AI Ethics Audit Best Practices Review

Written in Fall 2022
at San Francisco State University
by Darshil Dhameliya

Introduction

Modern AI algorithms are increasingly being deployed in a wide range of domains affecting our society in ways that are sometimes only apparent after deployment. AI models can be used in many settings from high-stakes medical applications to relatively harmless applications like music recommendation algorithms. With the possibility of these AI models resulting in unwanted side-effects, e.g., bias in algorithmic criminal risk assessment [4], it is crucial that such AI models are developed with rigorous auditing procedures to make sure they follow the developer organization's ethical principles.

Due to the recency of such expansion of AI models' usage in various domains, little research efforts are poured into ensuring that these models result in more expected upside than downside. It also becomes essential, then, that the organizations or teams within are held accountable when these models create high-risk situations. Standard auditing mandates should allow organizations to reduce such accountability gaps which also requires increased transparency to the intermediate and end users. Transparency also helps organizations in adhering to regulatory requirements as this information is readily accessible.

The following report is an attempt in drawing lessons from [1], [2], and [3] to identify the building blocks and broader pointers that shall help build an auditing procedure to address such high-stakes issues in both pre- and post-deployment scenarios. The report starts with a more detailed background primarily following [2] and [3]. Then, we propose a domain-agnostic framework that can aid in auditing any AI algorithm. We also try to explain the importance of explainability, transparency, and data provenance to reinforce the necessity of auditing AI algorithms. A list of all the references is available at the end of the report.

1. Background

In the 21st century, AI models have recently shown promising results in performing a wide range of tasks including playing complex multiplayer games like Dota 2 [5], or in predicting protein folding patterns which is traditionally very costly and time intensive [6]. Some applications involving high-impact tasks which are likely to have drastic societal implications need to be sufficiently evaluated while being developed and after deployment. Such high-impact algorithms are also likely to implicitly amplify social biases.

To hedge ourselves against such unwanted downsides, some regulations have already been proposed by government bodies like the EU [7]. Organizations still need to have a more detailed outlook of how the cutting-edge AI algorithms they develop might affect society and be aware of how to reduce the misusage of such algorithms. Thus, internal auditing while product development becomes essential for organizations so that the origins of hazards are easily identified when something goes wrong.

Paper [2] draws lessons from other industries like aerospace, finance, and medical devices where audit mandates are already in place. The authors take these learnings and combine them with standard software engineering practices also considering implications of AI disregarded by most of the prevailing practices in AI development to propose an auditing framework that can be used during the production cycle. Such audits shall allow regulators and organizations to identify interaction failures in these complex processes, monitor adverse outcomes and the possibility of wide-scale failure, and identify intervention strategies when needed.

Even though many organizations are aware of the ethical standards of how an AI should work, it is very hard to translate these ideals into practice. The authors of [2] try to outline an audit plan that would help convert these principles into practice by focusing on risk analysis based on failures of algorithms. It also becomes essential that the internal auditors and the developers adhere to audit integrity and try to follow ethical standards whether they are organizational or personal ethical stances. This integrity, in turn, makes the audit outcomes more legitimate when an algorithm needs to be scrutinized.

Aside from easily complying with external audit standards, an internal audit allows the organization to fully audit the processes as some of the proprietary assets might not be suited to be exposed to a general audience. Since external auditors don't have access to all the internal models and data, rigorous internal auditing allows organizations to incorporate significantly more security in the models as they would proactively look for all the loopholes instead of just reactively complying with external standards. Internal standards also make it seamless to include other organization members to audit such systems.

We will now try to outline some of the safety standards from adjacent industries: aerospace and medical devices. The aerospace industry has placed very rigorous auditing requirements for aviation devices, these standards have made airplanes very safe, and we see very few failure incidents in the industry. In the medical industry, auditing devices or drugs is very important as they can have a drastic positive or negative impact on the users. Which is why federal regulations are very rigorous around this industry.

People working in aerospace use *checklists* to help designers to be aware of important questions, edge cases, and failures. They use *traceability* in terms of building the broader context of how product requirements originated based on the sources and how it is translated into design. As described in [2], *Failure Modes and Effects Analysis (FMEA)* is done "to define, identify and eliminate potential failures or problems in different products, designs, systems, and services." FMEA is primarily derived from literature review and interviewing all the organization members from designers to managers.

