Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 7 Issue 10, October - 2020

A Design Framework For Auditing AI

Ayse Kok Arslan

***Abstract*—Rising concern for the societal implications of artificial intelligence systems has inspired a wave of academic and journalistic literature in which deployed systems are audited for harm by investigators from outside the organizations deploying the algorithms. However, it remains challenging for practitioners to identify the harmful repercussions of their own systems prior to deployment, and, once deployed, emergent issues can become difficult or impossible to trace back to their source. This research study introduces a framework for algorithmic auditing that supports artificial intelligence system development end-to-end, to be applied throughout the internal organization development lifecycle. Each stage of the audit yields a set of documents that together form an overall audit report, drawing on an organization's values or principles to assess the fit of decisions made throughout the process. The proposed auditing framework is intended to contribute to closing the accountability gap in the development and deployment of large-scale artificial intelligence systems by embedding a robust process to ensure audit integrity.**

1. INTRODUCTION

With the increased access to artificial intelligence (AI) development tools and Internet-sourced datasets, corporations, nonprofits and governments are deploying AI systems at an unprecedented pace, often in massive-scale production systems impacting millions if not billions of users [1]. In the midst of this widespread deployment, however, come valid concerns about the effectiveness of these automated systems for the full scope of users, and especially a critique of systems that have the propensity to replicate, reinforce or amplify harmful existing social biases [8, 37, 62]. External audits are designed to identify these risks from outside the system and serve as accountability measures for these deployed models. However, such audits tend to be conducted after model deployment, when the system has already negatively impacted users [26, 51].

This study presents internal algorithmic audits as a mechanism to check that the engineering processes involved in AI system creation and deployment meet declared ethical expectations and standards, such as organizational AI principles. The audit process is necessarily boring, slow, meticulous and methodical—antithetical to the typical rapid development pace for AI technology. However, it is critical to slow down as algorithms continue to be deployed in increasingly high-stakes domains [20]. Executed by a dedicated team of organization employees, internal audits operate within the product development context and can inform the ultimate decision to abandon the development of AI technology when the risks outweigh the benefits (see Figure 1).

Inspired by the practices and artifacts of several disciplines, a defined internal audit framework-SMACTR- is developed to guide practical implementations. The framework strives to establish interdisciplinarity as a default in audit and engineering processes while providing the much needed structure to support the conscious development of AI systems.

2. GOVERNANCE, ACCOUNTABILITY AND AUDITS

Accountability refers to the state of being responsible or answerable for a system, its behavior and its potential impacts [38]. Although algorithms themselves cannot be held accountable as they are not moral or legal agents [7], the organizations designing and deploying algorithms can be held accountable through governance structures. Proposed standard ISO 37000 defines this structure as "the system by which the whole organization is directed, controlled and held accountable to achieve its core purpose over the long term."

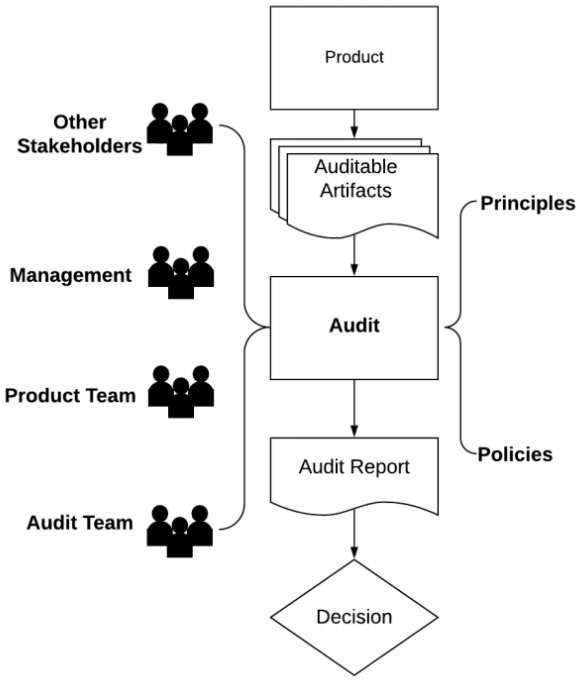

Figure 1: High-level overview of the context of an internal algorithmic audit.

As seen in Figure 1, the audit is conducted during product development and prior to launch. The audit team leads the product team, management and other

stakeholders in contributing to the audit. Policies and principles, including internal and external ethical expectations, also feed into the audit to set the standard for performance.

