI accountability after initial Abstract screening; 2) the inclusion of accountability practices/processes based on in-depth review. Exclusions were made for non-English sources and conference papers superseded by journal versions. he quality assessment was thorough, emphasizing the clarity of research aims, robustness of design, and the articulation of findings, contributions, and limitations. This rigorous process resulted in the selection of 40 pertinent studies. The distribution of these studies over recent years is as follows: 6 published in 2023, 14 in 2022, 12 in 2021, and 8 between 2017 and 2020, underscoring the growing interest and attention in AI accountability.

Data Extraction: Data extraction was conducted utilizing Zotero for review and annotation. We extracted information on processes for AI accountability (process metrics), tools and resources for AI accountability (resource metrics), and the resultant artefacts (product metrics). The extraction of resource and product metrics was contingent upon the identification of relevant process metrics.

3.1.2 Grey Literature Review (GLR). **Search Strategy**: Employing the same keywords as in SLR, our GLR involved searches on Google Search Engine, yielding 146 results. We also reviewed six internationally recognized AI governance/risk management frameworks based on their quality and representativeness [49]: EU Trustworthy AI Assessment List [75], NIST AI RMF [74], Canada Algorithmic Impact Assessment Tool [35], Australia NSW AI Assurance Framework [6], Microsoft Responsible AI Impact Assessment Template [58], and Singapore's AI Verify [3].

Inclusion and Exclusion Criteria & Quality Assessment & Data Extraction: The GLR adhered to the same rigorous criteria of Inclusion/Exclusion and Quality Assessment as the SLR, with an added emphasis on the credibility of grey literature sources [30]. This led to an initial selection of 7 studies, which was expanded to 24 through snowballing, acknowledging the interconnected nature of grey literature. The extracted data also focused on the same three types of metrics as in SLR. Similar to SLR, 5 of the studies were published in 2023, 7 in 2022, 4 in 2021, and 7 between 2018-2020, with 1 exception with no clearly stated publication date.

3.2 Thematic Coding

Our thematic coding process employed a hybrid approach, blending both deductive and inductive methods [25]. This approach leverages the strengths of both structured and emergent analysis, ensuring a comprehensive understanding of the data. This hybrid thematic coding approach allowed us to systematically categorize the extracted metrics into a structured yet flexible framework. The predefined criteria provided a clear direction for our analysis, while the emergent sub-criteria offered depth and detailed insights, leading to a robust and comprehensive understanding of AI accountability.

(1) **Deductive Coding**: We began with three predefined broad criteria (i.e., Responsibility, Auditability, Redressability), as described in Section 2.1.1. These criteria served as initial guides for categorizing the extracted metrics.
(2) **Emergent Sub-Criteria**: As we delved deeper into the coding process, we allowed for the emergence of sub-themes. These sub-themes were not predefined but were identified based on patterns, similarities, and differences in the data. This inductive aspect of our analysis allowed for a nuanced understanding of the metrics and their interrelations.
(3) **Refinement and Integration**: The emergent sub-criteria were continuously refined and integrated into the broader predefined themes. This iterative process ensured that our thematic structure accurately represented the complexities and nuances of the data.
(4) **Internal Validation**: To ensure the validity of our thematic structure, one author conducted the coding and three authors reviewed the results and provided feedback. Adjustments were made to make sure consensus among authors.

Upon completion of the MLR and thematic coding, our analysis distilled eleven key process metrics systematically categorized into five emergent sub-criteria, which align with three pre-defined overarching pillar criteria, as depicted in Table 2.

4 RESULTS - RESPONSIBILITY

Responsibility serves as a cornerstone of accountability in AI governance, specifying who is answerable for each phase of an AI system's lifecycle. It goes beyond mere task allocation to mandate that these responsible entities possess the necessary ethical awareness, technical knowledge, and expertise. This ensures competent decision-making, thereby strengthening accountability.

4.1 Sub-Criterion 1.1: RAI Oversight

This emphasizes the necessity for a well-defined *organizational structure* to oversee the AI lifecycle, from procurement and development to deployment and operations. Such a structure is pivotal in upholding high ethical standards and responsibility in AI solutions, both in-house and externally sourced.

4.1.1 ***Process Metric 1.1.1: Roles and Responsibilities.*** **Context and Importance**: The clear delineation of roles and responsibilities is critical for the ethical, responsible, and effective development, deployment, procurement, and governance of AI systems

Table 2: System-Level Metrics Catalogue for AI Accountability

Criteria	Sub-Criteria	Process Metrics	Key Considerations	Resource Metrics	Product Metrics
Responsibility	RAI Oversight	Roles and Responsibilities	• Comprehensive role clarity: - Design and development - Deployment and operations - Procurement and integration - Governance and compliance - AI as a service	• Soft laws (e.g., best practices, guidelines standards etc) • Hard laws (e.g., EU AI Act)	• Procedure Manuals • Contracts or agreements • Position descriptions • Recruitment practices • Workforce dev strategy
		RAI Governance Committee	• Multidisciplinary composition • Strategic leadership involvement		• Policy doc on Committee
		Organizational AI Risk Tolerance	• Tiered risk-based categorization • Balancing competing interests		• Policy doc on org's risk tolerance and mitigations
	RAI Competence	RAI Training	• Holistic training content • Targeted training for diverse roles • Adaptive and ongoing education		• Training certificates
		RAI Capability Assessment	• Multifaceted assessment • Standard alignment • Organizational RAI maturity • Continuous enhancement		• Assessment reports
Auditability	Systematic Oversight	Data Provenance	• Detailed data record-keeping • Data version control • Data integrity and risk mitigation • Legal and ethical compliance	• Soft laws (e.g., auditing guidelines and frameworks etc) • Hard laws (e.g., EU AI Act) • AI documentation tools (e.g., datasheets, model/system cards) • Technical tools (e.g., blockchain, knowledge graph)	• Provenance records • System features (e.g., auto-logging, version control
		Model Provenance	• Detailed model record-keeping • Model selection and validation • Model version control		
		System Provenance and Logging	• Detailed system record-keeping - System version control - Decision/Trade-off • Comprehensive operational logging - User interaction and system response - Incident and response - System configuration changes • Composition Management		• Provenance records (and logs) • System features (e.g., auto-logging, version control
	Compliance Checking	Auditing	• Diversified auditing strategy • Multi-dimensional audit techniques • Ethical and legal compliance • Regular audits • Verifiable audits • Audit-driven improvements		• Audit reports • Compliance certificates and licenses
Redressability	Redress-by-Design	Incident Reporting and Response	• Accessibility and Visibility • Structured Incident Management • Feedback Loop Integration	• Redundancy design case studies • Incident management tools	• Incident and response doc • System features (user feedback and report)
		Built-in Redundancy	• Multi-Modal Redundancy		• System features (redundant components/functionalities)

