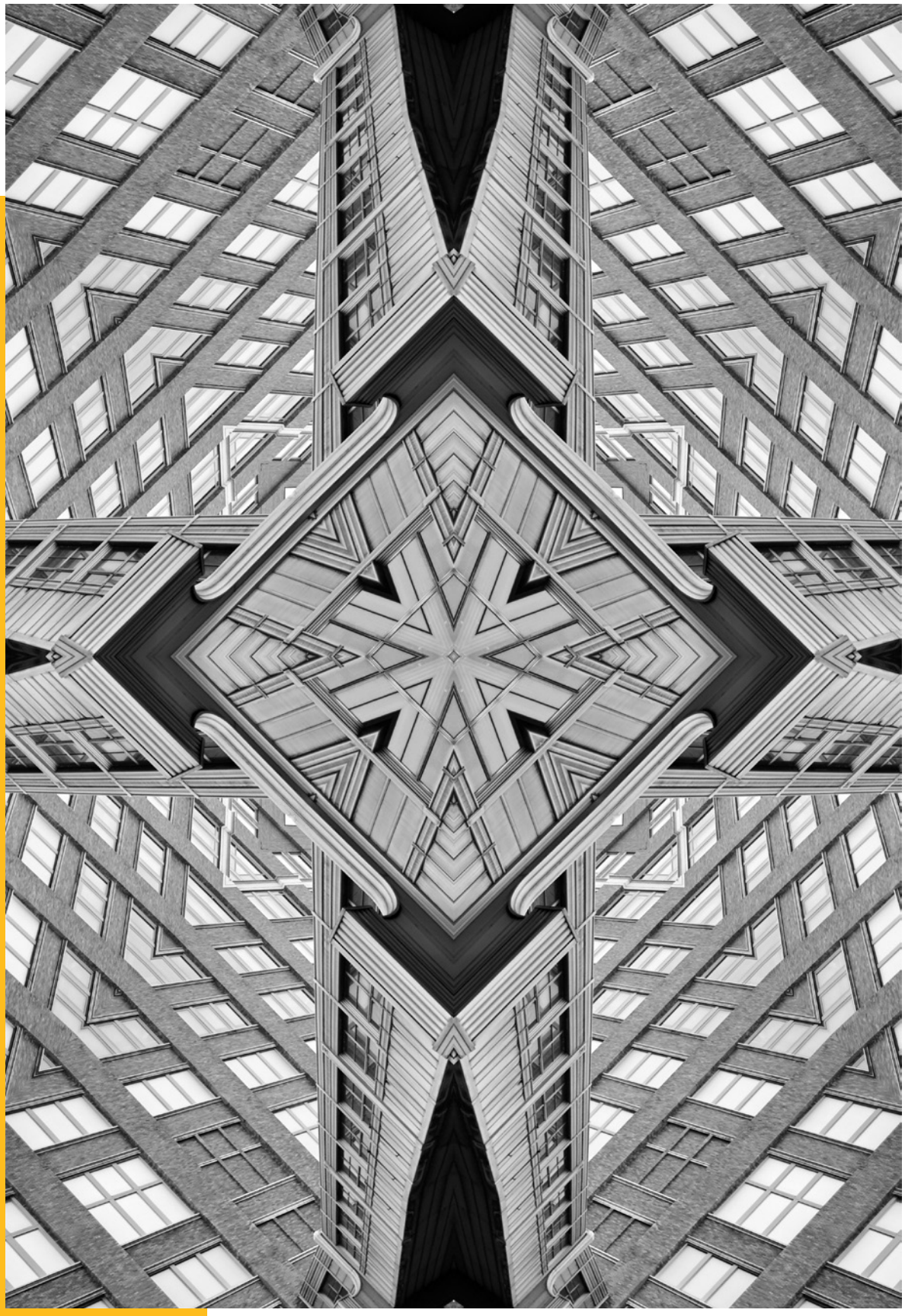
OBSERVER RESEARCH FOUNDATION

Occasional Paper



ISSUE NO. 460 JANUARY 2025

Audits as Instruments of Principled AI Governance

Anulekha Nandi

Abstract

Stakeholder groups have produced various guidelines on ethical Artificial Intelligence (AI) in recent years. However, translating principles into practice continues to be a massive challenge, as AI markets expand and AI risks are heightened. AI audits—or the process of investigating an algorithm against existing regulations and known harms—are emerging as a way of bridging the gap between principle and practice. This paper scans the landscape of AI audits and highlights the roles of industry organisations and technical bodies, governments, civil society organisations, academics, and researchers as well as the necessity of developing procedural standardisation, the skillsets required for audit teams, and determining the appropriate nature of regulation and compliance mechanisms.

Attribution: Anulekha Nandi, "Audits as Instruments of Principled AI Governance," *Occasional Paper No. 460*, January 2025, Observer Research Foundation.

Introduction

Artificial Intelligence (AI) governance is witnessing an "ethics boom".[1] A 2019 paper listed about 84 policy documents issued by institutional entities in both the public and private sectors that sought to define the values, tenets, and other guiding principles for ethical AI development and deployment.[2] In 2020 alone, 23 new sets of principles were created, a majority of them by private companies.[3] A 2023 study identified and analysed 200 ethical guidelines and AI governance policies released by public and academic institutions, private companies, and civil society organisations.[4] The period 2015-2020 saw 117 AI principles, with research and professional organisations being the first movers and private companies contributing the largest in terms of volume.[5] The heightened interest in ethical principles for AI governance is in parallel with the growth of the AI market, whose value increased by US$50 billion from 2023 to exceed US$184 billion in 2024.[6]

The proliferation of AI systems and services across sectors has added to the difficulty of operationalising high-level ethical principles, given the risks of harms arising in AI use, from concerns around bias and discrimination to social manipulation and misuse by malicious actors. The overarching challenge has been translating high-level principles to low-level technical and organisational measures.[7] This challenge arises largely from the focus of current debates on the 'what' of AI ethics (i.e., the broad principles that have emerged in the form of transparency, accountability, and non-discrimination) rather than the 'how' (i.e., practices or practical techniques required for the responsible management of AI).[8] This focus has resulted in a lack of alignment between the principles and operational requirements. Organisations that are developing and implementing AI systems require specific capabilities to be able to detect, identify, and remedy instances when AI systems in practice deviate from principles.[9]

The gulf between principles and practice widens due the lack of enforceability of principles, standardisation, and compliance mechanisms, alongside concerns around AI risks and harms across sectors and processes where AI is integrated, from hiring, security, and criminal justice, to housing, finance, and healthcare.[10]

AI audits[a] are increasingly gaining traction as one of the methods to bridge this gap by contributing to the overarching principles of transparency, safety, accountability, and non-discrimination that are highlighted in ethical

a This paper uses the term 'AI audit' although extant literature also uses the term 'algorithmic audit'. This is because the latter can indicate that auditing is restricted to the querying and evaluation of algorithms alone, while existing approaches and auditing mechanisms can also include documentation strategies, standards of practice, and risk assessments. The term 'AI audits' can potentially operate as an umbrella term to encapsulate different mechanisms that are used to audit AI systems, particularly given the need to understand how existing approaches need to be combined to provide a holistic approach towards the establishment of AI audits as a system of practice.