In environmental studies, Lynch and Veland [45] introduced the concept of urgent governance, distinguishing between auditing for system reliability versus societal harm. For example, a power plant can be consistently productive while causing harm to the environment through pollution [42]. Similarly, an AI system can be found technically reliable and functional through a traditional engineering quality assurance pipeline without meeting declared ethical expectations. A separate governance structure is necessary for the evaluation of these systems for ethical compliance. This evaluation can be embedded in the established quality assurance workflow but serves a different purpose, evaluating and optimizing for a different goal centered on social benefits and values rather than typical performance metrics such as accuracy or profit [39]. Although concerns about reliability are related, and although practices for testing production AI systems are established for industry practitioners [4], issues involving social impact, downstream effects in critical domains, and ethics and fairness concerns are not typically covered by concepts such as technical debt and reliability engineering.

2.1 Defining the Audit

Audits are tools for interrogating complex processes, often to determine whether they comply with company policy, industry standards or regulations [43]. The IEEE standard for software development defines an audit as “an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures” [32].

Building from methods of external auditing in investigative journalism and research [17, 62, 65], algorithmic auditing has started to become similar in spirit to the well-established practice of bug bounties, where external hackers are paid for finding vulnerabilities and bugs in released software [46]. These audits, modeled after intervention strategies in information security and finance [62], have significantly increased public awareness of algorithmic accountability. An external audit of automated facial analysis systems exposed high disparities in error rates among darker-skinned women and lighter-skinned men [8], showing how structural racism and sexism can be encoded and reinforced through AI systems. Such findings demonstrate the need for companies to understand the social and power dynamics of their deployed systems’ environments, and record such insights to manage their products’ impact.

2.2 External vs. Internal Audit

External auditing, in which companies are accountable to a third party [62], are fundamentally limited by lack of access to internal processes at the audited organizations. Although external audits conducted by credible experts are less affected by internal considerations, external auditors can only access model outputs, for example by using an API [65]. Auditors do not have access to intermediate models or training data, which are often protected as trade secrets [9].

Internal auditors’ direct access to systems can help extend traditional external auditing paradigms by incorporating additional information typically unavailable for external evaluations to reveal previously unidentifiable risks. Internal audits aim to evaluate how well the product candidate, once in real-world operation, will fit the expected system behavior encoded in standards. A modification in objective from a post-deployment audit to pre-deployment audit applied throughout the development process enables proactive ethical intervention methods, rather than simply informing reactive measures only implementable after deployment, as is the case with a purely external approach.

As the audit results can lead to ambiguous conclusions, it is critical to identify key stakeholders and decision makers who can drive appropriate responses to audit outcomes. Ultimately, internal audits complement external accountability, generating artifacts or transparent information [70] that third parties can use for external auditing, or even end-user communication.

2.3 Audit Integrity and Procedural Justice

Audit results are at times approached with skepticism since they are reliant on and vulnerable to human judgment. To establish the integrity of the audit itself as an independently valid result, the audit must adhere to the proper execution of an established audit process. This is a repeatedly observed phenomenon in tax compliance auditing, where several international surveys of tax compliance demonstrate that a fixed and vetted tax audit methodology is one of the most effective strategies to convince companies to respect audit results and pay their full taxes [22, 53].

Procedural justice implies the legitimacy of an outcome due to the admission of a fair and thorough process. Establishing procedural justice to increase compliance is thus a motivating factor for establishing common and robust frameworks through which independent audits can demonstrate adherence to standards.

2.4 AI Principles as Customized Ethical Standards

Important values such as ensuring AI technologies are subject to human direction and control, and avoiding the creation or reinforcement of unfair bias, have been included in many organizations’ ethical charters. However, the AI industry lacks proven methods to translate principles into practice [49], and AI principles have been criticized for being vague and providing little to no means of accountability [27, 82].

Nevertheless, such principles are becoming common methods to define the ethical priorities of an organization and thus the operational goals for which to aim [34, 83]. Therefore, in the absence of more formalized and universal standards, they can be used as a guide for the evaluation of the development lifecycle, and internal audits can investigate alignment with declared AI principles prior to model deployment.

This study proposes a framing of risk analyses centered on the failure to achieve AI principle objectives, outlining an audit practice that can begin translating ethical principles into practice.