[56, 70, 85]. This clarity is not only vital for defining responsibilities but also for managing potential overlaps and conflicts among roles. By proactively addressing these challenges, organizations can ensure cohesive and effective operation across all stages of the AI lifecycle. This approach is universally necessary, extending from traditional AI to GenAI.

Key Consideration-Comprehensive Role Clarity:

- **Design and Development**: Formalize roles associated with the initial stages of AI systems, covering plan and design, data collection and (pre)processing, model development, comprehensive testing (validation and verification), and fine-tuning in the case of large-scale GenAI models.
- **Deployment and Maintenance**: Define roles pertinent to the deployment, system monitoring, and operational management (e.g., issue resolution, upgrades) [59].
- **Procurement and Integration**: Establish clear roles for evaluating, selecting, and integrating AI systems, focusing on legal compliance, security, and ethical alignment with organizational values [59, 73]. These should include both technical (e.g., developers) and non-technical (e.g., procurement specialists) roles.
- **Governance and Compliance** [26, 73]: Assign roles for policy development, ethical standards adherence, and regulatory compliance. Include roles for internal and external audits to ensure continuous ethical assessment.
- **Service Provision (AI as a Service)**: For AI systems offered as a service [42], formalize roles for managing shared responsibilities and accountability [17], particularly for GenAI, ensuring clear agreements on usage policies [78], data governance, and liability.

4.1.2 ⚙ ***Process Metric 1.1.2: AI Governance Committee.*** **Context and Importance**: The establishment of an AI Governance Committee, or a similar oversight body, is ethically and effectively managing AI systems [57], especially in the dynamic and complex realm of GenAI [29]. This committee should oversee AI systems throughout their lifecycle, from conception to decommissioning, ensuring consistent and responsible governance [80]. Its role extends to efficient decision-making, particularly in ethical and risk

management scenarios, necessitating regular collaboration with internal and external stakeholders, including regulatory bodies, to stay aligned with best practices and evolving regulations.

Key Considerations:

- **Multidisciplinary Composition**: The committee should comprise experts from diverse fields like law, ethics, academia, and technology [10]. This diversity enables a comprehensive approach to AI risk management, facilitating the reconciliation of varied perspectives and addressing the complex ethical and operational challenges of (Gen)AI systems [73].
- **Strategic Leadership Involvement**: Incorporating senior leadership in the committee, akin to models like Microsoft's Responsible AI Council [61], is crucial. This ensures that strategic decisions are informed by a wide range of organizational perspectives and are aligned with RAI practices.

4.1.3 ⚙ ***Process Metric 1.1.3: Organizational AI Risk Tolerance***. **Context and Importance**: Defining an organizational risk tolerance/appetite is a critical aspect of risk-based approach to AI risk management [74], particularly for GenAI. This risk tolerance delineates the operational boundaries for AI systems, aligning them with ethical standards and operational requirements. It is essential for navigating ethical dilemmas like balancing fairness with accuracy and managing data privacy across different regulatory landscapes. The EU AI Act's [18] risk classification framework (i.e., Unacceptable/High/Limited/Minimal or no risk) offers a valuable model for categorizing AI systems based on risk levels and imposing transparency obligations specific to GenAI.

Key Considerations:

- **Tiered Risk-based Categorization**: Adopting a tiered risk-based categorization, aligned with standards and regulations like the EU AI Act, is recommended. This categorization should consider the nature, scope, and purpose of the AI system, as well as the organization's unique circumstances. For instance, healthcare AI systems may have different risk tolerances than retail AI, reflecting their varying ethical and privacy concerns.
- **Balancing Competing Interests**: It's crucial to balance competing interests such as data protection and operational efficiency. While data protection is paramount, overly rigid policies can stifle innovation. Organizations should aim for a balance that accommodates both ethical considerations and operational efficiency [38].

4.2 Sub-Criterion 1.2: RAI Competence

RAI Competence is essential for individuals and organizations to effectively implement and uphold RAI principles. In the GenAI context, this need is amplified due to its unique challenges and the fast pace of technological evolution. Ensuring RAI understanding across all organizational levels, from technical staff to executive leadership, is critical for maintaining RAI practices.

4.2.1 ⚙ ***Process Metric 1.2.1: RAI Training***. **Context and Importance**: Effective RAI training is indispensable for fostering an organizational awareness and culture that prioritizes RAI development and deployment [11, 70, 75, 81]. In GenAI, the high stakes associated with potential missteps due to advanced capabilities necessitate comprehensive and continuous training. This training should address the current state of technology and ethics while anticipating future developments and challenges, ensuring that all stakeholders are prepared to make responsible decisions in a rapidly changing environment.

Key Considerations:

- **Holistic Training Content**: Training should be holistic [81], including technical aspects, ethical considerations, legal compliance, and risk management. Specific modules could include data privacy, algorithmic fairness, and legal compliance.
- **Targeted Training for Diverse Roles**: Training should be customized to suit different roles within the organization. For instance, legal professionals may focus on compliance, while developers may delve into ethical coding practices [80]. Additional training should also be extended to procurement specialists and executive leadership, focusing on governance, risk management, and control [73]. This approach equips various stakeholders with the skills necessary to cultivate organizational RAI competence.
- **Adaptive and Ongoing Education** [21, 86]: The field of AI is continuously evolving, making ongoing education and adaptation essential. Training programs should not only include regular updates and feedback mechanisms but also adapt to new ethical challenges, technological advancements, and diverse global perspectives (i.e., future-proof).