AI principles.[11] AI audits bridge the gap between principles and practice by providing the evaluative component, i.e., highlighting how organisations that are developing and implementing AI systems are performing in terms of high-level principles. This evaluative function of AI audits supports AI governance by assessing the extent to which AI principles are observable in practice. This has led to the emergence of risk assessment mechanisms in regulations such as the European Union's (EU) AI Act,[12] New York City's bias audit law on Automated Employment Decision Tools (AEDT),[13] and frameworks like National Institute of Standards and Technology's (NIST) AI Risk Management Framework (RMF).[14]

However, AI audits have not developed into a coherent system of practice because of the diverse range of existing AI technologies; an AI system can consist of machine learning (ML), computer vision, or natural language processing combined with other software or mathematical and statistical approaches that enable a given functionality. One such approach involves systematically querying an algorithm with a range of inputs and statistically comparing the results. This approach was used by Harvard professor Latanya Sweeney, who found that Google ads were 25 percent more likely to suggest arrest records for names that "sounded" Black rather than for names that sounded White.[15]

Another seminal audit was performed by researchers at the Massachusetts Institute of Technology (MIT) and highlighted the biases in facial-recognition algorithms.[16] They used facial analysis benchmarks of gender and skin type to evaluate three commercial gender-classification systems. The results indicated that darker-skinned females were the most misclassified group, with error rates of up to 34.7 percent, compared to lighter-skinned males, with error rates of 0.8 percent.[17]

Although there are different approaches, most AI audits involve probing a product or process (e.g., facial-recognition systems or hiring processes) by providing the AI model with one or more inputs while changing the attributes of such inputs (e.g., gender and/or race). The aim is to assess whether a given AI system falls short of expected criteria such as bias, fairness, transparency, and regulatory compliance.[18]

However, despite the lack of systematisation within AI audits, different jurisdictions have been adopting different approaches for evaluating the safety of AI systems in the form of risk and impact assessments or mandating other forms of evaluation or transparency measures.[19] The EU AI Act, 2024 lays down a fourfold risk-based classification of AI systems, with high-risk systems having to undergo risk assessment before being put on the market as well as throughout their life cycle.[20] Generative AI models like ChatGPT, though not classified

as high risk, would be subject to transparency requirements. In the United States (US), NIST developed AI RMF following directions from the National Artificial Intelligence Initiative Act, 2020 to help "organizations designing, developing, deploying, or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems."[21] In August 2024, OpenAI and Anthropic signed an agreement with the US AI Safety Institute to obtain access to AI models from both companies prior to and following their release to enable feedback mechanisms on the basis of capability evaluation and safety risk assessment.[22] Earlier in 2024, the Indian government issued a revised advisory that required businesses to appropriately label AI-generated content to indicate their potential unreliability and contained provisions for firms to label and embed metadata and identifiers to enable traceability in the event of misuse.[23]

However, despite these developments and efforts to regulate the technology, the AI audit landscape remains fragmented, with advocates and practitioners highlighting the need for standardised frameworks to drive trust and legitimacy in the process.[24] This need is driven by a lack of consensus around terminologies as well as the lack of procedural and methodological standards, further compounded by the opaque and self-learning nature of the technology itself. A particular AI system performing a given function can be composed of a number of models and training data that is continuously self-learning from its interaction with humans. Therefore, there is a need to audit an AI system as a composition of technical sub-systems in conjunction with social elements stemming from human involvement in its design, development, training, and deployment. While data, model, and output render itself to bias evaluation and mitigation strategies, the source of bias from human feedback can be difficult to determine. This leads to an "interaction failure", wherein the constraints of the technology in its interface with social structures results in social harms or unjust outcomes for end users.[25]

Moreover, AI audits and evaluations are marked by diverse approaches that have not been procedurally standardised in relation to the nature of technology and its field of application. There is a lack of consensus on what constitutes AI audits, with the emergence of different communities of practice that have different priorities and viewpoints.[26]

This paper aims to capture some of these approaches and identify the ways forward from this fragmented landscape to identify the roles played by different stakeholders. Evaluations and investigations in various cases of algorithmic discrimination reinforce the need for ensuring the safety of these systems. However, a number of associated elements and concepts need to be factored in when discussing AI audits, including transparency, accountability,

Introduction

bias, fairness, and responsibility. This paper builds on the premise of audits as mechanisms to enable transparency and accountability to identify a range of audit approaches—from bias measurement and mitigation to risk assessment. Identifying these diverse approaches can help determine the roles and responsibilities of stakeholder groups in determining a standard of practice. This paper does not conduct in-depth expositions of these elements, as this has already been covered in extant grey and academic literature. Instead, it aims to provide an overview of AI audits, highlight tensions between technical constraints and legal and regulatory developments, and discuss efforts by the wider community in developing approaches to enable the better management and governance of AI systems.

The paper is restricted to AI audits in civilian AI technologies, as military applications may have different regulatory and safety criteria.[27] AI as a technology also differs from other kinds of critical technologies like nuclear technologies, which have their unique and clearly defined audit and safety mechanisms. AI risks evolve dynamically due to multiple sources of bias, from training data to human interactions as well as the nature of technology and the context of the application.

The following sections outline existing auditing practices to understand existing gaps and the social and technical sources of bias to highlight the complexities in designing any form of auditing mechanism. This is achieved by juxtaposing these complexities against existing auditing mechanisms and regulatory approaches. The paper also provides recommendations to consolidate AI auditing as a practice to both ease and enable compliance. The paper concludes by distilling the roles that different stakeholders have to play to establish a standard of practice and identify the way forward.

The Basics of AI Auditing

AI audits can broadly be classified into three types: first-party, second-party, and third-party.[28] These are conducted by internal teams, contracted parties, or external entities, respectively.

First-party audits have become increasingly common in Big Tech firms, which have internal responsible-AI teams. First-party audits benefit from internal and continuous access to technologies, wherein risks could be dynamically identified and addressed. In contrast, much like independent financial audits, second-party audits are performed by external entities that are contracted by companies to conduct audits of their AI systems.

However, the two known examples of second-party audits have been criticised for their lack of transparency and misrepresentation. A second-party audit performed for Pymetrics to test the performance of a hiring algorithm was criticised for not adequately disclosing the contractual relationship between the auditors.[b,29] Similarly, an audit performed by O'Neil Risk Consulting & Algorithmic Auditing (ORCAA) for HireVue was criticised for misrepresenting results while placing the final audit report under a non-disclosure agreement. HireVue's statement contained a partial quote and misleading phrasing to suggest that the audit had concluded its products to be free from bias.[30]

The audit for HireVue focused on its hiring assessments that are used to evaluate fresh college graduates but adopted a different approach to the Pymetrics audit, which had a greater focus on stakeholder interviews than the technical design of the algorithm. It also had a narrower use case, as opposed to the case of HireVue, which included the entire suite of algorithms.[31] These examples indicate that a lack of procedural standards can hamper the demand, uptake, and delivery of audit services.[32] They also highlight the lack of a standardised approach and clarity on disclosure requirements. Moreover, voluntary ad-hoc audits are inadequate without broader regulation to support and mandate them.