The medical industry has strictly defined product development stages. Such *design controls* are designed to make sure that designs and development processes are auditable so that the intended use and potential risks of the technology are anticipated and potentially mitigated. *Intended use* definition helps understand the risks in adjacent unintended use cases which could be developed by a third party, maybe a competitor. *Design history files*

(DHF) standards require the device makers to document design control processes at every stage of the development processes, also including extensive risk assessment and hazard analysis. *Structural vulnerability* analysis helps understand sociotechnical factors that might result in social biases, e.g., testing a device on people of a wide variety of genetic backgrounds.

Many of these lessons can be translated into the AI industry with slight modifications. The authors of [2] propose a *SMACTR* internal auditing framework in the paper. *SMACTR* stands for Scoping, Mapping, Artifact Collection, Testing, and Reflection. The authors of [3] developed a framework called "model cards" to encourage transparent model reporting. The model cards fit very nicely in the *SMACTR* framework. Before moving into the next section of the report, which focuses mostly on the *SMACTR* framework to define an audit plan, the rest of this section gives a broad overview of model cards as defined in [3].

Model cards succinctly describe benchmarking evaluations in various cultural, demographic, or phenotypic groups. Aside from that, it tries to disclose to the stakeholders the intended use cases, performance evaluations, and other information as outlined below. Model cards should also describe the motivation behind chosen performance metrics, group definitions, and other relevant factors. Similar to the FDA mandates of drug testing disaggregated by groups like age, race, and gender, the authors propose testing models through unitary and intersectional group disaggregation.

The model card as proposed in [3] has nine main sections. The *model details* section should describe the date, type, and version of the model, the developer or developer organization, the kind of model and all the parameters used, the resources used and citation details, licensing information, and contact information for the developers. The *intended use* section describes primary intended users and use cases and identifies potential out-of-scope use cases. The *factors* section consists of identifying relevant groups, environmental conditions in which the model was developed, technical attributes, and other evaluation factors used.

Furthermore, the *metrics* section would ideally reflect potential real-world impacts of the model describing model performance measures, decision thresholds and variation measurement approaches in building the model. *Evaluation data* gives details on the testing datasets, motivation, and preprocessing techniques used. *Training data* similarly describes the data used in training the model. The *quantitative analysis* section shows the analysis and test results disaggregated by the unitary and intersectional groups performed.

The *ethical considerations* section tries to report the kinds of ethical standards or assumptions considered while developing the model. This can include questions like what kinds of sensitive data were used, whether the model would impact human life directly or indirectly, what are the potential risks and harms and how to mitigate them. Lastly, the *caveats and recommendations* section would describe any extra information not covered by previous sections including assumptions made during the development process, and mention potential recommendations for further developments and experimentations. Combined with the *SMACTR* framework, the model card provides a robust process for internal auditing while production and after deployment.

2. Proposed Audit Plan

Let us now move into defining what a potential internal AI auditing plan might consist of. We will try to be as flexible as possible such that this plan may apply to any kind of AI algorithm. This plan can be further customized based on specific use cases. It would also be great to collaboratively collect feedback to improve the plan and create a standard organizational mechanism that would execute the auditing as efficiently as possible. The auditing process shall direct us towards responsible innovative practices, reducing accountability gaps and making the production cycle more consistent.

Some of the currently known AI principles revolve around "Transparency", "Justice, Fairness & Non-Discrimination", "Safety & Non-Maleficence", "Responsibility & Accountability" and "Privacy. [8]" These principles and other broader moral principles may guide the auditing process in making AI models safer by prioritizing scrutinization of the model interactions causing high-risk failures.

Let us now move into the details of the proposed audit plan inspired by the *SMACTR* framework from [2] which is shown in Image 1 below. The Gray colored blocks indicate processes, all the other colors are documents: documents marked by blue color are produced by product and engineering teams, and orange ones are created by the auditors. The green color shows documents created jointly by them.