4.2.2 ⚙ ***Process Metric 1.2.2: RAI Capability Assessment***. **Context and Importance**: Transitioning from being merely "trained" in RAI to being "competent" in their implementation is a critical evolution for organizations. This transition requires robust mechanisms to evaluate and certify individual staff competencies in RAI, as well as to assess and enhance the overall organizational RAI maturity. Such an approach ensures that both individuals and the organization are proficient in implementing RAI effectively, with a focus on empowering staff with necessary skills and equipping the organization for RAI practices.

Key Considerations:

- **Multifaceted Assessment**: Implement a variety of assessment methods to appraise RAI competence across diverse roles, including exams, practical exercises, and scenario analyses, covering both technical and ethical aspects of AI [81]. For instance, developers may be assessed on technical RAI knowledge, while legal professionals may be evaluated on compliance related aspects. This ensures a robust, role-specific measure of RAI competence.
- **Standard Alignment and Benchmarking**: The criteria and methodologies for assessment should align with evolving industry standards and best practices in AI ethics, drawing on guidelines from bodies like the IEEE SA [39].
- **Organizational Readiness and RAI Maturity**: Adopting or adapting an RAI maturity model (e.g., Microsoft RAI Maturity Model [93]) is critical to gauge and enhance the organization's overall capability in implementing RAI [53]. This model should evaluate aspects such as governance, technical

infrastructure, ethical alignment, and stakeholder engagement, serving as a tool for both diagnosis and continuous improvement.
- **Continuous Policy and Training Enhancement**: Leverage insights from the assessments to continuously refine training programs and organizational policies, ensuring they remain relevant in the rapidly changing AI landscape.

4.3 Resource & Product Metrics – Responsibility

Responsibility necessitates a strategic approach at the organizational level, integrating both resources and products to ensure ethical AI development and management. Resources provide the necessary tools and frameworks for ethical AI governance, while products represent the tangible outcomes of these efforts.

Essential ***Resources*** for *Responsibility*:

- **Soft Laws**: These encompass AI ethics best practices, guidelines, governance frameworks, and standards of care [20, 61], including both general principles and specific directives, such as procurement guidelines and aforementioned RAI maturity models. Examples include US NIST AI RMF [74] and IEEE 7000 Standard Series [39].
- **Hard Laws**: Adherence to legal frameworks is essential, especially for regulated sectors. These laws establish a legal baseline for responsible AI practices, ensuring compliance and ethical integrity.Legislation and regulations are crucial, especially for organizations in regulated sectors. Examples include the EU AI Act [18] and the US Blueprints for an AI Bill of Rights [90].

Products: The implementation of *Responsibility* requires the development of specific organizational policy **documentation** [3] related to different processes:

- **Roles and Responsibilities**: This includes internal procedure manuals, contracts, and written agreements that clearly delineate roles and responsibilities in AI system development and deployment [3, 26, 55, 69, 75]. It also encompasses position descriptions, recruitment practices, and workforce development strategies [73]. These documents are essential for establishing clear accountability and ethical guidelines within the organization.
- **AI Governance Committee**: Detailed documentation outlining the structure, function, and protocols of the Committee [3]. This includes its role in decision-making, oversight, and ensuring compliance with ethical standards.
- **Organizational AI Risk Tolerance**: Documents defining the organization's AI risk tolerance levels and mitigation strategies are crucial. This encompasses the assessment and management of different risk scenarios, aligning AI initiatives with the organization's overall risk framework.
- **RAI Training**: Certificates or other formal documents that validate the successful completion of RAI training, serving as a testament to the individual's proficiency in RAI.
- **RAI Capability Assessment**: Comprehensive reports that evaluate the RAI competencies at both individual and organizational levels, highlighting areas of strength and opportunities for improvement.

Weaving all these for *RAI Oversight* together, the organization can establish its own **RAI framework and structures**.

5 RESULTS - AUDITABILITY

Auditability enables thorough inspection and evaluation of AI systems, ensuring they conform to established standards and objectives. Auditability facilitates transparency and enables rigorous assessment, substantiating compliance and enhancing the ethical integrity of AI systems.

5.1 Sub-Criterion 2.1: Systematic Oversight

Systematic Oversight represents a comprehensive record-keeping and logging framework. It encompasses not just data and model lineage but extends to the entire AI system, encompassing development processes and operational dynamics. This holistic oversight is instrumental in maintaining system integrity and providing stakeholders with essential information for in-depth audits, ensuring every aspect of AI development and operation is subject to scrutiny.

5.1.1 ***Process Metric 2.1.1: Data Provenance.*** **Context and Importance**: Data Provenance is critical in AI risk management and governance, particularly for GenAI, where it bolsters transparency and ethical compliance. The dependency of GenAI on extensive datasets not only amplifies its importance but also presents unique challenges. Issues such as copyright, fair use, and privacy risks, as identified by Khan and Hanna [45], accentuate the necessity for robust Data Provenance. It's essential in verifying data integrity and ethical utilization in (Gen)AI systems, addressing these challenges to uphold the trustworthiness of AI.

Key Considerations:

- **Detailed Data Record-Keeping** [36, 56, 61, 66, 79]: Implement a thorough record-keeping system that documents all aspects of the data lifecycle, including sources, dataset characteristics, collection methods, preprocessing steps, and usage patterns. This comprehensive approach ensures that every aspect of data handling is auditable.
- **Data Version Control** [3]: Implement a robust data version control system to track changes and updates in the data used over time. This system should document each version's specific characteristics and modifications, providing clarity on the evolution of the data set.
- **Data Integrity and Risk Management** [27, 45, 66]: Ensure that the data used in AI systems is scrutinized for quality factors like representativeness, relevance, and accuracy and proactively address data-related risks.
- **Ethical and Legal Compliance** [27, 45]: The data provenance process could incorporate thorough assessments and documentation to ensure that an organization's data practices adhere to established ethical expectations and legal standards, like the EU GDPR.