---

b Pymetrics is hired by several large US firms like McDonald's, Boston Consulting Group, Heinz, and Colgate-Palmolive and uses AI-powered games designed as cognitive science experiments to score, screen, and shortlist job applicants.

The company contracted two academics and their team from Northeastern University through a grant of US$104,465, including US$64,813 for team salaries. Even though the researchers had editorial freedom, they had to first run negative findings by the company. See: https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/

Third-party audits conducted by independent organisations have flagged the biased operations and outputs of many AI implementations. These audits were conducted predominantly by rights and advocacy groups, journalists, or academics. ProPublica, Associated Press, and the MarkUp regularly conduct third-party audits, as well as the American Civil Liberties Union (ACLU) and institutions like NIST.[c,33] However, while third-party audits have highlighted significant biases and discrimination, the different documentations of audit methodologies are difficult to operationalise, standardise, and legitimise in practice, particularly due to the variety and variability in input data sources as well as the modelling approaches and AI techniques being used.

In addition to a lack of standards and standardised mechanisms, the legal regime around audits and auditors have a bearing on the future of AI audits. AI audits and transparency and accountability requirements can be conflicting and can compound considerations around intellectual property rights. This is because external auditors will have to gain access to granular details about the development process to be able to perform competent audits and ensure that relevant data is not blocked by non-disclosure agreements (NDAs), as this could hinder transparency requirements.[34] The issue of legal liability is a cause for concern, i.e., in financial audits, independent auditors are subject to legal liability to third parties and regulators for failure to identify omissions, misstatements, or knowingly abetting fraud. Unless standards, guidelines, and conditions of liability for auditors are clear, there might be uncertainty that hampers the development of the practice.[35] However, in addition to the organisational, institutional, and legal considerations, social and technical sources of bias are also key factors in formalising AI audits as a system of practice.

c ProPublica identified how the recidivism risk scoring system sold by Northpointe was biased against Black defendants. Joy Buolamwini of the Algorithmic Justice League identified the misclassification of people of colour in facial recognition systems. See: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm and https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

The AI development life cycle, ranging from problem definition to data selection, model development, refinement, and deployment, can be viewed as a sequential process of tasks and decisions.[36] In the initial stages of model development, the developer might not have the exact specifications for the product being developed due to an incomplete understanding of contextual or market needs. Given the lack of *a priori* specifications, model development requires arriving at an initial working version of the model, which then undergoes iterative development, testing, and evaluation.[37] This involves the ongoing selection and application of data, modelling approaches, and software packages along with operational domain knowledge and assumptions to arrive at a dependable functionality.[38] Development and fine-tuning of algorithms use a process of iterative problem-solving and abstraction of social reality to formulate precise mathematical relationships with well-defined parameters of causation to understand how well a set of variables can explain the phenomenon under consideration.[39]

Biases in AI systems can seep in through data, models, or human interaction.[40] Biases in data will transfer to the models as the model learns from them and returns outputs. The scale, speed, and sophistication with which AI systems operate can lead to compounding biases. Models trained on unrepresentative data can lead to the repeated production of unfair outcomes, errors, and the amplification of biases.

The modelling process itself can reinforce such biases and can be attributed to design and data choices.[41] These can stem from weighting factors assigned to different parameters within the model; for example, indicators like income or vocabulary might disadvantage historically underrepresented groups.[42] This pertains to how the problem is defined and the objectives of the model outlined. Moreover, AI systems are composed of multiple models and software that not only share technical interdependencies but learn from human interaction. An example is the 2016 case of Microsoft's AI chatbot Tay learning abusive and antisemitic behaviour from user interactions.[43] Despite technological developments to account for such behaviour, these outstanding concerns continue to persist as models develop 'overconfidence' (assigning high confidence or probability scores to incorrect predictions),[44] hallucinations (incorrect, misleading, or nonsensical predictions), and even sycophancy (models aligning themselves to users' views at the expense of accuracy) during the training phase, even with human-in-the-loop, as in the case of Reinforcement Learning through Human Feedback (RLHF). Models also remain susceptible to potential bias in the nature of the feedback itself, which highlights the tendency to seek approval from the humans training it or providing it feedback.[45]

Given the technical and social sources of bias and harm within AI systems, AI governance becomes both a technical and a social endeavour that requires an alignment between principle and practice.[46] In other words, it requires convergence between the established standards of desired behaviour, the means of their evaluation, and the assessment of their divergence. This becomes particularly difficult because of the diversity of AI approaches, varying contexts of their applications, and the self-learning and opaque nature of AI algorithms that defy codified evaluation and testing across technologies and contexts.

One of the approaches to identify and mitigate biases and harm within AI systems is to design "explainable AI"—whereby end users can understand and/or trace how an algorithm arrived at a particular decision.[47] One of the ways of approaching explainable AI is to design simple ML models in which it is easy for users to interpret and trace the causes behind outcomes. For more complex models, post-hoc explanations are arrived at through bias measurement and mitigation approaches. This can be model specific (i.e., for a neural network) or model agnostic (i.e., applicable to any model), with global explanations attempting to explain model behaviour as a whole while local predictions attempt to explain a single prediction.[48]

Besides concerns around bias mitigation and measurement requirements and approaches, many researchers have questioned whether AI systems even pass the functionality test, i.e., whether they actually work for the context for which they are designed and the problem that they are expected to solve. Some researchers argue that, even within policy documents and critiques of AI, the underlying assumption is that AI systems will fulfil their required functionality. There is limited acknowledgement of AI not working as intended, while balancing accuracy and fairness.[49] For example, when an algorithm to predict health support evaluates by health costs as opposed to illness, it discriminates on the basis of race and socio-economic status, as these disadvantaged groups are underrepresented due to their inability to access healthcare on account of high costs. As a result, such an algorithm can turn away severely ill patients from historically marginalised communities.[50]

Therefore, evaluations of bias are among the primary components of AI audits. This can range from mathematical and statistical approaches to evaluate and mitigate bias as well as documentation, standards of best practice, and regulatory mandates of risk assessment.

Approaches, Standards, and Regulatory Requirements

AI audits do not comprise a coherent set of practices. Auditing mechanisms span technical and mathematical modes of evaluation and mitigation, documentation techniques, standards, and regulatory requirements.

Bias evaluation and mitigation techniques represent abilities to identify and remedy deviations, while documentation approaches provide a trail for transparency and accountability. Standards provide guidance and norms against which to evaluate existing practices, with emerging regulatory approaches providing the support to assess and enforce norms and principles.