2.5 An Overview of Audit Approaches

Current software development practice in general, and artificial intelligence development in particular, does not typically follow the waterfall or verification-and-validation approach [16]. These approaches are still used, in combination with agile methods, in the above-mentioned industries because they are much more documentation-oriented, auditable and requirements-driven.

Agile artificial intelligence development is much faster and iterative, and thus presents a challenge to auditability. However, applying agile methodologies to internal audits themselves is a current topic of research in the internal audit profession. Most internal audit functions outside of heavily regulated industries tend to take a risk-based approach. They work with product teams to ask "what could go wrong" at each step of a process and use that to build a risk register [59]. This allows risks to rise to the surface in a way that is informed by the people who know these processes and systems the best.

Moreover, there is a dynamic complex interaction between users as sources of data, data collection, and model training and updating. Additionally, governance processes based solely on risk have been criticized for being unable to anticipate the most profound impacts from technological innovation, such as the financial crisis in 2008, in which big data and algorithms played a large role [52, 54, 57].

As Scully et al. point out, AI models create entanglement and make the isolation of improvements effectively impossible [67], which they call ‘Change Anything Change Everything’. One suggestion might be to have explicit documentation about the purpose, data, and model space, potential hazards which could be identified earlier in the development process. Selbst and Barocas argue that “one must seek explanations of the process behind a model‘s development, not just explanations of the model itself” [68].

Also, as AI is at times considered a “general purpose technology” with multiple and dual uses [78], the lack of reliable standardization poses significant challenges to governance efforts. This challenge is compounded by increasing customization and variability of what an AI product development lifecycle looks like depending on the anticipated context of deployment or industry.

3. AN INTERNAL AUDIT FRAMEWORK: SMACTR

This paper will now outline the components of an initial internal audit framework, which can be framed as encompassing five distinct stages— Scoping, Mapping, Artifact Collection, Testing and Reflection (SMACTR)— all of which have their own set of documentation requirements and account for a different level of the analysis of a system. Figure 2 illustrates the full set of artifacts recommended for each stage.

In Figure 2, the color gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed. To illustrate the utility of this framework, this paper contextualizes descriptions with the hypothetical example of Company X Inc., a large multinational software engineering consulting firm, specializing in developing custom AI solutions for a diverse range of clients.

Let’s imagine this company has designated five AI principles, paraphrased from the most commonly identified AI principles in a current online English survey [34]–"Transparency", "Justice, Fairness & Non-Discrimination", "Safety & Non-Maleficence", "Responsibility & Accountability" and "Privacy".Let’s assume that the corporate structure of Company X is typical of any technical consultancy, and design our stakeholder map by this assumption.

Let’s imagine the pilot implementation of the SMACTR internal audit framework on two hypothetical client projects:

- The first (hypothetical) client wishes to develop a child abuse screening tool similar to that of the real cases extensively studied and reported on [11, 14, 15, 21, 25, 36]. This complex case intersects heavily with applications in high-risk scenarios with dire consequences. This scenario demonstrates how, for algorithms interfacing with high-risk contexts, a structured framework can allow for the careful consideration of all the possibilities and risks with taking on the project, and the extent of its understood social impact.

- The second invented client is Happy-Go-Lucky, Inc., an imagined photo service company looking for a smile detection algorithm to automatically trigger the cameras in their installed physical photo booths. In this scenario, the worst case is a lack of customer satisfaction—the stakes are low and the situation seems relatively straightforward. This scenario demonstrates how in even seemingly simple and benign cases, ethical consideration of system deployment can reveal underlying issues to be addressed prior to deployment, especially when the model is contextualized within the setting of the

product and deployment environment. An end-to-end working example of the audit framework would ideally include demonstrative templates of all recommended documentation, specific process files such as any experimental results, interview transcripts, a design history file and the summary report. Workable templates can also be accessed as an online resources.

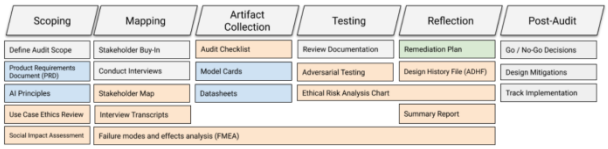

Figure 2: Overview of Internal Audit Framework

3.1 The Governance Process

To design the audit procedure, formal risk assessment methodologies can be complemented with ideas from responsible innovation, which stresses four key dimensions: anticipation, reflexivity, inclusion and responsiveness [73], as well as system-theoretic concepts that help grapple with increasing complexity and coupling of AI systems with the external world [42].