5.1.2 ***Process Metric 2.1.2: Model Provenance.*** **Context and Importance**: Model Provenance, especially for GenAI, is indispensable for maintaining ethical integrity and transparency. The inherent complexity of GenAI models, characterized by sophisticated algorithms, necessitates meticulous documentation of their

development and deployment processes. This is crucial for ensuring that (Gen)AI models adhere to ethical standards and regulatory norms throughout their lifecycle.

Key considerations:

- **Detailed Model Record-Keeping** [21, 27, 56, 66, 98]: Maintain extensive records that capture model characteristics, algorithmic details, decision-making thresholds, and data pathways leading to final decisions is essential. This documentation should include intended and unintended usages, known limitations, and associated risks.
- **Model Selection and Validation** [21, 81]: Document the rationale for model selection, with an emphasis on fairness, explainability, and robustness, which are particularly pertinent in GenAI. A well-defined validation strategy is also essential, ensuring clear accountability for model design and implementation.
- **Model Version Control** [3]: Implement systematic version control is critical for tracking changes and updates made to the model over time, including each version's specific features and modifications.

5.1.3 ***Process Metric 2.1.3: System Provenance and Logging.*** **Context and Importance**: System Provenance and Logging embodies a holistic approach to understanding and overseeing AI systems at a systemic level, including both AI and non-AI elements [36, 55]. This consolidates the comprehensive documentation of an AI system's developmental history with the nuanced monitoring of its operational dynamics. It ensures a profound grasp of the entire lifecycle of AI systems, from their inception and architectural evolution to their real-time responses and operational adjustments.

Key Considerations:

- **Detailed System Record-Keeping** [4, 19, 26, 78, 98]:
  - **System Version Control**: Thorough records of the system's architecture and components, documenting the evolution of both AI and non-AI elements. Document their evolution (e.g., code versioning) over time.
  - **Decision/Trade-off**: Capture decisions made by human agents and automated processes that have influenced the overall AI system. Clarify the causal relationships between different stages of system development to provide a clear understanding of how decisions impact system behavior and outcomes.
- **Comprehensive Operational Logging** [35, 37, 60, 73]:
  - **User Interactions and System Responses**: Log all user queries/prompts and system responses to track usage patterns and system performance.
  - **Incident and Response Logging**: Maintain a detailed log of operational incidents, including system errors or failures, and document the steps taken for resolution. This log should align with the organization's risk management and incident response strategies.
  - **System Configuration Changes**: Keep records of all changes made to the system's configuration and operational parameters in real-time, ensuring a clear trail for audit and review.
- **Composition Analysis and Vulnerability Management** [73]: Establish mechanisms for the systematic identification and analysis of AI components, complementing traditional software composition analysis. This process should attend to the unique challenges inherent in integrated AI systems, such as due to more frequent updates or retraining.

5.2 Sub-Criterion 2.2: Compliance Checking

Compliance Checking serves as a critical extension of *Systematic Oversight*, taking the maintained provenance records and operational logs and subjecting them to rigorous evaluation and audit trails. This process enables stakeholders, such as the RAI Governance Committee, to conduct in-depth reviews and audits, ensuring the AI system's adherence to RAI.

5.2.1 ***Process Metric 2.2.1: Auditing.*** **Context and Importance**: Auditing is essential for affirming the ethical and legal compliance of AI systems, gaining particular significance in the rapidly advancing field of large-scale GenAI. World-leading AI and governance experts suggest auditing as a key measure for managing risks in these advanced systems [9]. Additionally, it's increasingly becoming a regulatory requirement globally, as evidenced by initiatives from Singapore [3], EU [27], Australia NSW [6], and Canada [40]. Broadly, all metrics within this paper can be leveraged to conduct ethics-based auditing [64], a process that rigorously assesses an entity's adherence to moral principles (i.e., Accountability).

Key Considerations:

- **Diversified Auditing Strategy**: Develop a comprehensive auditing strategy that incorporates both internal and external audits [34, 69, 75, 87]. Tailor this strategy to the AI system's context, considering its risk levels, complexity, stakeholder involvement, and regulatory requirements. Ensure the independence of audits to provide unbiased evaluations [52, 73, 78, 87].
- **Multi-Dimensional Audit Techniques** [34]: Employ various audit techniques, including technical (focused on data and code), empirical (centered on measuring inputs and outputs), and governance-oriented (evaluating procedures and decisions).
- **Ethical and Legal Alignment**: Adhere strictly to established standards and ethical/legal frameworks [27, 34, 70].
- **Regular Auditing**: Incorporate a schedule for regular auditing [27, 47, 73] that includes pre-deployment [27, 61], post-deployment [61, 70], and post-incident audits [59, 61].
- **Verifiable Audits**: Ensure that audit results are transparent and verifiable, using mechanisms like **licenses and certifications** [9, 78, 94]. This is particularly important for large-scale GenAI models in high-stakes domains.
- **Audit-Driven Improvements** [56]: Establish a structured process for incorporating audit findings into ongoing system improvements. Prioritize actions based on the findings' severity, impact, and feasibility of implementation.

5.3 Resource & Product Metrics – Auditability

Auditability ensures AI systems are compliant with ethical and legal standards. The resources and products under this criterion facilitate thorough and effective auditing, enabling organizations to demonstrate the integrity and reliability of their AI systems.