Bias Evaluation and Mitigation

AI audits can involve repeatedly querying the algorithm to draw conclusions about its workings.[51] AI audits range from bias evaluation and mitigation, to the mapping and documentation of the modelling process, as well as mandates for risks assessments stemming from risk-based regulatory approaches adopted in different jurisdictions. Big Tech firms have released bias and fairness evaluation toolkits—such as IBM Fairness 360, Microsoft Fairlearn, Amazon SageMaker, and Meta's ROBBIE[d]—most of which include tools for bias evaluation and bias mitigation.

Bias evaluation metrics involve statistical or mathematical approaches to evaluate whether a model or output treats privileged and underprivileged individuals and groups similarly.[52] Bias mitigation techniques offer a range of computation and statistical techniques to mitigate bias within an AI system. Bias mitigation can be deployed in the pre-processing (removing bias from datasets before using them as input for ML models), in-processing (modifying or manipulating algorithms to improve fairness while training), and post-processing stages (applied to scenarios of limited access to training data or the model; comparatively less popular than the other two), depending on the life-cycle stage in which mitigation strategies come into play.[53]

d IBM Fairness 360 is an open-source toolkit to "examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle." The toolkit provides bias evaluation metrics and bias mitigation algorithms (https://aif360.res.ibm.com/). Fairlearn was an open-source project started by Microsoft Research, New York. It now includes other partners, from Eindhoven University to Hugging Face (https://fairlearn.org/v0.10/about/index.html). It also contains toolkits to assess and mitigate biases. Like Fairlearn and Fairness 360, Amazon SageMaker Clarify contains assessment and detection algorithms while allowing evaluations through its Bedrock platform that provides access to foundational models from Amazon and other leading start-ups. Meta developed ROBBIE (Robust Bias Evaluation of Large Generative Language Models), a benchmarking and mitigating approach using prompt-based datasets across text domains and demographics (https://ai.meta.com/research/publications/robbie-robust-bias-evaluation-of-large-generative-language-models/). Meta further released two datasets that included a comprehensive set of terms across 12 different axes beyond race, gender, and ethnicity with the aim of ensuring fair model development procedure (https://ai.meta.com/blog/measure-fairness-and-mitigate-ai-bias/).

Documentation

There are additional approaches, besides computational and statistical approaches for bias detection and mitigation, for the documentation and mapping of the modelling process to facilitate transparency. These include datasheets for datasets, model cards, model development factsheets, and system cards. Datasheets for datasets can address the lack of a standardised approach for the documentation of datasets within ML.[54] This approach proposes that every dataset used in ML is accompanied by a datasheet that contains the characteristics of the dataset, thus providing an indication of the data provenance, with information on motivation, composition, collection process, preprocessing/cleaning/labelling, uses, distribution (how and to whom the datasets will be distributed), and maintenance (entities responsible for managing or maintaining the database, updating, and fixing errors, with an option for external contribution).

Different iterations of documentation processes, such as datasheets have been adopted by product teams in companies like Google, Microsoft, and IBM. Google also worked on model cards[e] aimed at documenting modelling processes as well as a data card that represented a more lightweight version of the datasheet-building process.[55]

IBM further suggested factsheets to document AI-service characteristics in tandem with the rising need for better documentation of AI products and services. Within this approach, IBM suggested and implemented a methodology for developing a documentation process called Factsheets, which were then deployed for a dozen models.[56] The methodology involved incorporating the "facts and lineage from all phases of the 'life of the model'". The factsheet methodology outlined a seven-step process within which a factsheet team responsible for anchoring the process coordinates with factsheet producers (such as a model developer) and consumers (such as a model validator) to understand the kinds of information generated and their requirements to produce and populate a factsheet template based on these the inputs. Producers are then aided to fill a factsheet of their own, which is then evaluated with consumers. Based on this internal process, a refined factsheet can be created for dissemination among a wider set of consumers such as external review boards, regulators, sales personnel, and end users.

e Model cards are aimed at conveying the characteristics of model performance to promote transparency about the nature of technology and its potential use. Model cards contain important information about a model's performance metrics that are benchmarked across parameters such as race, age, gender, and geographic location, which are relevant to the area of application. This needs to be adapted to the context of application and type of technology (e.g., ML, computer vision, large language models). See: https://arxiv.org/abs/1810.03993

Meta's system cards aimed to define how and why AI systems operate the way they do.[57] System cards build on earlier documentation methods to explain how ML models and non-AI technologies work together within a given AI system to achieve specific tasks. System cards aim to provide stakeholders with an overview of an AI system, how its different components operate, and how they interact, along with how information is used and consumed within the system.

In India, a seven-layer approach combining the evaluation and documentation approach informed the development of the standard for Fairness Assessment and Rating of Artificial Intelligence Systems (TEC 57050:2023) released by the Telecommunications Engineering Centre (TEC), Department of Telecommunication. This approach attempted to standardise AI fairness assessment across the AI development life cycle[58] and includes checklists that cover the following:

- The requirements, context, and purpose layer to understand the fit between the context of the problem, the technological solution, and the mode of solving it to understand stakeholders' conceptions of the problems, availability of existing technology, and whether tolerance limits for fairness and bias have been decided.

- Data collection and selection layer to establish the provenance of data and data sources, representativeness, age of the dataset, adequate labelling procedures, and quality assurance and verification procedures.

- Pre-processing and feature-engineering layer to detect and deploy bias-mitigation techniques by identifying protected attributes and their relationship and relevance to the area of application, including transparency around how features were selected, who was involved in the selection, and whether this process of inclusion/exclusion disadvantages any particular population group.

- Algorithm layer, which involves transparency in the selection of given algorithms for given contexts, the composition of the team, checks and reviews with regard to the creeping of individual biases, selection of the fairness evaluation, and congruency of the model in relation to the requirements specified in the first layer.

- AI-system-training layer to test the fairness of both the training dataset and the output of the model. This involves the deployment of appropriate fairness metrics for bias evaluation, such as statistical parity difference, equal (mis)opportunity, disparate impact, and equal opportunity.

- Independent audit layer to evaluate the independence of the auditing process and the following of standardised processes. This involves a system of scoring fairness of the overall system rather than individual models within the system.

- Usage layer to mitigate biases by monitoring fairness key performance indicators, re-training the system when required, and mitigating bias from new data as appropriate.

Risk Management

AI risk assessment and management are becoming increasingly important concepts within AI governance.[59] These are guided by AI ethical principles documents such as that by the Organisation for Economic Co-operation and Development (OECD), which help guide the priorities and direction of risk assessment. The risk-based approach has become a key aspect of AI regulation and governance, finding traction with lawmakers globally.[60] However, risk regulation works best on quantifiable problems and harms,[61] and AI risks have both technical and social sources, some of which are quantifiable, whereas others defy accurate detection such as through bias resulting from usage.

