Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing

Inioluwa Deborah Raji*
Partnership on AI
deb@partnershiponai.org

Andrew Smart*
Google
andrewsmart@google.com

Rebecca N. White
Google

Margaret Mitchell
Google

Timnit Gebru
Google

Ben Hutchinson
Google

Jamila Smith-Loud
Google

Daniel Theron
Google

Parker Barnes
Google

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.

In this paper, we introduce 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.

CCS CONCEPTS

• **Social and professional topics** → **System management**; **Technology audits**; • **Software and its engineering** → **Software development process management**.

KEYWORDS

Algorithmic audits, machine learning, accountability, responsible innovation

*Both authors contributed equally to this paper. This work was done by Inioluwa Deborah Raji as a fellow at Partnership on AI (PAI), of which Google, Inc. is a partner. This should not be interpreted as reflecting the official position of PAI as a whole, or any of its partner organizations.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02.
https://doi.org/10.1145/3351095.3372873

ACM Reference Format:
Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3351095.3372873

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].

In this paper, we present 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. By considering historical examples across industries, we make the case that such audits can be leveraged to anticipate potential negative consequences before they occur, in addition to providing decision support to design mitigations, more clearly defining and monitoring potentially adverse outcomes, and anticipating harmful feedback loops and system-level risks [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 from the practices and artifacts of several disciplines, we go further to develop SMACTR, a defined internal audit framework meant to guide practical implementations. Our 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

We use *accountability* to mean 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 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."[1] If the responsible development of artificial intelligence is a core purpose of organizations creating AI, then a governance system by which the whole organization is held accountable should be established.

[1]https://committee.iso.org/sites/tc309/home/projects/ongoing/ongoing-1.html

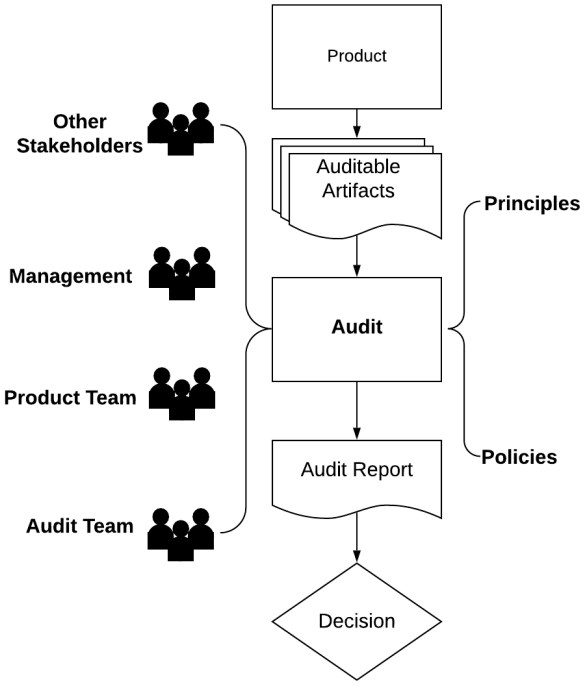

Figure 1: High-level overview of the context of an internal algorithmic audit. The audit is conducted during product development and prior to launch. The audit team leads the product team, management and other stakeholders in contributing to the audit. Policies and principles, including internal and 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 vs 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 What is an 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. [8] reveals *interaction failures*, in which the production and deployment of an AI system interacts with unjust social structures to contribute to biased predictions, as Safiya Noble has described [54]. 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 AI Principles as Customized Ethical Standards

According to Mittelstadt [49], at least 63 public-private initiatives have produced statements describing high-level principles, values and other tenets to guide the ethical development, deployment and governance of AI. 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]. Thus, in the absence of more formalized and universal standards, they can be used as a North Star to guide the evaluation of the development lifecycle, and internal audits can investigate alignment with declared AI principles prior to model deployment. We propose 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.