Essential ***Resources*** for *Auditability*:

- **Soft Laws**: Comprehensive audit frameworks, guidelines, and standards that provide a structured approach to auditing AI systems. These may include best practices for AI auditing and methodologies for conducting thorough evaluations. For example, the UK Information Commissioner's Office published a guideline for conducting AI audits [91]. The capAI framework [27] aligns with the EU AI Act, providing a template for conformity auditing of AI systems. Henrisken et al. [37] introduced a framework for end-to-end internal algorithmic auditing.
- **Hard Laws**: Legal and regulatory resources that provide up-to-date information on requirements relevant to AI systems. Examples include the proposed EU AI Act [18] and the US Blueprints for an AI Bill of Roghts [90].
- **AI Documentation Tools**: Employ various tools and frameworks for AI documentation, including data documentation (e.g., Datasheet [31], Data Readiness Report [1]), model documentation (e.g., Model Cards [63], Model Info Sheet [44]), and system-level documentation (e.g., Reward Reports [33], System Cards [36, 77], FactSheets [5], Software Bill of Materials [95], AI Bill of Materials [97]). These tools are crucial for ensuring comprehensive and transparent AI system documentation, aiding effective auditing.
- **Technical Tools**: Techniques such blockchain [50, 51] and knowledge graphs [65, 66] have also been explored for enabling provenance, auditing, and governance of AI systems.

Key ***Products*** indicating effective implementation of *Auditability* include documentations [3] and built-in system features:

- **Provenance Records**: Detailed documentation on the data/model/system etc., resulting from the utilized AI documentation tools.
- **Audit Reports**: Detailed reports and documentation generated from auditing activities, including audit findings, recommendations, and action plans for rectification.
- **Compliance Certificates and Licenses**: Certificates or licenses could be issued upon successful completion of audits, indicating compliance with specific standards or regulations. These serve as formal recognition of the AI system's adherence to established norms.
- **Integrated System Features**: Built-in system functionalities such as automated logging features and version control.

6 RESULTS - REDRESSABILITY

Redressability serves as the linchpin for actionable recourse. It necessitates the creation of formal mechanisms that facilitate stakeholder remediation for adverse impacts, not only fulfilling ethical and legal imperatives but also engender stakeholder trust.

6.1 Redress-by-Design

This refers to the preemptive incorporation and establishment of mechanisms for issue and error detection, management, and rectification within the AI system, which fortifies accountability by enabling timely and effective redress.

6.1.1 ***Process Metric 3.1.1: Incident Reporting and Response.*** **Context and Importance**: Effective Incident Reporting and Response mechanisms are critical for mitigating adverse impacts of (Gen)AI systems [87]. These processes not only rectify issues but also facilitate the continuous evolution of AI systems. Their implementation is vital for maintaining stakeholder trust and adhering to ethical and legal standards.

Key Considerations:

- **Accessibility and Visibility** [75]: Reporting channels must be accessible to all stakeholders, offering user-friendly options like online forms and hotlines. Prominent visibility of these channels encourages engagement and reporting.
- **Structured Incident Management** [73, 75]: Develop a structured process for handling reported incidents. This should include initial assessment, severity categorization, in-depth investigation, response planning, and execution of corrective actions, with provisions for progress tracking.
- **Feedback Loop Integration** [21, 70]: Systematically incorporate feedback from incident reports into the AI system's development and operational processes. This integration is crucial for ongoing system refinement, enhancing performance, and reducing future risks.

6.1.2 ***Process Metric 2.2.4: Built-in Redundancy.*** **Context and Importance**: Built-in redundancy is a fundamental aspect of designing (Gen)AI systems to ensure their resilience and facilitate redress. It involves creating multiple layers of fallback mechanisms and alternative procedures, which are critical for the system's fault tolerance [75]. This aspect of design is indispensable for continuous operation, effective problem detection, and the rectification of errors or system failures, safeguarding against potential disruptions and upholding the integrity of AI systems.

Key Considerations:

Multi-Modal Redundancy [75]: Implement redundancy across various dimensions of the (Gen)AI system, such as data storage, computational resources, and operational procedures. This ensures that the system remains functional and efficient, even in the face of component failures or external disruptions. Examples include OpenAI's global infrastructure redundancy plan [76].

6.2 Resource & Product Metrics – Redressability

Redressability underscores the importance of resources and products that facilitate prompt and effective responses to incidents, ensuring swift and effective response measures.

Essential ***Resources*** for *Redressability*:

- **Redundancy Design Case Studies**: Utilize examples from existing AI systems to inform redundancy strategies. For instance, Tesla's autopilot system [88], which employs dual AI chips for decision consensus, serves as a practical model for effective redundancy implementation in AI designs.
- **Incident Management Tools**: Utilize integrated tools designed for comprehensive reporting, tracking, and managing incidents related to AI systems. These tools should facilitate efficient communication among stakeholders and responsible teams, ensuring effective responses.

Key ***Products*** indicative of effective *Redressability*:

- **Incident and Response Documentation**: Detailed records of incidents, the responses undertaken, corrective action

plans, and outcomes. Unlike "Incident and Response Logging," which focuses on real-time logging, this documentation provides a post-event analysis, including a transparent record of actions, timelines, and responsible parties.
- **Integrated System Features**: This include built-in system functionalities that allow users to provide feedback and report incidents directly, and redundant components or functionalities as part of the system's design [53].

7 DISCUSSION

7.1 Accountability and Liability

In the context of (Gen)AI, distinguishing between liability and accountability is crucial yet complex, involving intricate legal and ethical dimensions. Liability pertains to the *legal responsibility* for damages caused by AI systems, often encompassing financial reparations and associated with *legal risks*. Accountability, on the other hand, relates to *social expectations* and *public answerability*, and is associated with both *legal* and *reputational risks*.

The integration of GenAI with external services and API-based interactions further complicates this landscape, introducing complex AI supply chains [17]. For instance, training subsequent models based on outputs from upstream models can obscure the origins of data and decisions, thereby complicating the assignment of accountability and liability. These scenarios exemplify how the scale and intricacy of GenAI systems present unique challenges that may not be fully addressed by existing legal frameworks. In response, initiatives like the European Commission's AI Liability Directive [23] aim to adapt liability rules for AI. This directive introduces concepts such as a rebuttable "presumption of causality" for victims and a strict liability regime for high-risk AI systems.