The EU AI Act mandates that conformity assessments are required to institute accountability for the development and deployment of AI systems that are classified as high-risk under the Act. The Act classifies AI systems as low or minimal risk, limited risk, high risk, or unacceptable risk, with proportional obligations. Unacceptable-risk AI systems are prohibited, while high-risk systems are subject to stringent obligations. Conformity assessments are meant to demonstrate whether a company or developer of an AI system is meeting the requirements set out in Title III Chapter 2 of the Act.[62] These assessments must be conducted before developers release these AI systems on the market. The requirement for conformity assessments highlights the need for internal audits or notified assessment bodies that are responsible for performing such evaluations. However, these requirements depend on the presence and usage of harmonised standards.[63]

The European Data Protection Board (EPDB) initiated the AI auditing project within the framework of the Support Pool of Experts programme, at the initiative of the Spanish Data Protection Authority, to help all parties understand and assess data protection safeguards within the context of the AI Act.[64] However, the EU AI Act does not provide for access for third-party audits. Researchers argue that neither the Digital Services Act nor the AI Act support the establishment of a third-party AI audit ecosystem by providing for access to data and models for civil society and investigative journalists, thereby undermining processes of social accountability.[65]

In the United States, the NIST AI RMF was developed in response to the National Artificial Intelligence Initiative Act of 2020.[66] The NIST AI RMF aims to help organisations map (recognising the context and associated risks), measure (identify, analyse, and track risks), manage (prioritise and address risks), and govern (develop a "culture of risk management") risks arising from AI systems. The RMF aims to help business align their practices to values around "human centricity, social responsibility, and sustainability".[67] The expectation of professional responsibility on the part of AI developers stems from the ISO standard ISO/IEC TR 24368:2022, which requires such developers to recognise the consequences of their actions. The RMF is intended to be voluntary, flexible, and sector-agnostic to aid its uptake and use by a diverse set of organisations of varying sizes.[68]

The RMF was followed with a profile released in July 2024 on generative AI.[f] The profile demonstrated an implementation of RMF categories for a particular class of technology—in this case, generative AI.[69] This reinforced the life-cycle approach to risk management undertaken by the RMF that attempts to align such practices within business goals, legal and regulatory requirements, and best practices.

Since then, there have been efforts to draw synergies or 'crosswalks' between the NIST RMF and other frameworks released by non-profits, industry associations, standards bodies, policies, and legislations.[g,70]

Apart from national-level frameworks, several state laws require AI safety and risk assessments. The New York City Bias Audit Law requires employers and employment agencies deploying automated technologies for hiring—ranging from resume filtering to advanced candidate assessment—to audit annually for bias with regard to protected characteristics by an independent third-party.[71] Colorado Senate Bill 21-169, enacted in 2021, aims to prevent discriminatory practices in the insurance sector using algorithms and predictive models trained on external sources of consumer information and requires insurers to establish a risk-management framework to identify and mitigate such risks.[72] Colorado Senate Bill 24-205, signed into law in May 2024, aims to establish consumer-protection regimes for users of AI systems, placing requirements on developers of high-risk systems to protect users from risks of algorithmic discrimination.[73]

f As per Section 4.1(a)(i)(A) of the Executive Order 14110 on Safe, Secure, and Trustworthy AI.

g These include the BSA Framework to Build Trust in AI; ISO/IEC FDIS23894 Information Technology- AI—Guidance on risk management; Executive Order 13960 (Promoting the use of trustworthy AI in federal government); OECD Recommendations on AI; the (then) proposed EU AI Act; the Blueprint for AI Bill of Rights; Singapore's AI Verify Testing Framework; ISO 5338/5339; Japan's AI Guidelines for Business, and CLTC UC Berkeley's Taxonomy on Trustworthiness of AI. See: https://airc.nist.gov/AI_RMF_Knowledge_Base/Crosswalks

While the EU has taken a horizontal approach to AI regulation, the US's market-driven landscape presents a more mixed bag, with NIST's voluntary risk assessments combined with vertical sector-specific state legislations. Each of these approaches has procedural variations with regard to compliance and non-standard discovery, notice, and disclosure requirements. Moreover, the use of general-purpose technologies like generative AI defies classification into a singular risk category because of the diversity of application across contexts.[74] Further, definitions of risk categories and AI systems tend to be broad, which can potentially include other, "simpler" software systems with predictive power within their ambit along with onerous and unmerited requirements.[75]

Some researchers argue that local laws such as New York City's are not instrumental in creating an effective audit regime due to definitional issues around key components and practices, such as who constitutes an independent auditor and differences between vendors creating the AI products under consideration and the companies deploying them.[76] Unlike critiques of the EU AI Act, criticisms of these laws indicate that, due to industry lobbying, their definitions of what constitutes an automated decision-making tool has been narrowed down to an extent that results in most tools being exempt.[77] While New York law requires the provision of notice of the use of these technologies to candidates and employees, along with allowing them to request alternative selection or evaluation procedures, employers are not obligated to provide such alternatives.[78]

Standards

Standards provide a "formula that describes the best way of doing something", providing guidance to developers and adopters without formal legal constraints, thereby adopting a soft law and regulation approach.[79] Standards are particularly helpful in the absence of sovereign regulations and legislations governing development and use. In the case of AI technologies, standards aim to provide pathways for the traceability of normative values commonly found in AI principles and ethics documents. Standards bodies have primarily focused on AI risk assessment and management, as follows:

- ISO/IEC 23894: 2023 (Published): Draws on existing international standards on risk management (ISO 31000:2018) and AI concepts and terminologies (ISO/IEC 22989:2022) to offer guidance and concrete examples of risk management across the AI development life cycle by setting out "vertical and horizontal pathways for implementing the principles, processes, and frameworks that can be adapted to any organisation."[80]

- ISO/IEC 42001: 2023 (Published): Pertains to improving AI management systems within organisations with a focus on responsible development and management, enhancing trustworthiness, supporting compliance with legal and regulatory requirements, and fostering innovations within the above framework.[81]

- ISO/IEC 38507:2022 (Governance implications of the use of AI by organisations) (Published): Helps establish a chain of responsibility and accountability by laying down guidance for members of the governing body within an organisation.[82]

Apart from ISO standards, the Institute of Electrical and Electronics Engineers (IEEE) has released practice guidelines that range from organisational AI governance (IEEE P2863) to model processes for addressing ethical concerns during system design (IEEE 7000-2021).[83] The former is under development and aims to establish process steps for the implementation of governance criteria such as safety, transparency, accountability, responsibility, and bias reduction through auditing, training, and compliance in the development and deployment of AI within an organisation. The latter highlights considerations for ethical inclusion through transparent communication with stakeholders. It also enables the translation of value-based conceptions into design characteristics through elicitation from relevant stakeholders throughout the process, thereby enabling the traceability of ethical values within a framework of ethical risk-based design.

The TEC, Department of Telecommunications, India, has released standard 57050:2023, which outlines procedures for assessing and rating AI systems for fairness. It combines documentation processes in the form the seven-layer approach with a bias evaluation metric using a combination of mathematical and statistical approach. Certification under this standard involves a three-step process of bias assessment, determining the threshold for metrics and testing for bias under different scenarios to ensure that it performs equally well for all individuals.[84] While the TEC standard adopts a holistic approach that defines the thresholds against which performance metrics are supposed to be measured, in most cases, such criteria (e.g., transparency, bias, fairness) are often not clearly outlined.