Risk-based assessments can be limited in their ability to capture social and ethical stakes, and they should be complemented by anticipatory questions such as, “what if.?”. The aim is to increase ethical foresight through systematic thinking about the larger socio-technical system in which a product will be deployed [50]. At a minimum, the internal audit process should enable critical reflections on the potential impact of a system, serving as internal education and training on ethical awareness in addition to leaving what is referred to as a “transparency trail” of documentation at each step of the development cycle (see Figure 2).

Each industry has a way to judge what requires a full audit, but that process is discretionary and dependent on a range of contextual factors pertinent to the industry, the organization, audit team resourcing, and the case at hand.

3.2 The Scoping Stage

The goal of the scoping stage is to clarify the objective of the audit by reviewing the motivations and intended impact of the investigated system, and confirming the principles and values meant to guide product development. This is the stage in which the risk analysis begins by mapping out intended use cases and identifying analogous deployments either within the organization or from competitors or adjacent industries.

In the case of the smile-triggered phone booth, a smile detection model is required, providing a simple product, with not a broad scope of considerations as the potential for harm does not go much beyond inconvenience or customer exclusion and dissatisfaction.

For the child abuse detection product, there are many more approaches to solving the issue and many more options for how the model interacts with the broader system. The use case itself involves many ethical considerations, as an ineffective model may result in serious consequences like death or family separation.

The key artifacts developed by the auditors from this stage include an ethical review of the system use case and a social impact assessment. Pre-requisite documents from the product and engineering team should be a declaration or confirmation statement of ethical objectives, standards and AI principles.

3.2.1 Artifact: Ethical Review of System Use Case

When a potential AI system is in the development pipeline, it should be reviewed with a series of questions that first and foremost check to see, at a high level, whether the technology aligns with a set of ethical values or principles. This can take the form of an ethical review that considers the technology from a responsible innovation perspective by asking who is likely to be impacted and how.

Algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values. Thus it is not always possible for individual technology workers to identify or assess their own biases or faulty assumptions [33]. For this reason, a critical range of viewpoints is included in the review process. The essential inclusion of independent domain experts and marginalized groups in the ethical review process "has the potential to lead to more rigorous critical reflection because their experiences will often be precisely those that are most needed in identifying problematic background assumptions and revealing limitations with research questions, models, or methodologies" [33].

3.2.2 Artifact: Social Impact Assessment

Social impact assessments are commonly defined as a method to analyze and mitigate the unintended social consequences, both positive and negative, that occur when a new development, program, or policy engages with human populations and communities [79].

It describes how the use of an artificial intelligence system might change people’s ways of life, their culture, their community, their political systems, their environment, their health and well-being, their personal and property rights, and their experiences (positive or negative) [79]. The social impact assessment includes two primary steps: an assessment of the severity of the risks, and an identification of the relevant social, economic, and cultural impacts and harms that an artificial intelligence system applied in context may create.

3.3 The Mapping Stage

The mapping stage is not a step in which testing is actively done, but rather a review of what is already in place and the perspectives involved in the audited

system. This is also the time to map internal stakeholders, identify key collaborators for the execution of the audit, and orchestrate the appropriate stakeholder buy-in required for execution.

For the child abuse detection algorithm, the initial identification of failure modes reveals the high stakes of the application, and immediate threats to the "Safety & Non-Maleficence" principle. False positives overwhelm staff and may lead to the separation of families that could have recovered. False negatives may result in a dead or injured child that could have been rescued. For the smile detector, failures disproportionately impact those with alternative emotional expressions—those with autism, different cultural norms on the formality of smiling, or different expectations for the photograph who are then excluded from the product by design.

The key artifacts from this stage include a stakeholder map and collaborator contact list, a system map of the product development lifecycle, and the engineering system overview, especially in cases where multiple models inform the end product. Finally, it includes a report or interview transcripts on key findings from internal ethnographic fieldwork involving the stakeholders and engineers.

3.3.1 Artifact: Stakeholder Map

Who was involved in the system audit and collaborators in the execution of the audit should be outlined. Clarifying participant dynamics ensures a more transparent representation of the provided information, giving further context to the intended interpretation of the final audit report.