2.3 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. In addition, audit integrity is best established when auditors themselves live up to an ethical standard, vetted by adherence to an expected code of conduct or norm in how the audit is to be conducted. In finance, for example, it became clear that any sense of dishonesty or non-transparency in audit methodology would lead audit targets to dismiss rather than act on results [66].

2.4 The 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 organization-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 thus help extend traditional external auditing paradigms by incorporating additional information typically unavailable for external evaluations to reveal previously unidentifiable risks.

The goals of an internal audit are similar to quality assurance, with the objective to enrich, update or validate the risk analysis for product deployment. Internal audits aim to evaluate how well the product candidate, once in real-world operation, will fit the expected system behaviour 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. Because there is an increased level of system access in an internal audit, identified gaps in performance or processes can be mapped to sociotechnical considerations that should be addressed through joint efforts with product teams. 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.

Additionally, with an internal audit, because auditors are employees of the organization and communicate their findings primarily to an internal audience, there is opportunity to leverage these audit outcomes for recommendations of structural organizational changes needed to make the entire engineering development process auditable and aligned with ethical standards. 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. Internal audits can thus enable review and scrutiny from additional stakeholders, by enforcing transparency through stricter reporting requirements.

3 LESSONS FROM AUDITING PRACTICES IN OTHER INDUSTRIES

Improving the governance of artificial intelligence development is intended to reduce the risks posed by new technology. While not without faults, safety-critical and regulated industries such as aerospace and medicine have long traditions of auditable processes and design controls that have dramatically improved safety [77, 81].

3.1 Aerospace

Globally, there is one commercial airline accident per two million flights [63]. This remarkable safety record is the result of a joint and concerted effort over many years by aircraft and engine manufacturers, airlines, governments, regulatory bodies, and other industry stakeholders [63]. As modern avionic systems have increased in size and complexity (for example, the Boeing 787 software is estimated at 13 million lines of code [35]), the standard 1-in-1,000,000,000 per use hour maximum failure probability for critical aerospace systems remains an underappreciated engineering marvel [19].

However, as the recent Boeing 737 MAX accidents indicate, safety is never finished, and the qualitative impact of failures cannot be ignored—even one accident can impact the lives of many and is rightfully acknowledged as a catastrophic tragedy. Complex systems tend to drift toward unsafe conditions unless constant vigilance is maintained [42]. It is the sum of the tiny probabilities of individual events that matters in complex systems—if this grows without bound, the probability of catastrophe goes to one. The *Borel-Cantelli* Lemmas are formalizations of this statistical phenomenon [13], which means that we can never be satisfied with safety standards. Additionally, standards can be compromised if competing business interests take precedence. Because the non-zero risk of failure grows over time, without continuous active measures being developed to mitigate risk, disaster becomes inevitable [29].

3.1.1 Design checklists. Checklists are simple tools for assisting designers in having a more informed view of important questions, edge cases and failures [30]. Checklists are widely used in aerospace for their proven ability to improve safety and designs. There are several cautions about using checklists during the development of complex software, such as the risk of blind application, the broader

context and nuanced interrelated concerns are not considered. However, a checklist can be beneficial. It is good practice to avoid yes/no questions to reduce the risk that the checklist becomes a box-ticking activity, for example by asking designers and engineers to describe their processes for assessing ethical risk. Checklist use should also be related to real-world failures and higher-level system hazards.

3.1.2 Traceability. Another key concept from aerospace and safety-critical software engineering is *traceability*—which is concerned with the relationships between product requirements, their sources and system design. This practice is familiar to the software industry in requirements engineering [2]. However, in AI research, it can often be difficult to trace the provenance of large datasets or to interpret the meaning of model weights—to say nothing of the challenge of understanding how these might relate to system requirements. Additionally, as the complexity of sociotechnical systems is rapidly increasing, and as the speed and complexity of large-scale artificial intelligence systems increase, new approaches are necessary to understand risk [42].