7.2 Implications

Practical Implications: Our study provides a comprehensive metrics catalogue of AI accountability, synthesizing insights from both academic and grey literature. It offers practitioners and policymakers in AI a nuanced understanding of AI accountability, critical for developing robust AI governance and risk management frameworks. The identified metrics for AI accountability are particularly valuable for organizations seeking to operationalize ethical AI principles, emphasizing not only the technical aspects but also the ethical, legal, and societal dimensions. Additionally, the inclusion of the GenAI perspective provides a timely insight into AI accountability in a rapidly evolving landscape. Furthermore, the practical application of these metrics in tools like web-based portals can significantly aid organizations in assessing and managing AI risks.

Theoretical Implications: Our research contributes to the evolving discourse on AI governance and risk management by offering a refined conceptual framework for AI accountability. This framework emphasizes the interconnectedness of responsibility, auditability, and redressability, laying a foundation for future academic exploration in GenAI accountability.

7.3 Limitations and Future Work

Limitations: While the metrics presented in this study form a comprehensive catalogue for AI accountability, they currently primarily focus on a binary (yes/no) assessment, which may not fully capture the complexities of AI systems. The varying levels of RAI maturity across industries, coupled with the evolving legal landscape surrounding AI, especially GenAI, pose challenges to the universal applicability of these metrics. Additionally, the implementation of these process metrics necessitates concerted efforts from various stakeholders, including policymakers, AI developers, end-users, as well as the general public.

Future Work: Future research includes empirically validating the proposed metrics through user testing, which is an ongoing effort by the authors. Adapting the catalogue to accommodate diverse contexts, including varying stakeholders, lifecycle stages, and regional or global legislative standards, is essential to enhance its utility across different domains. Future research could also expand the scope of this study to include other major AI ethics principles. Furthermore, the development of more sophisticated metrics and scoring mechanisms, beyond the current binary (yes/no) approach, could significantly enhance the precision and applicability of our findings. This advancement would allow for a more nuanced assessment of AI accountability, taking into account the quality and effectiveness of the implemented processes.

7.4 Threats to Validity

Internal Validity: The primary threat to internal validity is the potential for subjective interpretation during thematic coding. While measures were taken to mitigate this (e.g., multiple reviewers), some degree of subjectivity is inevitable. Additionally, while our methodology was rigorous, there is a possibility that relevant works might not have been included, which could affect the comprehensiveness of our findings. **External Validity**: The rapidly evolving nature of AI technology means that our findings may need updating as new developments emerge. **Construct Validity**: The construct validity hinges on the appropriateness and comprehensiveness of the AI accountability metrics. While developed through a rigorous methodology, empirical validation is necessary to establish their effectiveness. Moreover, the classification into process, resource, and product metrics presents its own challenges. The identification of resource and product metrics is less explicit compared to process metrics, as they are often implicitly mentioned in the literature, which could affect the clarity and precision of our categorization.

8 CONCLUSION

In this paper, we present a comprehensive metrics catalogue for AI accountability, with a specifically tailored focus of GenAI. Our approach synthesizes insights from both academic and grey literature, resulting in a robust framework that encapsulates the critical dimensions of *Responsibility*, *Auditability*, and *Redressability* in AI systems. This catalogue, comprising process, resource, and product metrics, serves as a practical tool for AI governance and risk assessment, particularly in the rapidly evolving domain of GenAI.

Our work contributes to the discourse on AI accountability, bridging the gap between theoretical concepts and practical applications. By integrating both technical and non-technical aspects, such as ethical, legal, and societal considerations, our findings offer a holistic view of accountable AI systems. Our work not only enriches the academic discourse but also provides actionable guidance for practitioners and fosters AI accountability.

REFERENCES

[1] Shazia Afzal, C Rajmohan, Manish Kesarwani, Sameep Mehta, and Hima Patel. 2021. Data readiness report. In *2021 IEEE International Conference on Smart Data Services (SMDS)*. IEEE, 42–51.

[2] Credo AI. 2023. Credo AI. https://www.credo.ai/.

[3] AI Verify Foundation. 2023. What is AI Verify? https://aiverifyfoundation.sg/what-is-ai-verify/

[4] B Akhgar, PS Bayerl, K Bailey, R Dennis, H Gibson, S Heyes, A Lyle, A Raven, and F Sampson. 2022. *Accountability Principles for Artificial Intelligence (AP4AI) in the Internal Security Domain: AP4AI Framework Blueprint.* Europol Innovation Lab.

[5] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[6] Australia NSW Government. 2022. NSW Artificial Intelligence Assurance Framework. https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework

[7] Australian Government, Department of Industry, Sicence, and Resources. [n. d.]. Australia's AI Ethics Principles. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

[8] Prachi Bagave, Marcus Westberg, Roel Dobbe, Marijn Janssen, and Aaron Yi Ding. 2022. Accountable AI for Healthcare IoT Systems. In *2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*. IEEE, 20–28.

[9] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. 2023. Managing AI Risks in an Era of Rapid Progress. *arXiv preprint arXiv:2310.17688* (2023).

[10] Reid Blackman. 2022. Why You Need an AI Ethics Committee. *Harvard Business Review* (2022). https://hbr.org/2022/07/why-you-need-an-ai-ethics-committee

[11] Veronika Bogina, Alan Hartman, Tsvi Kuflik, and Avital Shulner-Tal. 2021. Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *International Journal of Artificial Intelligence in Education* (2021), 1–26.

[12] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The Foundation Model Transparency Index. *arXiv preprint arXiv:2310.12941* (2023).

[13] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework 1. *European law journal* 13, 4 (2007), 447–468.

[14] Madalina Busuioc. 2021. Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review* 81, 5 (2021), 825–836.

[15] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 4209.

[16] Center for Long-Term Cybersecurity (CLTC), University of California, Berkeley. 2023. A Taxonomy of Trustworthiness for Artificial Intelligence. https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/.

[17] Jennifer Cobbe, Michael Veale, and Jatinder Singh. 2023. Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1186–1197.

[18] European Commission. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[19] A Feder Cooper, Karen Levy, and Christopher De Sa. 2021. Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.

[20] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 864–876.