3.3.2 Artifact: Ethnographic Field Study

As Leveson points out, bottom-up decentralized decision making can lead to failures in complex socio-technical systems [42]. Each local decision may be correct in the limited context in which it was made, but can lead to problems when these decisions and organizational behaviors interact. Therefore, ethnography-inspired fieldwork methodology based on how audits are conducted in other industries, such as finance [74] and healthcare [64] is useful to get a deeper and qualitative understanding of the engineering and product development process.

Traditional metrics for AI like loss may conceal fairness concerns, social impact risks or abstraction errors [69]. Taking metrics measured in isolation risks recapitulating the abstraction error that [69] point out, "To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore to make a category error, or an abstraction error."

What is considered as data is already an interpretation, highly subjective and contested [23]. During the interviews, auditors should capture and pay attention to what falls outside the measurements and metrics, and to render explicit the assumptions and values the metrics apprehend [75]. For example, the decision about whether to prioritize the false positive rate over false negative rate (precision/recall) is a question about values and cannot be answered without stating the values of the organization, team or even engineer within the given development context.

3.4 The Artifact Collection Stage

At this stage, all the required documentation from the product development process is identified and collected in order to prioritize opportunities for testing. Often this implies a record of data and model dynamics though application-based systems can include other product development artifacts such as design documents and reviews, in addition to systems architecture diagrams and other implementation planning documents and retrospectives.

The key artifact from auditors during this stage is the audit checklist, one method of verifying that all documentation pre-requisites are provided in order to commence the audit.

3.4.1 Artifact: Design Checklist

This checklist is a method of taking inventory of all the expected documentation to have been generated from the product development cycle.

3.4.2 Artifacts: Datasheets and Model Cards

Two recent standards can be leveraged to create auditable documentation, model cards and datasheets [24, 48].

To clarify the intended use cases of AI models and minimize their usage in contexts for which they are not well suited, Mitchell et al. recommend that released models be accompanied by documentation detailing their performance characteristics [48], called a model card. This should include information about how the model was built, what assumptions were made during development, and what type of model behavior might be experienced by different cultural, demographic or phenotypic groups. A model card is also extremely useful for internal development purposes to make clear to stakeholders details about trained models that are included in larger software pipelines, which are parts of internal organizational dynamics, which are then parts of larger socio-technical logics and processes.

Model cards are intended to complement "Datasheets for Datasets" [24].A critical part of the datasheet covers the data collection process required to make informed decisions about using the dataset: what mechanisms or procedures were used to collect the data? Was any ethical review process conducted? Does the dataset relate to people? This documentation feeds into the auditors' assessment process.

3.5 The Testing Stage

This stage is where the majority of the auditing team's testing activity is done—when the auditors execute a

series of tests to gauge the compliance of the system with the prioritized ethical values of the organization. Auditors review the documentation collected from the previous stage and begin to make assessments of the likelihood of system failures to comply with declared principles. High variability in approach is likely during this stage, as the tests that need to be executed change dramatically depending on organizational and system context. Auditors might employ counterfactual adversarial examples designed to confuse the model and find problematic failure modes.

For the child prediction model, performance on a selection of diverse user profiles can be tested. These profiles can also be treated for variables that correlate with vulnerable groups to test whether the model has learned biased associations with race or socio-economic status. For the ethical risk analysis chart, by looking at the principles one might realize that there might be immediate risks to the "Privacy" principle—with one case involving juvenile data, which is sensitive, and the other involving face data, a biometric. Also, in the smiling booth case, there might be disproportionate performance for certain underrepresented user subgroups, thus jeopardizing the "Justice, Fariness & Non-Discrimination" principle.

The main artifacts from this stage of the auditing process are the results of tests such as adversarial probing of the system and an ethical risk analysis chart.

3.5.1 Artifact: Adversarial Testing

In general, adversarial testing attempts to simulate what a hostile actor might do to gain access to a system, or to push the limits of the system into edge case or unstable behavior to elicit very-low probability but high-severity failures.