The development and diffusion of international standards and their legitimacy are bound up in centres of innovation and regulatory power. This results in large parts of the world being subject to standards that are not adequately contextualised in relation to actors (public and private) and normative objectives enshrined in principles documents.[85] Moreover, the

inequality across nations in terms of AI components and capabilities inhibits some nations' ability to develop AI capabilities. Additionally, the power exerted by regulatory and innovation centres require transnational firms to adopt their compliance benchmarks, which leads to these firms incurring additional costs when providing customised product and service offerings across markets. The flexibility and adaptability offered by the TEC standard could offer a Global South approach that can be contextualised for developing countries.

Despite the existence of multiple approaches, the lack of coherence in practice inhibits the mainstreaming of AI audits. Bias evaluation and mitigation techniques often do not apply to intersectional identities,[86] documentation approaches may remain hidden behind NDAs and proprietary information, standards might not rise above prescriptive norms of practice into levels of operational execution, and regulations without the adequate triangulation of the other three components might not ensure compliance towards the responsible management of AI systems. This necessitates that any systematisation of the auditing practice ensures a holistic approach that can effectively evaluate the technical and social aspects of risk and bias.

Stakeholder Roles

Different stakeholders play critical roles within the auditing processes towards mainstreaming audits as a bridge between principles and practice. As illustrated above, private companies have been instrumental in developing fairness assessment toolkits, while governments around the world are increasingly incorporating AI audits in their regulations and assessment frameworks. Key issues of bias and discrimination as well as AI harms being brought to the fore highlight the work being done by independent researchers, civil society, academia, and journalists. Further, given the role of standards as a guiding mechanism, the role of standards-setting bodies in establishing norms of practice is also critical.

- **Industry bodies and technical organisations**: Industry bodies and technical organisations bring together the developers and deployers of AI systems. They have the capacity to work towards industry standards and best practices for *ex-ante* intervention in minimising and mitigating risks before such systems reach end users. Industry bodies and technical organisations can be instrumental in driving the consensus around documentation, vendor compliances in the supply chain, and safe practices and self-regulatory compliance mechanisms in response to changes and evolution of technology.

- **Government**: The government has a role to play in developing regulatory guidelines, specifying compliance and disclosure requirements, establishing standards, and instituting mandates and procedures for companies to demonstrate compliance. Working with the industry, it can be responsive to the evolving nature of AI technologies, sectoral priorities, and innovations. This involves specifying the nature of audits, the legal liability of auditors and their limitations, and actions to be taken if audited entities are found to be in breach of established safety standards.

- **Standards-setting bodies**: Standards-setting bodies help establish legitimacy towards modes of practice. Current standards on AI governance need to be complemented with standards on AI audits, including documentation and compliance checklists that can help streamline the audit processes.

- **Civil society organisations, academics, and researchers**: These stakeholders are instrumental as third-party auditors to provide appropriate checks and balances for first- and second-party audits. Third-party audits, such as for facial recognition and recidivism, have highlighted the need for the responsible development of AI systems and can reveal hitherto undetected safety concerns about the operations of these systems. In addition, multilateral organisations like the United Nations, the United Nations Educational, Scientific and Cultural Organization, and the International Telecommunication Union are instrumental in coordinating international norms and supporting capacity building and technical assistance.

AI governance is both a social and a technical phenomenon. Therefore, modes of safety assessment need to incorporate social and technical evaluation along with regulatory compliance. AI risks emerge from the interactions and interface between technical elements like data and algorithms, which can be unrepresentative and driven by design decisions that might define what is evaluated and what is excluded, and social elements like human feedback for training, live data received in use, and the nature of understanding the problem and the context for which the AI system is designed.

At present, AI audits, despite best efforts, do not reflect a joint optimisation of these two elements, which is vital to evaluate the existing state of the system with respect to its risk of social harm, bias, and discrimination. Moreover, there is limited consensus on the legal status of audits, disclosure requirements, and procedural standardisation of auditing practices.[87]

This factor is compounded by the diversity of governance mechanisms and the nature of the technology itself, which inhibits the lack of standardisation. This highlights the role of different stakeholders within the ecosystem and the role that they play in codifying a system of practice such as AI audits, ensuring procedural standardisation, consolidating compliance mechanism, and managing the diversity of expertise that is required for conducting AI audits.

- **Developing procedural standardisation**: Developing procedural standards are critical for ensuring the legitimacy of audit processes and ensuring that they fulfil their expected functions of maintaining the transparency and accountability of AI systems. Procedural standards ensure that audit requirements across given AI systems follow a unified and homogeneous process. Process standardisation would ideally standardise the conditions for establishing the evaluation metric and identifying the appropriate mitigation strategies. It would further triangulate against available documentation and qualitative information about the development process that is collected through checklists.

- **Outlining the expertise of AI audit teams**: The risks associated with AI systems are both technical and social. This highlights the needs for diverse skillsets, from computational mathematicians and statisticians, data scientists, specialists in the AI sub-domain, context specialists, compliance and regulatory experts, design researchers, and social scientists familiar with modes of discrimination and exclusion.

- **Determining the nature of regulation**: AI systems are a composition of sub-systems and models and differ on the basis of the nature of the technology, such as ML or computer vision. Moreover, they have diverse applications across different sectors such as healthcare, finance, and education. This highlights the need for careful considerations around the nature of regulation, i.e., whether it needs to be sector agnostic or sector-specific or technology agnostic or technology specific.

Anulekha Nandi *is a Fellow in Technology, Economy and Society at ORF.*

Endnotes

1 Nicholas Kluge Corrêa et al., "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance," *Patterns* 4, no. 5 (2023), https://doi.org/10.1016/j.patter.2023.100857

2 Brent Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence* 1 (2019), https://doi.org/10.1038/s42256-019-0114-4; Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Learning Intelligence* 1 (2019), https://doi.org/10.1038/s42256-019-0088-2

3 Statista, "Number of New Artificial Intelligence Ethics Principles from 2015 to 2020, by Organization Type," https://www.statista.com/statistics/1286900/ai-ethics-principles-by-organization-type/

4 Corrêa et al., "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance"

5 Stanford University Human-Centered Artificial Intelligence, *Artificial Intelligence Index Report, 2021*, Stanford University, https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf

6 Statista, "Artificial Intelligence (AI) Market Size Worldwide from 2020 to 2030," https://www.statista.com/forecasts/1474143/global-ai-market-size

7 Andrea Guillen and Emma Teodoro, "Embedding Ethical Principles into AI Predictive Tools for Migration Management in Humanitarian Action," *Social Sciences* 12, no. 2 (2023), https://www.mdpi.com/2076-0760/12/2/53

8 Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics* 26 (2020), https://link.springer.com/article/10.1007/s11948-019-00165-5; Pouria Akbarighatar, "Operationalizing Responsible AI Principles through Responsible AI Capabilities," *AI and Ethics* (2024), https://link.springer.com/article/10.1007/s43681-024-00524-4