3.1.3 Failure Modes and Effects Analysis. Finally, a standard tool in safety engineering is a *Failure Modes and Effects Analysis* (FMEA), methodical and systematic risk management approach that examines a proposed design or technology for foreseeable failures [72]. The main purpose of a FMEA is to define, identify and eliminate potential failures or problems in different products, designs, systems and services. Prior to conducting a FMEA, known issues with a proposed technology should be thoroughly mapped through a literature review and by collecting and documenting the experiences of the product designers, engineers and managers. Further, the risk exercise is based on known issues with relevant datasets and models, information that can be gathered from interviews and from extant technical documentation.

FMEAs can help designers improve or upgrade their products to reduce risk of failure. They can also help decision makers formulate corresponding preventive measures or improve reactive strategies in the event of post-launch failure. FMEAs are widely used in many fields including aerospace, chemical engineering, design, mechanical engineering and medical devices. To our knowledge, however, the FMEA method has not been applied to examine ethical risks in production-scale artificial intelligence models or products.

3.2 Medical devices

Internal and external quality assurance audits are a daily occurrence in the pharmaceutical and medical device industry. Audit document trails are as important as the drug products and devices themselves. The history of quality assurance audits in medical devices dates from several medical disasters in which devices, such as infusion pumps and autoinjectors, failed or were used improperly [80].

3.2.1 Design Controls. For medical devices, the stages of product development are strictly defined. In fact, federal law (Code of Federal Regulations Title 21) mandates that medical-device makers establish and maintain "design control" procedures to ensure that design requirements are met and designs and development processes are auditable. Practically speaking, design controls are a documented method of ensuring that the end product matches the intended use, and that potential risks from using the technology have been anticipated and mitigated [77]. The purpose is to ensure that anticipated risks related to the use of technology are driven down to the lowest degree that is reasonably practicable.

3.2.2 Intended Use. Medical-device makers must maintain procedures to ensure that design requirements meet the "intended use" of the device. The intended use of a "device" (or, increasingly in medicine, an algorithm—see [60] for more) determines the level of design control required: for example, a tongue depressor (a simple piece of wood) is the lowest class of risk (Class I), while a deep brain implant would be the highest (Class III). The intended use of a tongue depressor could be "to displace the tongue to facilitate examination of the surrounding organs and tissues", differentiating a tongue depressor from a Popsicle stick. This may be important when considering an algorithm that can be used to identify cats or to identify tumors; depending on its intended use, the same algorithm might have drastically different risk profiles, and additional risks arise from unintended uses of the technology.

3.2.3 Design History File. For products classified as medical devices, at every stage of the development process, device makers must document the design input, output, review, verification, validation, transfer and changes—the design control process (section 3.2.1). Evidence that medical device designers and manufacturers have followed design controls must be kept in a design history file (DHF), which must be an accurate representation and documentation of the product and its development process. Included in the DHF is an extensive risk assessment and hazard analysis, which must be continuously updated as new risks are discovered. Companies also proactively maintain "post-market surveillance" for any issues that may arise with safety of a medical device.

3.2.4 Structural Vulnerability. In medicine there is a deep acknowledgement of socially determinant factors in healthcare access and effectiveness, and an awareness of the social biases influencing the dynamic of prescriptions and treatments. This widespread acknowledgement led to the framework of operationalizing structural vulnerability in healthcare contexts, and effectively the design of an assessment tool to record the anticipated social conditions surrounding a particular remedy or medical recommendation [61]. Artificial intelligence models are equally subject to social influence and social impact, and undergoing such assessments on more holistic and population- or environment-based considerations is relevant to algorithmic auditing.

3.3 Finance

As automated accounting systems started to appear in the 1950s, corporate auditors continued to rely on manual procedures to audit "around the computer". In the 1970s, the Equity Funding Corporation scandal and the passage of the Foreign Corrupt Practices Act spurred companies to more thoroughly integrate internal controls throughout their accounting systems. This heightened the need to audit these systems directly. The 2002 Sarbanes-Oxley Act introduced sweeping changes to the profession in demanding greater focus on financial reporting and fraud detection [10].