In this process, direct non-statistical testing uses tailored inputs to the model to see if they result in undesirable outputs. These inputs can be motivated by an intersectional analysis, for example where an ML system might produce unfair outputs based on demographic and phenotypic groups that might combine in non-additive ways to produce harm, or over time recapitulate harmful stereotypes or reinforce unjust social dynamics (for example, in the form of opportunity denial). This is distinct from adversarially attacking a model with human-imperceptible pixel manipulations to trick the model into misidentifying previously learned outputs [28], yet these approaches can be complementary.

3.5.2 Artifact: Ethical Risk Analysis Chart

The ethical risk analysis chart considers the combination of the likelihood of a failure and the severity of a failure to define the importance of the risk. Highly likely and dangerous risks are considered the most high-priority threats. Each risk is assigned a severity indication of "high", "mid" and "low" depending on their combination of these features. Failure likelihood is estimated by considering the occurrence of certain failures during the adversarial testing of the system and the severity of the risk is identified in earlier stages, from informative processes such as the social impact assessment and ethnographic interviews.

3.6 The Reflection Stage

This phase will reflect on product decisions and design recommendations that could be made following the audit results.

For the smile detection algorithm, the decision could be to train a new version of the model on more diverse data before considering deployment, and add more samples of underrepresented populations to the training data. It could be decided that the use case does not necessarily define affect, but treats smiling as a favorable photo pose. Design choices for other parts of the product outside the model should be considered—for instance, an opt-in functionality with user permissions required on the screen before applying the model-controlled function, and the default being that the model-controlled trigger is disabled. There could also be an included disclaimer on privacy, assuring users of safe practices for face data storage and consent.

For the child abuse detection model—this is a more complex decision. Given the ethical considerations involved, the project may be stalled or even cancelled, requiring further inquiry into the ethics of the use case, and the capability of the team to complete the mitigation plan required to deploy an algorithm in such a high risk scenario.

3.6.1 Artifact: Algorithmic Use-related Risk Analysis

The risk analysis should be informed by the social impact assessment and known issues with similar models. Careful attention must be paid to the distinction between the designers' mental models of the artificial intelligence system and the user's mental model. The designers' mental models are an idealization of the AI system before the model is released. Significant differences exist between this ideal model and how the actual system will behave or be used once deployed. This may be due to many factors, such as distributional drift [41] where the training and test set distributions differ from the real-world distribution, or intentional or unintentional misuse of the model for purposes other than those for which it was designed. Therefore, the user's mental model of the system should be anticipated and taken into consideration.

Christin points out "the importance of studying the practices, uses, and implementations surrounding algorithmic technologies. Intellectually, this involves establishing new exchanges between literatures that may not usually interact, such as critical data studies, the sociology of work, and organizational analysis".

3.6.2 Artifact: Remediation and Risk Mitigation Plan

After the audit is completed and findings are presented to the leadership and product teams, it is important to develop a plan for remediating these problems. The goal is to drive down the risk of ethical concerns or potential negative social impacts to the extent reasonably practicable. For the concerns raised in any audit against ethical values, a technical team might want to know: what is the threshold for acceptable performance? If auditors discover, for example, unequal classifier performance across subgroups, how close to parity is necessary to say the classifier is acceptable?

3.6.3 Artifact: Algorithmic Design History File

An algorithmic design history file (ADHF) would collect all the documentation from the activities outlined above related to the development of the algorithm. It could point to the documents necessary to demonstrate that the product or model was developed in accordance with an organization's ethical values, and that the benefits of the product outweigh any risks identified in the risk analysis process. The ADHF could also assist with an audit trail, enabling the reconstruction of key decisions and events during the development of the product.

3.6.4 Artifact: Algorithmic Audit Summary Report

The report aggregates all key audit artifacts, technical analyses and documentation, putting this in one accessible location for review.

4. LIMITATIONS OF INTERNAL AUDITS

The audit is never isolated from the practices and people conducting the audit, just as AI systems are not independent of their developers or of the larger socio-technical system. Audits are not unified or monolithic processes with an objective "view from nowhere", but must be understood as a "patchwork of coupled procedures, tools and calculative processes" [74]. To avoid audits becoming simply acts of reputation management for an organization, the auditors should be mindful of their own and the organizations' biases and viewpoints. Internal audits are only one important aspect of a broader system of required quality checks and balances.

5. CONCLUSION

AI has the potential to benefit the whole of society, however there is currently an inequitable risk distribution such that those who already face patterns of structural vulnerability or bias disproportionately bear the costs and harms of many of these systems. Fairness, justice and ethics require that those bearing these risks are given due attention and that organizations that build and deploy AI systems internalize and proactively address these social risks as well, being seriously held to account for system compliance to declared ethical principles.