9 Akbarighatar, "Operationalizing Responsible AI Principles through Responsible AI Capabilities"

10 Alex Hanna and Emily M. Bender, "AI Causes Real Harm. Let's Focus on that Over the End-of-Humanity Hype," *Scientific American*, August 12, 2023, https://www.scientificamerican.com/article/we-need-to-focus-on-ais-real-harms-not-imaginary-existential-risks/

11 Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini, "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem" (paper presented in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) of the Association for Computing Machinery, New York, USA, 1571–1583. https://doi.org/10.1145/3531146.3533213); Corrêa et al., "Worldwide AI Ethics"

Endnotes

12 "EU AI Act: First Regulation on Artificial Intelligence," European Parliament, June 6, 2023, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

13 "Local Law 144 of 2021: Automated Employment Decision Tools (AEDT)," *NYC Consumer and Worker Protection*, https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page; "New York City Bias Audit," Holistic AI, https://www.nycbiasaudit.com/

14 NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0), January 2023*, U.S. Department of Commerce, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

15 Danaë Metaxa et al., "Auditing Algorithms: Understanding Algorithmic Systems from the Outside In," *Foundations and Trends in Human–Computer Interaction* 14, no. 4 (2021), https://ieeexplore.ieee.org/document/9627858

16 Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018), https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

17 Buolamwini and Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification"

18 Metaxa et al., "Auditing Algorithms: Understanding Algorithmic Systems from the Outside In"; Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

19 Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

20 "EU AI Act: First Regulation on Artificial Intelligence," European Parliament, June 18, 2024, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

21 "Artificial Intelligence Risk Management Framework (AI RMF 1.0)"

22 "Safety First: AI Models by OpenAI, Anthropic to Undergo Testing before US Rollouts," *The Indian Express*, September 2, 2024, https://indianexpress.com/article/technology/artificial-intelligence/ai-models-openai-anthropic-safety-testing-us-9546332/

23 Aditi Agarwal, "In Revised AI Advisory, IT Ministry Removes Requirement for Govt Permission," *Hindustan Times,* March 15, 2024, https://www.hindustantimes.com/india-news/in-revised-ai-advisory-it-ministry-removes-requirement-for-government-permission-101710520296018.html

24 Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

25 Raji et al., "Closing the AI Accountability Gap: Defining an End-to-End

Endnotes

Framework for Internal Algorithmic Auditing" (paper presented at FAT* '20: Conference on Fairness, Accountability, and Transparency, 2020), https://arxiv.org/abs/2001.00973

26 Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

27 Samir Saran, Anulekha Nandi, and Sameer Patil, *'Moving Horizons': A Responsive and Risk-Based Regulatory Framework for A.I.*, Observer Research Foundation, 2024, https://www.orfonline.org/research/moving-horizons-a-responsive-and-risk-based-regulatory-framework-for-a-i

28 Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

29 Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

30 Alex C. Engler, "Independent Auditors are Struggling to Hold AI Companies Accountable," Fast Company, January 26, 2021, https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue

31 Engler, "Independent Auditors are Struggling to Hold AI Companies Accountable"

32 Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

33 Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

34 Ellen P. Goodman and Julia Trehu, "Algorithmic auditing: Chasing AI Accountability," *Santa Clara High Technology Law Journal* 39, no. 3 (2023), https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?params=/context/chtlj/article/1689/&path_info=11_Goodman___Algorithmic_Auditing_Chasing_AI_Accountability_PUBLISHED.pdf

35 Goodman and Trehu, "Algorithmic Auditing: Chasing AI Accountability"

36 Jeff Saltz, "What is the AI Life Cycle?," Data Science Process Alliance, March 31, 2024, https://www.datascience-pm.com/ai-lifecycle/

37 Antti Lyyra, "From Components to Compositions: (De-)construction of Computer-Controlled Behaviour with the Robot Operating System" (PhD diss., London School of Economics, 2018), https://etheses.lse.ac.uk/3837/

38 Lyyra, "From Components to Compositions: (De-)construction of Computer-Controlled Behaviour with the Robot Operating System"; Jannis Kallinikos, "The Order of Technology: Complexity and Control in a Connected World," *Information and Organisation* 15, no. 3 (2005), https://doi.org/10.1016/j.infoandorg.2005.02.001

Endnotes

39 Dimitris Chorafas, *Systems and Simulation* (New York: Academic Press Inc., 1965); Kallinikos, "The Order of Technology: Complexity and Control in a Connected World"

40 Svetoslav Nizhnichenkov, Rahul Nair, Elizabeth Daly, and Brian Mac Namee, "Explaining Knock-On Effects of Bias Mitigation," arXiv, 2023, https://arxiv.org/html/2312.00765v1

41 Nizhnichenkov et al., "Explaining Knock-On Effects of Bias Mitigation," https://arxiv.org/html/2312.00765v1

42 "Shedding Light on AI Bias with Real World Examples," IBM, October 16, 2023, https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples

43 Jane Wakefield, "Microsoft Chatbot is Taught to Swear on Twitter," *BBC*, March 24, 2016, https://www.bbc.com/news/technology-35890188; Alex Hearn, "Microsoft Scrambles to Limit PR Damage over Abusive AI Bot Tay," *The Guardian*, March 24, 2016, https://www.theguardian.com/technology/2016/mar/24/microsoft-scrambles-limit-pr-damage-over-abusive-ai-bot-tay

44 Giskard Documentation, "Overconfidence," Giskard, https://docs.giskard.ai/en/stable/knowledge/key_vulnerabilities/overconfidence/index.html

45 Karen He, "AI Innovation and Ethics with AI Safety and Alignment," *Fiddler*, March 7, 2024, https://www.fiddler.ai/blog/ai-innovation-and-ethics-with-ai-safety-and-alignment

46 Emre Kazim, Adriano Soares Koshiyama, Airlie Hilliard, and Roseline Polle, "Systematizing Audit in Algorithmic Recruitment," *Journal of Intelligence* 9, no. 3 (2021), https://doi.org/10.3390/jintelligence9030046; NIST, *Artificial Intelligence Risk Management Framework*

47 Amazon Web Services, *Amazon AI Fairness and Explainability Whitepaper*, Amazon, https://pages.awscloud.com/rs/112-TZM-766/images/Amazon.AI.Fairness.and.Explainability.Whitepaper.pdf

48 "Amazon AI Fairness and Explainability Whitepaper"

49 Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst, "The Fallacy of AI Functionality" (paper presented at FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June, 2022), https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533158

50 Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366, no. 6464 (2019), https://pubmed.ncbi.nlm.nih.gov/31649194/

51 Danaë Metaxa et al., "Auditing Algorithms: Understanding Algorithmic Systems from the Outside In," *Foundations and Trends in Human–Computer Interaction* 14, no. 4 (2021), https://ieeexplore.ieee.org/document/9627858

Endnotes

52 Franklin Cardenoso Fernandez, "Bias Mitigation Strategies and Techniques for Classification Tasks," Holistic AI, June 8, 2023, https://www.holisticai.com/blog/bias-mitigation-strategies-techniques-for-classification-tasks

53 Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro, "Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey," *ACM Journal on Responsible Computing* 1, no. 2 (2024), https://dl.acm.org/doi/10.1145/3631326