Financial auditing had to play catch-up as the complexity and automation of financial business practices became too unwieldy to manage manually. Stakeholders in large companies and government

regulators desired a way to hold companies accountable. Concerns among regulators and shareholders that the managers in large financial firms would squander profits from newly created financial instruments prompted the development of financial audits [74].

Additionally, as financial transactions and markets became more automated, abstract and opaque, threats to social and economic values were answered increasingly with audits. But financial auditing lagged behind the process of technology-enabled financialization of markets and firms.

3.3.1 Audit Infrastructure. In general, internal financial audits seek assurance that the organization has a formal governance process that is operating as intended: values and goals are established and communicated, the accomplishment of goals is monitored, accountability is ensured and values are preserved. Further, internal audits seek to find out whether significant risks within the organization are being managed and controlled to an acceptable level [71].

Internal financial auditors typically have unfettered access to necessary information, people, records and outsourced operations across the organization. IIA Performance Standard 2300, Performing the Engagement [55], states that internal auditors should identify, analyze, evaluate and record sufficient information to achieve the audit objectives. The head of internal audit determines how internal auditors carry out their work and the level of evidence required to support their conclusions.

3.4 Discussion and Challenges

The lessons from other industries above are a useful guide toward building internal accountability to society as a stakeholder. Yet, there are many novel and unique aspects of artificial intelligence development that present urgent research challenges to overcome.

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.[2]

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. Internal audits can also leverage relevant experts from within the company to facilitate such discussions and provide additional insight on potential risks [3].

Large-scale production AI systems are extraordinarily complex, and a critical line of future research relates to addressing the interaction of highly complex coupled sociotechnical systems. 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].

With artificial intelligence systems it can be difficult to trace model output back to requirements because these may not be explicitly documented, and issues may only become apparent once systems are released. However, from an ethical and moral perspective it is incumbent on producers of artificial intelligence systems to anticipate ethics-related failures before launch. However, as [58] and [31] point out, the design, prototyping and maintenance of AI systems raises many unique challenges not commonly faced with other kinds of intelligent systems or computing systems more broadly. For example, *data entanglement* results from the fact that artificial intelligence is a tool that mixes data sources together. As Scully et al. point out, artificial intelligence models create entanglement and make the isolation of improvements effectively impossible [67], which they call *Change Anything Change Everything*. We suggest that by having explicit documentation about the purpose, data, and model space, potential hazards 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]. As a relatively young community focused on fairness, accountability, and transparency in AI, we have some indication of the system culture requirements needed to normalize, for example, an adequately thorough documentation procedure and guidelines [24, 48]. Still, we lack the formalization of a standard model development template or practice, or process guidelines for when and in which contexts it is appropriate to implement certain recommendations. In these cases, internal auditors can work with engineering teams to construct the missing documentation to assess practices against the scope of the audit. Improving documentation can then be a remediation for future work.

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.

We thus combine learnings from prior practice in adjacent industries while recognizing the uniqueness of the commercial AI industry to identify key opportunities for internal auditing in our specific context. We do so in a way that is appropriate to the requirements of an AI system.

4 SMACTR: AN INTERNAL AUDIT FRAMEWORK

We 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.

To illustrate the utility of this framework, we contextualize our descriptions with the hypothetical example of Company X Inc.,

[2] https://deloitte.wsj.com/riskandcompliance/2018/08/06/mind-over-matter-implementing-agile-internal-audit/