REFERENCES

[1] Omar Y Al-Jarrah, Paul D Yoo, Sami Muhaidat, George K Karagiannidis, and Kamal Taha. 2015. Efficient machine learning for big data: A review. Big Data Research 2, 3 (2015), 87–93.

[2] Amel Bennaceur, Thein Than Tun, Yijun Yu, and Bashar Nuseibeh. 2019. Requirements Engineering. In Handbook of Software Engineering. Springer, 51–92.

[3] Li Bing, Akintola Akintoye, Peter J Edwards, and Cliff Hardcastle. 2005. The allocation of risk in PPP/PFI construction projects in the UK. International Journal of project management 23, 1 (2005), 25–35.

[4] Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, and D Sculley. 2017. The ml test score: A rubric for ml production readiness and technical debt reduction. In 2017 IEEE International Conference on Big Data (Big Data). IEEE, 1123–1132.

[5] Shona L Brown and Kathleen M Eisenhardt. 1995. Product development: Past research, present findings, and future directions. Academy of management review 20, 2 (1995), 343–378.

[6] Chad Brubaker, Suman Jana, Baishakhi Ray, Sarfraz Khurshid, and Vitaly Shmatikov. 2014. Using Frankencerts for Automated Adversarial Testing of Certificate Validation. In in SSL/TLS Implementations,âĂİ IEEE Symposium on Security and Privacy. Citeseer.

[7] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. Artificial Intelligence and Law 25, 3 (2017), 273–291.

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency. 77–91.

[9] Jenna Burrell. 2016. How the machine "thinks": Understanding opacity in machine learning algorithms. Big Data & Society 3, 1 (2016), 2053951715622512.

[10] Paul Eric Byrnes, Abdullah Al-Awadhi, Benita Gullvist, Helen Brown-Liburd, Ryan Teeter, J Donald Warren Jr, and Miklos Vasarhelyi. 2018. Evolution of Auditing: From the Traditional Approach to the Future Audit 1. In Continuous Auditing: Theory and Application. Emerald Publishing Limited, 285–297.

[11] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Conference on Fairness, Accountability and Transparency. 134–148.

[12] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. Big Data & Society 4, 2 (2017), 2053951717718855.

[13] Kai Lai Chung and Paul Erdös. 1952. On the application of the Borel-Cantelli lemma. Trans. Amer. Math. Soc. 72, 1 (1952), 179–186.

[14] Rachel Courtland. 2018. Bias detectives: the researchers striving to make algorithms fair. Nature 558, 7710 (2018), 357–357.

[15] Stephanie Cuccaro-Alamin, Regan Foust, Rhema Vaithianathan, and Emily Putnam-Hornstein. 2017. Risk assessment and decision making in child protective services: Predictive risk modeling in context. Children and Youth Services Review 79 (2017), 291–298.

[16] Michael A Cusumano and Stanley A Smith. 1995. Beyond the waterfall: Software development at Microsoft. (1995).

[17] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. (2014).

[18] Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli. 2018. A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. arXiv preprint arXiv:1807.00553 (2018).

[19] Kevin Driscoll, Brendan Hall, Håkan Sivencrona, and Phil Zumsteg. 2003. Byzantine fault tolerance, from theory to reality. In International Conference on Computer Safety, Reliability, and Security. Springer, 235–248.

[20] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway feedback loops in predictive policing. arXiv preprint arXiv:1706.09847 (2017).

[21] Virginia Eubanks. 2018. A child abuse prediction model fails poor families. Wired Magazine (2018).

[22] Sellywati Mohd Faizal, Mohd Rizal Palil, Ruhanita Maelah, and Rosiati Ramli. 2017. Perception on justice, trust and tax compliance behavior in Malaysia. Kasetsart Journal of Social Sciences 38, 3 (2017), 226–232.

[23] Jonathan Furner. 2016. “Data“: The data. In Information Cultures in the Digital Age. Springer, 287–306.

[24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. arXiv preprint arXiv:1803.09010 (2018).

[25] Jeremy Goldhaber-Fiebert and Lea Prince. 2019. Impact Evaluation of a Predictive Risk Modeling Tool for Allegheny CountyâĂŹs Child Welfare Office. Pittsburgh: Allegheny County.[Google Scholar] (2019).