54 Timnit Gebru et al., "Datasheets for Datasets," *Communications of the ACM* 64, no. 12 (2021), https://arxiv.org/abs/1803.09010

55 Margaret Mitchell et al., "Model Cards for Model Reporting" (paper presented in FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019), https://arxiv.org/abs/1810.03993; Timnit Gebru et al., "Datasheets for Datasets"

56 John Richards, David Piorkowski, Michael Hind, Stephanie Houde, and Aleksandra Mojsilović, "A Methodology for Creating AI FactSheets," arXiv, 2020, https://arxiv.org/pdf/2006.13796

57 Nekesha Green, Chavez Procope, Adeel Cheema, and Adekunle Adediji, "System Cards, a New Resource for Understanding how AI Systems Work," *Meta*, February 23, 2022, https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/

58 Telecommunications Engineering Centre, *Fairness Assessment and Rating of Artificial Intelligence Systems*, Department of Telecommunications, New Delhi, 2024, https://www.tec.gov.in/pdf/SDs/TEC%20Draft%20Standard%20for%20fairness%20assessment%20and%20rating%20of%20AI%20systems%20final%202022_12_27.pdf; Avinash Agarwal and Harsh Agarwal, "A Seven-Layer Model with Checklists for Standardising Fairness Assessment throughout the AI Lifecycle," *AI Ethics* 4, (2024), https://doi.org/10.1007/s43681-023-00266-9

59 Ernst & Young and Trilateral Research, "A Survey of Artificial Intelligence Risk Assessment Methodologies," Ernst & Young, 2022, https://www.trilateralresearch.com/wp-content/uploads/2022/01/A-survey-of-AI-Risk-Assessment-Methodologies-full-report.pdf

60 Margot E. Kaminski, "Regulating the Risks of AI," *Boston University Law Review* 103, no. 1347 (2023), https://www.bu.edu/bulawreview/files/2023/11/KAMINSKI.pdf

61 Kaminski, "Regulating the Risks of AI"

62 Siddhant Chatterjee, "Conformity Assessments in the EU AI Act: What You Need to Know," Holistic AI, August 7, 2023, https://www.holisticai.com/blog/conformity-assessments-in-the-eu-ai-act

63 Chatterjee, "Conformity Assessments in the EU AI Act: What You Need to Know"

64 "AI Auditing," European Data Protection Board, https://www.edpb.europa.eu/

Endnotes

our-work-tools/our-documents/support-pool-experts-projects/ai-auditing_en

65 David Hartmann, José Renato Laranjeira de Pereira, Chiara Streitbörger, and Bettina Berendt, "Addressing the Regulatory Gap: Moving Towards an EU AI Audit Ecosystem Beyond the AIA by Including Civil Society," arXiv, 2024, https://arxiv.org/html/2403.07904v1

66 NIST, *Artificial Intelligence Risk Management Framework*

67 NIST, *Artificial Intelligence Risk Management Framework*

68 NIST, *Artificial Intelligence Risk Management Framework*

69 NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)," July 2024, U.S. Department of Commerce, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

70 NIST, "Crosswalk Documents," *Trustworthy & Responsible AI Resource Center*, https://airc.nist.gov/AI_RMF_Knowledge_Base/Crosswalks

71 "Local Law 144 of 2021"; "New York City Bias Audit"

72 "SB21-169 - Protecting Consumers from Unfair Discrimination in Insurance Practices," *Colorado Department of Regulatory Agencies, Department of Insurance*, https://doi.colorado.gov/for-consumers/sb21-169-protecting-consumers-from-unfair-discrimination-in-insurance-practices

73 "SB24-205 Consumer Protections for Artificial Intelligence: Concerning Consumer Protections in Interactions with Artificial Intelligence Systems," *Colorado General Assembly 2024 Regular Session*, https://leg.colorado.gov/bills/sb24-205

74 Tamlin T. Higgins, "The EU AI Act: Concerns and Criticism," Clifford Chance, April 6, 2023, https://www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2023/04/the-eu-ai-act--concerns-and-criticism.html

75 Higgins, "The EU AI Act: Concerns and Criticism"

76 Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait, "Auditing Work: Exploring the New York City Algorithmic Bias Audit Regime" (paper presented at the Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), Association for Computing Machinery, New York, NY, USA, 1107–1120, https://doi.org/10.1145/3630106.3658959

77 Groves et al., "Auditing Work: Exploring the New York City Algorithmic Bias Audit Regime"

78 James A. Sherer and Brittany A. Yantis, "What About the Robots that are Already Here? New York City to Begin Enforcement of Artificial Intelligence Applications Related to Applicants and Employees Through the NYC Automated Employment Decision Tools Law on July 5, 2023," Baker Hostetler, April 7, 2023, https://www.

Endnotes

bakerdatacounsel.com/blogs/what-about-the-robots-that-are-already-here-new-york-city-begin-enforcement-artificial-intelligence-applications-july-5-2023/

79 Alicia Solow-Niederman, "Can AI Standards Have Politics?," *UCLA Law Review* (2024), https://www.uclalawreview.org/can-ai-standards-have-politics/

80 Tim McGarr, "ISO/IEC 23894 – A New Standard for Risk Management of AI," *AI Standards Hub*, https://aistandardshub.org/a-new-standard-for-ai-risk-management; "ISO/IEC 23894:2023 Information Technology — Artificial Intelligence — Guidance on Risk Management," *ISO*, 2023, https://www.iso.org/standard/77304.html

81 "ISO/IEC 42001:2023 Information technology — Artificial Intelligence — Management System," *ISO*, 2023, https://www.iso.org/standard/81230.html; Reto P. Grubenmann and Flavia Masoni, "ISO/IEC 42001: The Latest AI Management System Standard," KPMG, April 10, 2024, https://kpmg.com/ch/en/insights/technology/artificial-intelligence-iso-iec-42001.html

82 "ISO/IEC 38507:2022 Information Technology — Governance of IT — Governance Implications of the Use of Artificial Intelligence by Organizations," *ISO*, 2022, https://www.iso.org/standard/56641.html; Ernst & Young and Trilateral Research, "A Survey of Artificial Intelligence Risk Assessment Methodologies"

83 Ernst & Young and Trilateral Research, "A Survey of Artificial Intelligence Risk Assessment Methodologies"

84 Telecommunications Engineering Centre, *Fairness Assessment and Rating of Artificial Intelligence Systems*

85 Alicia Solow-Niederman, "Can AI Standards Have Politics?"

86 Wei Guo and Aylin Caliskan, "Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases" (paper presented at Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021), https://arxiv.org/abs/2006.03955

87 Costanza-Chock et al., "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem"

Images used in this paper are from Getty Images/Busà Photography.

Endnotes

ORF OBSERVER RESEARCH FOUNDATION

Ideas . Forums . Leadership . Impact

20, Rouse Avenue Institutional Area,
New Delhi - 110 002, INDIA
Ph. : +91-11-35332000. Fax : +91-11-35332005
E-mail: contactus@orfonline.org
Website: www.orfonline.org