[21] Google DeepMind. 2023. Google's Response to NTIA Request for Comment: Artificial Intelligence Accountability. https://www.regulations.gov/comment/NTIA-2023-0005-1308

[22] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[23] European Parliament. 2023. Artificial Intelligence Liability Directive. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf

[24] Norman Fenton and James Bieman. 2014. *Software metrics: a rigorous and practical approach*. CRC press.

[25] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92.

[26] Lajla Fetic, Torsten Fleischer, Paul Grünke, Thilo Hagendorf, Sebastian Hallensleben, Marc Hauer, Michael Herrmann, Rafaela Hillerbrand, Carla Hustedt, Christoph Hubig, et al. 2020. From Principles to Practice. An interdisciplinary framework to operationalise AI ethics. (2020).

[27] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. 2022. capAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. *Available at SSRN 4064091* (2022).

[28] Jade S Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P Bennett, Jamie McCusker, and Deborah L McGuinness. 2022. An Ontology for Fairness Metrics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 265–275.

[29] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. , 277–304 pages.

[30] Vahid Garousi, Michael Felderer, and Mika V. Mäntylä. 2019. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and Software Technology* 106 (Feb 2019), 101–121. https://doi.org/10.1016/j.infsof.2018.09.006

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

[32] German Federal Office for Information Security (BSI). 2021. AI Cloud Service Compliance Criteria Catalogue (AIC4). https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.html.

[33] Thomas Krendl Gilbert, Nathan Lambert, Sarah Dean, Tom Zick, Aaron Snoswell, and Soham Mehta. 2023. Reward reports for reinforcement learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 84–130.

[34] Ellen P Goodman and Julia Trehu. 2022. AI audit washing and accountability. *Available at SSRN 4227350* (2022).

[35] Government of Canada. 2023. Algorithmic Impact Assessment tool . https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[36] Furkan Gursoy and Ioannis A Kakadiaris. 2022. System cards for AI-based decision-making for public policy. *arXiv preprint arXiv:2203.04754* (2022).

[37] Anne Henriksen, Simon Enni, and Anja Bechmann. 2021. Situated accountability: Ethical principles, certification standards, and explanation methods in applied AI. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 574–585.

[38] U.K. Information Commissioner's Office (ICO). 2020. Guidance on the AI Auditing Framework: Draft Guidance for Consultation. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

[39] IEEE Standards Association. [n. d.]. IEEE portfolio of AIS technology and impact standards and standards projects. https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/.

[40] Innovation, Science and Economic Development Canada. 2023. Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems. https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems.

[41] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. 2022. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences* 12, 3 (2022), 1353.

[42] Seyyed Ahmad Javadi, Richard Cloete, Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh. 2020. Monitoring misuse for accountable'artificial intelligence as a service'. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 300–306.

[43] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[44] Sayash Kapoor and Arvind Narayanan. 2022. Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048* (2022).

[45] Mehtab Khan and Alex Hanna. 2022. The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability. (2022).

[46] Barbara Kitchenham, O Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology* 51, 1 (2009), 7–15.

[47] Florian Königstorfer and Stefan Thalmann. 2022. AI Documentation: A path to accountability. *Journal of Responsible Technology* 11 (2022), 100043.

[48] Thomson Reuters Practical Law. [n. d.]. Accountability principle. https://uk.practicallaw.thomsonreuters.com/w-014-8164?transitionType=