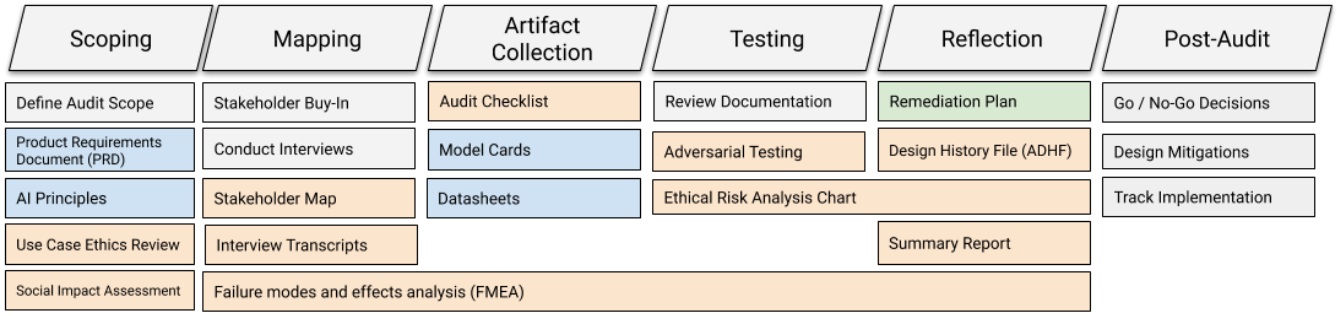

Figure 2: Overview of Internal Audit Framework. Gray indicates a process, and the colored sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.

a large multinational software engineering consulting firm, specializing in developing custom AI solutions for a diverse range of clients. We 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, Fariness & Non-Discrimination", "Safety & Non-Maleficence", "Responsibility & Accountability" and "Privacy". We also assume that the corporate structure of Company X is typical of any technical consultancy, and design our stakeholder map by this assumption.

Company X has decided to pilot the SMACTR internal audit framework to fulfill a corporate mandate towards responsible innovation practice, accommodate external accountability and operationalize internal consistency with respect to its identified AI principles. The fictional company thus pilots the 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 we contextualize the model within the setting of the product and deployment environment.

An end-to-end worked example of the audit framework is available as supplementary material to this paper for the Happy-Go-Lucky, Inc. client case. This includes demonstrative templates of all recommended documentation, with the exception of 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 resource here.

4.1 The Governance Process

To design our audit procedure, we suggest complementing formal risk assessment methodologies 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 artificial intelligence 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 sociotechnical system in which a product will be deployed [50]. There are also intersections between this framework and just effective product development theory [5], as many of the components of audit design refocus the product development process to prioritize the user and their ultimate well-being, resulting in a more effective product performance outcome.

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 we refer to as a "transparency trail" of documentation at each step of the development cycle (see Figure 2). To shift the process into an actionable mechanism for accountability, we present a validated and transparently outlined procedure that auditors can commit to. The thoroughness of our described process will hopefully engage the trust of audit targets to act on and acknowledge post-audit recommendations for engineering practices in alignment with prescribed AI principles.

This process primarily addresses how to conduct internal audits, providing guidance for those that have already deemed an audit necessary but would like to further define the scope and execution details. Though not covered here, an equally important process is determining what systems to audit and why. 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. Risk prioritization and the necessary variance in scrutiny is a separately interesting and rich research topic on its own. The process outlined below can be applied in full or in a lighter-weight formulation, depending on the level of assessment desired.

4.2 The Scoping Stage

For both clients, a product or request document is provided to specify the requirements and expectations of the product or feature. 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. The goal is to anticipate areas to investigate as potential sources of harm and social impact. At this stage, interaction with the system should be minimal.

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. The product team should also provide a Product Requirements Document (PRD), or project proposal from the initial planning of the audited product.

4.2.1 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.

Importantly, we stress standpoint diversity in this process. **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]. Another method to elicit implicit biases or motivated cognition [40] is to ask people to reflect on their preliminary assessment and then ask whether they might have reason to regret the action later on. This can shed light on how our position in society biases our assumptions and ways of knowing [18].

An internal ethics review board that includes a diversity of voices should review proposed projects and document its views. Internal ethics review boards are common in biomedical research, and the purpose of these boards is to ensure that the rights, safety, and well-being of all human subjects involved in medical research are protected [56]. Similarly, the purpose of an ethics review board for AI systems includes safeguarding human rights, safety, and well-being of those potentially impacted.