[26] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-theloop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 90–99.

[27] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences.

[28] Shixiang Gu and Luca Rigazio. 2014. Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068 (2014).

[29] John Haigh. 2012. Probability: A very short introduction. Vol. 310. Oxford University Press.

[30] Brendan Hall and Kevin Driscoll. 2014. Distributed System Design Checklist. (2014).

[31] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need? arXiv preprint arXiv:1812.05239 (2018).

[32] IEEE. 2008. IEEE Standard for Software Reviews and Audits. IEEE Std 1028-2008 (Aug 2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584

[33] Kristen Intemann. 2010. 25 years of feminist empiricism and standpoint theory: Where are we now? Hypatia 25, 4 (2010), 778–796.

[34] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. Artificial Intelligence: the global landscape of ethics guidelines. arXiv preprint arXiv:1906.11668 (2019).

[35] Paul A Judas and Lorraine E Prokop. 2011. A historical compilation of software metrics with applicability to NASA‘s Orion spacecraft flight software sizing. Innovations in Systems and Software Engineering 7, 3 (2011), 161–170.

[36] Emily Keddell. 2019. Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice. Social Sciences 8, 10 (2019), 281.

[37] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. arXiv preprint arXiv:1805.04508 (2018).

[38] Nitin Kohli, Renata Barreto, and Joshua A Kroll. 2018. Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems. In 1st Conference on Fairness, Accountability, and Transparancy. New York, NY, USA. 7.

[39] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. U. Pa. L. Rev. 165 (2016), 633.

[40] Arie W Kruglanski. 1996. Motivated social cognition: Principles of the interface. (1996).

[41] Joel Lehman. 2019. Evolutionary Computation and AI Safety: Research Problems Impeding Routine and Safe Real-world Application of Evolution. arXiv preprint arXiv:1906.10189 (2019).

[42] Nancy Leveson. 2011. Engineering a safer world: Systems thinking applied to safety. MIT press.

[43] Jie Liu. 2012. The enterprise risk management and the risk oriented internal audit. Ibusiness 4, 03 (2012), 287.

[44] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision. 3730–3738.

[45] Amanda H Lynch and Siri Veland. 2018. Urgency in the Anthropocene. MIT Press.

[46] Thomas Maillart, Mingyi Zhao, Jens Grossklags, and John Chuang. 2017. Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. Journal of Cybersecurity 3, 2 (2017), 81–90.

[47] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. arXiv preprint arXiv:1901.10436 (2019).

[48] 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. ACM, 220–229.

[49] Brent Mittelstadt. 2019. AI Ethics: Too Principled to Fail? SSRN (2019).

[50] Brent Daniel Mittelstadt and Luciano Floridi. 2016. The ethics of big data: current and foreseeable issues in biomedical contexts. Science and engineering ethics 22, 2 (2016), 303–341.

[51] Laura Moy. 2019. How Police Technology Aggravates Racial Inequity: A Taxonomy of Problems and a Path Forward. Available at SSRN 3340898 (2019).

[52] Fabian Muniesa, Marc Lenglet, et al. 2013. Responsible innovation in finance: directions and implications. Responsible Innova-tion: Managing the Responsible Emergence of Science and Innovation in Society. Wiley, London (2013), 185–198.

[53] Kristina Murphy. 2003. Procedural justice and tax compliance. Australian Journal of Social Issues (Australian Council of Social Service) 38, 3 (2003).

[54] Safiya Umoja Noble. 2018. Algorithms of oppression: How search engines reinforce racism. nyu Press.

[55] Institute of Internal Auditors. Research Foundation and Institute of Internal Auditors. 2007. The Professional Practices Framework. Inst of Internal Auditors.

[56] General Assembly of the World Medical Association et al. 2014. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. The Journal of the American College of Dentists 81, 3 (2014), 14.

[57] Cathy O'neil. 2016. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

[58] Charles Parker. 2012. Unexpected challenges in large scale machine learning. In Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications. ACM, 1–6.

[59] Fiona D Patterson and Kevin Neailey. 2002. A risk register database system to aid the management of project risk. International Journal of Project Management 20, 5 (2002), 365–374.

[60] W Price and II Nicholson. 2017. Regulating black-box medicine. Mich. L. Rev. 116 (2017), 421.