Default&contextData=(sc.Default).
[49] Sung Une Lee, Harsha Perera, Boming Xia, Yue Liu, Qinghua Lu, Liming Zhu, Olivier Salvado, and Jon Whittle. 2023. QB4AIRA: A Question Bank for AI Risk Assessment. *arXiv preprint arXiv:2305.09300* (2023).
[50] Yue Liu, Qinghua Lu, Liming Zhu, and Hye-Young Paik. 2023. Decentralised Governance for Foundation Model based Systems: Exploring the Role of Blockchain in Responsible AI. *arXiv preprint arXiv:2308.05962* (2023).
[51] Sin Kit Lo, Yue Liu, Qinghua Lu, Chen Wang, Xiwei Xu, Hye-Young Paik, and Liming Zhu. 2022. Toward trustworthy ai: Blockchain-based architecture design for accountability and fairness of federated learning systems. *IEEE Internet of Things Journal* 10, 4 (2022), 3276–3284.
[52] Michele Loi and Matthias Spielkamp. 2021. Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 757–766.
[53] Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and Aurelie Jacquet. 2023. Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. *ACM Comput. Surv.* (2023). https://doi.org/10.1145/3626234
[54] Laura Lucaj, Patrick van der Smagt, and Djalel Benbouzid. 2023. AI Regulation Is (not) All You Need. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1267–1279.
[55] Milan Markovic, Iman Naja, Pete Edwards, and Wei Pang. 2021. The accountability fabric: A suite of semantic tools for managing ai system accountability and audit. In *CEUR Workshop Proceedings*.
[56] Jeanna Matthews. 2020. Patterns and anti-patterns, principles and pitfalls: accountability and transparency in AI. *AI Magazine* 41, 1 (2020), 82–89.
[57] Bahar Memarian and Tenzin Doleck. 2023. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review. *Computers and Education: Artificial Intelligence* (2023), 100152.
[58] Microsoft. 2022. Microsoft Responsible AI Impact Assessment Template. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf
[59] Microsoft. 2022. Microsoft Responsible AI Standard v2: General Requirements. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf
[60] Microsoft. 2023. Log and Monitor Azure OpenAI. https://learn.microsoft.com/en-us/azure/architecture/ai-ml/openai/architecture/log-monitor-azure-openai.
[61] Microsoft. 2023. Microsoft response to AI Accountability Policy Request for Comment, NTIA-2023–07776. https://www.regulations.gov/comment/NTIA-2023-0005-1337
[62] Microsoft. n.d.. Responsible AI Principles and Approach. https://www.microsoft.com/en-us/ai/principles-and-approach
[63] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
[64] Jakob Mökander and Luciano Floridi. 2023. Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics* 3, 2 (2023), 451–468.
[65] Iman Naja, Milan Markovic, Peter Edwards, and Caitlin Cottrill. 2021. A semantic framework to support AI system accountability and audit. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*. Springer, 160–176.
[66] Iman Naja, Milan Markovic, Peter Edwards, Wei Pang, Caitlin Cottrill, and Rebecca Williams. 2022. Using Knowledge Graphs to Unlock Practical Collection, Integration, and Audit of AI Accountability Information. *IEEE Access* 10 (2022), 74383–74411.
[67] Mina Narayanan and Christian Schoeberl. 2023. A Matrix for Selecting Responsible AI Frameworks. https://cset.georgetown.edu/publication/a-matrix-for-selecting-responsible-ai-frameworks/
[68] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2 (1996), 25–42.
[69] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2023. Accountability in artificial intelligence: what it is and how it works. *AI & SOCIETY* (2023), 1–12.
[70] OECD. 2023. Advancing accountability in AI. 349 (2023). https://doi.org/https://doi.org/10.1787/2448f04b-en
[71] OECD. n.d.. OECD AI Principles overview. https://oecd.ai/en/ai-principles
[72] OECD.AI. 2023. Catalogue of Tools & Metrics for Trustworthy AI. https://oecd.ai/en/catalogue/metrics.
[73] U.S Department of Energy. [n. d.]. DOE AI Risk Management Playbook (AIRMP). https://www.energy.gov/ai/doe-ai-risk-management-playbook-airmp
[74] US National Institute of Standards and Technology (NIST). 2023. AI Risk Management Framework (AI RMF 1.0). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
[75] European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG). 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
[76] OpenAI. [n. d.]. Production Best Practices. https://platform.openai.com/docs/guides/production-best-practices/improving-latencies.
[77] OpenAI. 2023. *GPT-4 System Card*. Technical Report. OpenAI. https://cdn.openai.com/papers/gpt-4-system-card.pdf
[78] OpenAI. 2023. OpenAI's Response to NTIA Request for Comment: Artificial Intelligence Accountability. https://www.regulations.gov/comment/NTIA-2023-0005-1245
[79] José A Peregrina, Guadalupe Ortiz, and Christian Zirpins. 2022. Towards a metadata management system for provenance, reproducibility and accountability in federated machine learning. In *European Conference on Service-Oriented and Cloud Computing*. Springer, 5–18.
[80] Hong Kong Privacy Commissioner for Personal Data. 2021. Guidance on the Ethical Development and Use of Artificial Intelligence. https://www.pcpd.org.hk/english/resources_centre/publications/files/guidance_ethical_e.pdf
[81] Ashwin Kumar Raja and Jianlong Zhou. 2023. AI Accountability: Approaches, Affecting Factors, and Challenges. *Computer* 56, 4 (2023), 61–70.
[82] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.
[83] John Ratzan and Noushi Rahman. 2023. Measuring responsible artificial intelligence (RAI) in banking: a valid and reliable instrument. *AI and Ethics* (2023), 1–19.
[84] Amanda Sinclair. 1995. The chameleon of accountability: Forms and discourses. *Accounting, organizations and Society* 20, 2-3 (1995), 219–237.
[85] Helen Smith. 2021. Clinical AI: opacity, accountability, responsibility and liability. *Ai & Society* 36, 2 (2021), 535–545.
[86] Jaemarie Solyst, Shixian Xie, Ellia Yang, Angela EB Stewart, Motahhare Eslami, Jessica Hammer, and Amy Ogan. 2023. "I Would Like to Design": Black Girls Analyzing and Ideating Fair and Accountable AI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.
[87] Ramya Srinivasan and Beatriz San Miguel González. 2022. The role of empathy for artificial intelligence accountability. *Journal of Responsible Technology* 9 (2022), 100021.
[88] Tesla. [n. d.]. Autopilot. https://www.tesla.com/autopilot.
[89] The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
[90] The White House Office of Science and Technology Policy. 2023. Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.
[91] UK Information Commissioner's Office (ICO). [n. d.]. A Guide to AI Audits. https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf.
[92] US National Institute of Standards and Technology (NIST). [n. d.]. AI Measurement and Evaluation. https://www.nist.gov/ai-measurement-and-evaluation.
[93] Mihaela Vorvoreanu, Amy Heger, Samir Passi, Shipi Dhanorkar, Zoe Kahn, and Ruotong Wang. 2023. *Responsible AI Maturity Model*. Technical Report MSR-TR-2023-26. Microsoft Research.
[94] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, explainable, and accountable AI for robotics. *Science robotics* 2, 6 (2017), eaan6080.
[95] Boming Xia, Tingting Bi, Zhenchang Xing, Qinghua Lu, and Liming Zhu. 2023. An Empirical Study on Software Bill of Materials: Where We Stand and the Road Ahead. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. 2630–2642. https://doi.org/10.1109/ICSE48619.2023.00219
[96] Boming Xia, Qinghua Lu, Harsha Perera, Liming Zhu, Zhenchang Xing, Yue Liu, and Jon Whittle. 2023. Towards Concrete and Connected AI Risk Assessment (C^2AIRA): A Systematic Mapping Study. In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*. IEEE, 104–116.
[97] Boming Xia, Dawen Zhang, Yue Liu, Qinghua Lu, Zhenchang Xing, and Liming Zhu. 2023. Trust in Software Supply Chains: Blockchain-Enabled SBOM and the AIBOM Future. *arXiv preprint arXiv:2307.02088* (2023).
[98] Agne Zainyte and Wei Pang. 2021. Challenges and Future Directions for Accountable Machine Learning. In *CEUR Workshop Proceedings*, Vol. 2894. CEUR-WS, 40–47.
[99] Dawen Zhang, Boming Xia, Yue Liu, Xiwei Xu, Thong Hoang, Zhenchang Xing, Mark Staples, Qinghua Lu, and Liming Zhu. 2023. Tag Your Fish in the Broken Net: A Responsible Web Framework for Protecting Online Privacy and Copyright. *arXiv preprint arXiv:2310.07915* (2023).