4.2.2 Artifact: Social Impact Assessment. A social impact assessment should inform the ethical review. 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]. In it, we describe 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. The severity of risk is the degree to which the specific context of the use case is assessed to determine the degree in which potential harms may be amplified. The severity assessment proceeds from the analysis of impacts and harms to give a sense of the relative severity of the harms and impacts depending on the sensitivity, constraints, and context of the use case.

4.3 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. At this stage, the FMEA (Section 3.1.3) should begin and risks should be prioritized for later testing.

As Company X is a consultancy, this stage mainly requires identifying the stakeholders across product and engineering teams anchored to this particular client project, and recording the nature of their involvement and contribution. This enables an internal record of individual accountability with respect to participation towards the final outcome, and enables the trace of relevant contacts for future inquiry.

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. Additionally, this stage includes a design history file review of all existing documentation of the development process or historical artifacts on past versions of the product. Finally, it includes a report or interview transcripts on key findings from internal ethnographic fieldwork involving the stakeholders and engineers.

4.3.1 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.

4.3.2 Artifact: Ethnographic Field Study. As Leveson points out, bottom-up decentralized decision making can lead to failures in complex sociotechnical 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. With modern large-scale artificial intelligence projects and API development, it can be difficult to gain a shared understanding at the right level of system description to understand how local decisions, such as the choice of dataset or model architecture, will impact final system behavior.

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. As in internal financial auditing, access to key people in the organization is important. This access involves semi-structured interviews with a range of individuals close to the development process and documentation gathering to gain an understanding of possible gaps that need to be examined more closely.

Traditional metrics for artificial intelligence like loss may conceal fairness concerns, social impact risks or abstraction errors [69]. A key challenge is to assess how the numerical metrics specified in the design of an artificial intelligence system reflect or conform with these values. Metrics and measurement are important parts of the auditing process, but should not become aims and ends in themselves when weighing whether an algorithmic system under audit is ethically acceptable for release. 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 as we posit here, an abstraction error." What we consider data is already an interpretation, highly subjective and contested [23]. Metrics must be understood in relation to the engineering context in which they were developed and the social context into which they will be deployed. 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.

4.4 The Artifact Collection Stage

Note that the collection of these artifacts advances adherence to the declared AI principles of the organization on "Responsibility & Accountability" and "Transparency".

In this stage, we identify and collect all the required documentation from the product development process, 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.

At times documentation can be distributed across different teams and stakeholders, or is missing altogether. In certain cases, the auditor is in a position to enforce retroactive documentation requirements on the product team, or craft documents themselves.

The model card for the smile detection model is the template model card from the original paper [48]. A hypothetical datasheet for this system is filled out using studies on the CelebA dataset, with which the smile detector is built [44, 47]. In the model card, we identify potential for misuse if smiling is confused for positive affect. From the datasheet for the CelebA dataset, we see that although the provided binary gender labels seem balanced for this dataset (58.1% female, 42% male), other demographic details are quite skewed (77.8% aged 0-45, 22.1% aged over 46 and 14.2% lighter-skinned, 85.8% darker-skinned)[47].

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. Those pre-requisites can include model and data transparency documentation.

4.4.1 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. It ensures that the full scope of expected product processes and that the corresponding documentation required to be completed before the audit review can begin are finished. This is also a procedural evaluation of the development process for the system, to ensure that appropriate actions were pursued throughout system development ahead of the evaluation of the final system outcome.

4.4.2 Artifacts: Datasheets and Model Cards. Two recent standards can be leveraged to create auditable documentation, model cards and datasheets [24, 48]. Both model cards and datasheets are important tools toward making algorithmic development and the algorithms themselves more auditable, with the aim of anticipating risks and harms with using artificial intelligence systems. Ideally, these artifacts should be developed and/or collected by product stakeholders during the course of system development.

To clarify the intended use cases of artificial intelligence 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 sociotechnical logics and processes. A robust model card is key to documenting the intended use of the model as well as information about the evaluation data, model scope and risks, and what might be affecting model performance.