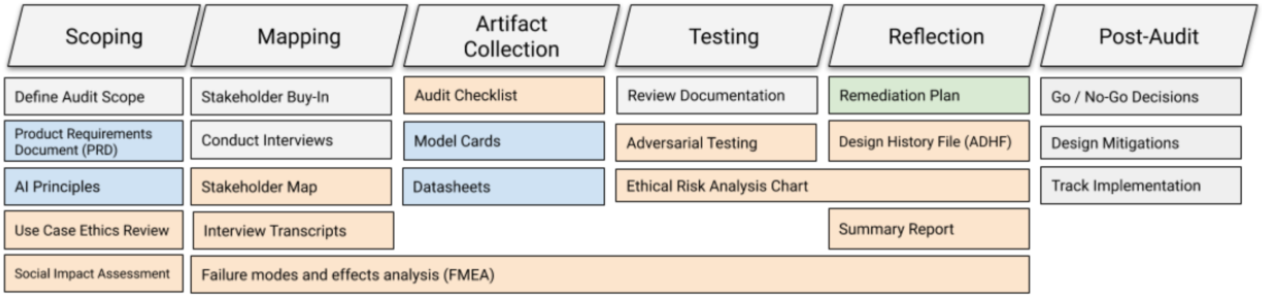

Image 1: Overview of the SMACTR Framework as illustrated in [2].

3.1 Scoping

The scoping stage sets up the expectations of the product including the requirements. It mentions the intended use cases and other adjacent use cases where it might not be appropriate to deploy. This would help us identify the negative social impact the given model can have. Some of the required documents before this stage begins are ethical objectives and AI principles to adhere to when auditing the algorithm. Another such document is a Product Requirement Document (PRD). All these documents are supplied by product and/or engineering teams.

The scoping stage creates two documents or artifacts: The *Ethical Review of the System Use Case* and the *Social Impact Assessment*. The *ethical review* considers various perspectives on who are likely to be impacted and how. As the paper [2] mentions, “Al algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values.” So, it becomes important to consult experts from the domain as well. While in this stage, it is important to know that our personal social stance may impact or bias our thinking.

The *Social Impact Assessment* delineates potential positive or negative impacts on society due to the deployment of the tool being developed. It could include several things like how it can change people’s thinking and behavior, their social interactions, their environments, etc. This stage has two steps: identifying the relevant impact and harms, and the severity of them: maybe tagging them as low, medium, or high impact. Based on the sensitivity and the context of deployment.

3.2 Mapping

This step lays out every detail that is available so far and maps the internal stakeholders and primary actors or collaborators who are auditing. We also need to educate the stakeholders appropriately to set up the right context of the audit for them. This stage involves recording individual accountability towards the final product which would help identify individuals most appropriate for future inquiry. The *Failure Modes and Effects Analysis (FMEA)* will be initiated and will be updated in the later stage as more information is available through auditing for risk analysis and prioritization.

This stage creates two other artifacts. The *Stakeholder Map* outlines stakeholders involved in the audit process and everyone from other teams who collaborated in helping in the audit. It can also be helpful to show how the various members participated in the audit to provide more context to external auditors or stakeholders. The purpose of the *Ethnographic Field Study* is to get a diversified perspective of the social landscape where the application might have a drastic impact.

“Bottom-up decentralized decision-making can lead to failures in complex sociotechnical systems. 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.” [2] [9] The *Ethnographic Field Study* consists of interviewing key individuals related to the development process to understand possible gaps needed to be scrutinized further. Here, It becomes important to understand that traditional performance metrics may not inform us about the bias in the model or social impact risks.

3.3 Artifact Collection

The third stage tries to gather various documents from the development process on things like model dynamics, the data used, design reviews, planning, and such documents from previous production cycles. If such documents are not produced by teams, the auditors should have a right to enforce such documentation from various teams. This stage creates two types of documents: *Design Checklist* and *Datasheets and Model Cards.*

The *Design Checklist* is an audit document that lists of all the standards for designing an AI algorithm. This step also involves making sure the development process adheres to these ethical standards. The *Model Cards* have been discussed in the report previously which helps in understanding the potential risks of a given AI system. The *Datasheets cover* details about the data and how it was obtained to provide more context on how the model might perform on unseen data.