Model cards are intended to complement "Datasheets for Datasets" [24]. Datasheets for machine learning datasets are derived by analogy from the electronics hardware industry, where a datasheet for an electronics component describes its operating characteristics, test results, and recommended uses. A critical part of the datasheet covers the data collection process. This set of questions are intended to provide consumers of the dataset with the information they need 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.

4.5 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 engage with the system in various ways, and produce a series of artifacts to demonstrate the performance of the analyzed system at the time of the audit. Additionally, 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. Testing should be based on a risk prioritization from the FMEA.

For the smile detector, we might employ counterfactual adversarial examples designed to confuse the model and find problematic failure modes derived from the FMEA. For the child prediction model, we test performance on a selection of diverse user profiles. 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 SES.

For the ethical risk analysis chart, we look at the principles and realize that there are immediate risks to the "Privacy" principle—with one case involving juvenile data, which is sensitive, and the other involving face data, a biometric. This is also when it becomes clear that in the smiling booth case, there is 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.

4.5.1 Artifact: Adversarial Testing . Adversarial testing is a common approach to finding vulnerabilities in both pre-release and post-launch technology, for example in privacy and security testing [6]. 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], but these approaches can be complementary. This approach is more generally defined—encompassing a range of input options to try in an active attempt to fool the system and incite identified failure modes from the FMEA.

Internal adversarial testing prior to launch can reveal unexpected product failures before they can impact the real world. Additionally, proactive adversarial testing of already-launched products can be a best practice for lifecycle management of released systems. The FMEA should be updated with these results, and the relative changes to risks assessed.

4.5.2 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.

4.6 The Reflection Stage

This phase of the audit is the more reflective stage, when the results of the tests at the execution stage are analyzed in juxtaposition with the ethical expectations clarified in the audit scoping. Auditors update and formalize the final risk analysis in the context of test results, outlining specific principles that may be jeopardized by the AI system upon deployment. This phase will reflect on product decisions and design recommendations that could be made following the audit results.

Additionally, key artifacts at this stage may include a mitigation plan or action plan, jointly developed by the audit and engineering teams, that outlines prioritized risks and test failures that the engineering team is in a position to mitigate for future deployments or for a future version of the audited system.

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 in CelebA to the training data. It could be decided that the use case does not necessarily define affect, but treats smiling as a favourable 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. Once these conditions are met, Company X could be confident to greenlight developing this product for the client.

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.

4.6.1 Artifact: Algorithmic Use-related Risk Analysis and FMEA. The risk analysis should be informed by the social impact assessment and known issues with similar models. Following Leveson's work on safety engineering [42], we stress that 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 artificial intelligence 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. Reasonable and foreseeable misuse of the model should be anticipated by the designer. Therefore, the *user's mental model* of the system should be anticipated and taken into consideration. Large gaps between the *intended* and *actual* uses of algorithms have been found in contexts such as criminal justice and web journalism [12].

This adds complexity to anticipated hazards and risks, nevertheless these should be documented where possible. 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". We propose that known use-related issues with deployed systems be taken into account during the design stage. The format of the risk analysis can be variable depending on context, and there are many valuable templates to be found in *Failure Modes and Effects Analysis* (Section 3.1.3) framing and other risk analysis tools in finance and medical deployments.

4.6.2 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. This plan can be reviewed by the audit team and leadership to better inform deployment decisions.

For the concerns raised in any audit against ethical values, a technical team will 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? In safety engineering, a risk threshold is usually defined under which the risk is considered tolerable. Though a challenging problem, similar standards could be established and developed in the ethics space as well.

4.6.3 Artifact: Algorithmic Design History File. Inspired by the concept of the design history file from the medical device industry [77], we propose an algorithmic design history file (ADHF) which would collect all the documentation from the activities outlined above related to the development of the algorithm. It should 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.

This design history file would form the basis of the final audit report, which is a written evaluation by the organization's audit team. The ADHF should assist with an audit trail, enabling the reconstruction of key decisions and events during the development of the product. The algorithmic report would then be a distillation and summary of the ADHF.

4.6.4 Artifact: Algorithmic Audit Summary Report. The report aggregates all key audit artifacts, technical analyses and documentation, putting this in one accessible location for review. This audit report should be compared qualitatively and quantitatively to the expectations outlined in the given ethical objectives and any corresponding engineering requirements.

5 LIMITATIONS OF INTERNAL AUDITS

Internal auditors necessarily share an organizational interest with the target of the audit. While it is important to maintain an independent and objective viewpoint during the execution of an audit, we awknowledge that this is challenging. The audit is never isolated from the practices and people conducting the audit, just as artificial intelligence systems are not independent of their developers or of the larger sociotechnical 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. Although long-standing internal auditing practices for quality assurance in the financial, aviation, chemical, food, and pharmaceutical industries have been shown to be an effective means of controlling risk in these industries [76], the regulatory dynamics in these industries suggest that internal audits are only one important aspect of a broader system of required quality checks and balances.

6 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 artificial intelligence 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-the-loop 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.

[61] James Quesada, Laurie Kain Hart, and Philippe Bourgois. 2011. Structural vulnerability and health: Latino migrant laborers in the United States. *Medical anthropology* 30, 4 (2011), 339–362.

[62] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society*.

[63] Clarence Rodrigues and Stephen Cusick. 2011. *Commercial aviation safety 5/e*. McGraw Hill Professional.

[64] G Sirgo Rodríguez, M Olona Cabases, MC Martin Delgado, F Esteban Reboll, A Pobo Peris, M Bodí Saera, et al. 2014. Audits in real time for safety in critical care: definition and pilot study. *Medicina intensiva* 38, 8 (2014), 473–482.

[65] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22 (2014).

[66] David Satava, Cam Caldwell, and Linda Richards. 2006. Ethics and the auditing culture: Rethinking the foundation of accounting and auditing. *Journal of Business Ethics* 64, 3 (2006), 271–284.

[67] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, and Michael Young. 2014. Machine learning: The high interest credit card of technical debt. (2014).

[68] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.

[69] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 59–68.

[70] Hetan Shah. 2018. Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 20170362.

[71] Dominic SB Soh and Nonna Martinov-Bennie. 2011. The internal audit function: Perceptions of internal audit roles, effectiveness and evaluation. *Managerial Auditing Journal* 26, 7 (2011), 605–622.

[72] Diomidis H Stamatis. 2003. *Failure mode and effect analysis: FMEA from theory to execution*. ASQ Quality press.

[73] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a framework for responsible innovation. *Research Policy* 42, 9 (2013), 1568–1580.

[74] Alexander Styhre. 2015. *The financialization of the firm: Managerial and social implications*. Edward Elgar Publishing.

[75] Alexander Styhre. 2018. The unfinished business of governance: towards new governance regimes. In *The Unfinished Business of Governance*. Edward Elgar Publishing.

[76] JohnK Taylor. 2018. *Quality assurance of chemical measurements*. Routledge.

[77] Marie B Teixeira, Marie Teixeira, and Richard Bradley. 2013. *Design controls for the medical device industry*. CRC press.

[78] Manuel Trajtenberg. 2018. *AI as the next GPT: a Political-Economy Perspective*. Technical Report. National Bureau of Economic Research.

[79] Frank Vanclay. 2003. International principles for social impact assessment. *Impact assessment and project appraisal* 21, 1 (2003), 5–12.

[80] Tim Vanderveen. 2005. Averting highest-risk errors is first priority. *Patient Safety and Quality Healthcare* 2 (2005), 16–21.

[81] Ajit Kumar Verma, Srividya Ajit, Durga Rao Karanki, et al. 2010. *Reliability and safety engineering*. Vol. 43. Springer.

[82] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, and Stephen Cave. 2019. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA*. 27–28.

[83] Yi Zeng, Enmeng Lu, and Cunqing Huangfu. 2018. Linking Artificial Intelligence Principles. *arXiv preprint arXiv:1812.04814* (2018).