3.4 Testing

This is the stage where most of the testing is conducted. Testing would ideally be done following the FMEA risk prioritization order. This stage provides us with two artifacts. The *Adversarial Training* document tries to identify loopholes for both pre- and post-deployment scenarios through extensive adversarial testing. This may reveal outlier high-risk situations that might occur in the real world. The intersectional and unitary approach to various groups identified earlier would be handy while testing for various demographics. The *Ethical Risk Analysis Chart* is like the *FMEA* which shows the likelihood and severity of possible known failures of the system.

3.5 Reflection

The final stage of the internal audit process tries to understand the test results concerning the ethical standards defined during the audit scope. These analyses would inform us about possible remedies or risk mitigation plans to make the model more appropriate for deployment. It is considered a good practice to ask for user permissions and show disclaimers to the users on how their private data is processed.

This stage finalizes the *FMEA* and *Ethical Risk Analysis* documents. In addition to that, this stage creates three final artifacts for completing the auditing process. Following [9], the *Risk Analysis* and *FMEA* should consider the gap between the designer's mental models of the given model and the user's mental model of how it may affect them. This is because the idealized AI model envisioned by the designer may be drastically disparate from how it behaves after deployment. We should always consider how unseen data may result in intentional or unintentional misusage of the model.

The *Remediation and Risk Mitigation Plan*'s goal is to reduce the potential social impact by proposing remediation techniques. This would be done through a collaboration of all the stakeholders mapped previously. The *Algorithmic Design History File (ADHF)* should gather all the documents related to the development processes for future reference. It is also a good way to check whether the upsides from the model deployment are higher than the potential downsides. It should now be apparent that the model shall only be deployed if the benefits are much higher.

The *Algorithmic Audit Summary Report* is an aggregation of the previously created artifacts and analyses results. This final task would be informed by the ADHF to provide the results of the auditing procedure and compare the results with the ethical expectations of the organizations as identified previously. The whole audit process itself should always be scrutinized regularly to improve it making it more aligned with the organizational values and external regulations. So, the audit planning committee should always be open to critical feedback from stakeholders. The auditing process shall not only proactively mitigate the risks associated but should also be initiated if we identify hazardous behavior after post-deployment.

3. Why Explainability and Transparency Matters

The authors of [2] provide an artifact requirement to include datasheets for data provenance purposes. Data provenance efforts usually consist of understanding and documenting the trail of how the data is processed to identify potential hazards created because of the data. Data provenance through datasheets would also increase transparency toward external stakeholders and regulators. Datasheets can also serve third-party users of the model so they can make better judgments about how the model might be utilized or what the potential risks associated with using such a model are.

Transparency refers to documenting how a model was developed which can help others replicate the model to verify the results of a given model. [3] outlines the concept of Model Cards which are useful in providing the context in which the model was developed and details on its use cases. Furthermore, benchmarking using standard practices can help internal auditors understand adversarial scenarios. These results can be included in the model cards to increase the transparency of the development of an AI model. Furthermore, Transparency can help in the "responsible democratization of machine learning and related artificial intelligence technology." [3]

Something that both [2] and [3] have not shed light on is, how the explainability of AI algorithms can be important in understanding the future performance of a given AI model. The Explainability of a model helps us understand the underlying representation of the given problem created by the algorithm which ultimately helps us understand how the model may behave. Explainability is a much more robust measure of the performance of a model than traditional accuracy or performance measures. Explainability also generalizes well in predicting how models might perform under distributional shifts in real-world data. This implies prioritizing explainability efforts may help us perform risk analysis through a much better lens.

References:

[1] CSC 859 Class Slides by Prof. D. Petkovic, Fall 2022, San Francisco State University.

[2] Raji et al. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2020).* arXiv:2001.00973.

[3] Mitchell et al. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2019*). arXiv:1810.03993v2.

[4] Julia et al. "Machine Bias." https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).

[5] Berner et al. "Dota 2 with large scale deep reinforcement learning." arXiv:1912.06680 (2019).

[6] Jumper *et al.* "Highly accurate protein structure prediction with AlphaFold." *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2.

[7] "Ethics Guidelines for Trustworthy Artificial Intelligence." https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2018).

[8] Jobin et al. "Artificial Intelligence: the global landscape of ethics guidelines." arXiv:1906.11668 (2019).

[9] Nancy Leveson. "Engineering a safer world: Systems thinking applied to safety." (2011) MIT Press.