AI AUDIT

Aigos AI Security Blueprint Series

aigos

Securedbyaigos.com

Contents

Introduction

What is AI Audit?

AI audit refers to the evaluation of AI systems to ensure they work securely, without bias or discrimination, and are aligned with ethical and legal standards. In essence, it is a methodical examination of an entire AI system, encompassing algorithms, data sources, model training, and deployment protocols. The overarching objective of an AI audit is to validate the functionality of AI applications, ensuring that they not only meet performance benchmarks but also adhere to fundamental principles of fairness and transparency. By scrutinizing the inner workings of AI systems, organizations can proactively identify and address potential vulnerabilities, thereby fostering a culture of responsible and ethical AI deployment. This process extends beyond mere technical assessment, incorporating legal compliance and ethical considerations to guarantee that AI technologies contribute positively to organizational goals while upholding the highest standards of integrity and accountability.

Reasons for AI Audit

Companies undertake AI audits for a spectrum of reasons, encompassing not only regulatory compliance but also a strategic pursuit of sustainable competitive advantages. In the realm of compliance, AI audits serve as a proactive measure to align with evolving legal frameworks and ethical standards, guarding against potential pitfalls and ensuring robust security protocols. This mirrors established practices in cybersecurity, where standards like SOC-2 and ISO 27001 are embraced for similar reasons. However, astute organizations recognize that the significance of AI audits extends beyond mere regulatory adherence. Whether integrating AI systems internally to bolster productivity or offering AI solutions as services, these forward-thinking entities perceive audits as a pivotal step toward optimizing algorithms, fostering transparency, and building trust. In essence, akin to established standards in cybersecurity, AI audits become a strategic investment, positioning companies as pioneers in responsible AI deployment and fostering a competitive edge in an ever-evolving technological landscape.

Beyond Security

Contrary to common misconceptions, AI audits extend far beyond the realm of security concerns. While security is undeniably crucial, these audits also delve into legal compliance, ensuring that AI applications adhere to relevant regulations and ethical standards. Moreover, AI audits scrutinize the reliability of AI models, assessing their performance across various scenarios and identifying potential biases. The scope further extends to the ethical implications of AI applications, emphasizing responsible and fair use. By embracing a comprehensive approach, AI audits become a powerful tool to enhance not only security but also legal adherence, reliability, and ethical considerations.

Who is AI Audit for?

AI audits transcend the exclusive realm of technology experts, catering to a diverse range of stakeholders, from C-suite executives to legal teams, data scientists, and developers, the insights gleaned from AI audits hold particular importance for these innovative entities.

- For startups and service providers, the audit process becomes a strategic tool, aligning their AI applications with business objectives and ensuring responsible, secure, and unbiased AI deployment.
- For larger enterprises, AI audits serve as an internal check and balances mechanism for in-house system builds, reinforcing the reliability and ethical use of AI. Moreover, these audits seamlessly integrate into the standard vendor security assessment process, becoming a crucial component when procuring AI-related services and capabilities externally.

This dual perspective underscores the comprehensive role AI audits play in both internal governance and external partnerships, fostering a culture of responsible AI innovation across the entire organizational spectrum.

Components of AI Audit

We now explore the different components of a comprehensive AI audit, including real world examples of how the audit process either helps to mitigate risk, improve AI performance, adheres to existing standards or improves transparency.

System Security

It is crucial to recognize that AI system security is fundamentally different from traditional application security. This is primarily because of the dynamic nature of how AI models work, which is different from the algorithmic and static nature of traditional software applications.

A comprehensive security audit assesses AI-specific risks, such as: prompt injection, model poisoning, how vector embeddings are secured, model biases that may expose backdoors or supply chain vulnerabilities given the models used. Real-world examples illustrate the significance of system security audits; for instance, a financial institution might conduct an AI security audit to safeguard customer data from cyber threats. By implementing security measures and regularly auditing them, organizations can fortify their AI systems against potential breaches, instilling confidence in users and stakeholders.

Data Protection and Privacy Compliance

Data protection and privacy compliance are critical components of an AI audit, especially in the era of stringent data regulations. The audit assesses how the AI system handles, stores, and processes sensitive user information, ensuring alignment with privacy laws like GDPR or HIPAA. For instance, a healthcare AI application must undergo a meticulous audit to guarantee patient data confidentiality. By prioritizing data protection, organizations not only mitigate legal risks but also build trust with users. The audit process helps identify and rectify potential privacy breaches, fostering a secure environment for the responsible use of data in AI applications.

Fairness and Biases

The fairness and bias component of an AI audit focuses on evaluating whether the AI system treats all individuals fairly and without bias. Through a systematic examination, auditors analyze training data, model outputs, and decision-making processes to identify and rectify biases that may disproportionately impact certain groups. Real-world applications include recruitment AI tools, where audits help uncover and rectify biases that might perpetuate gender or ethnic disparities. By addressing fairness concerns, organizations not only uphold ethical standards but also enhance the inclusivity and trustworthiness of their AI systems, promoting equitable outcomes for diverse user groups.

Accuracy and Reliability

In an AI audit, accuracy and reliability are paramount considerations, assessing the precision and dependability of the AI model in its intended context. Auditors meticulously scrutinize how well the model aligns with its specified purpose, examining key performance metrics such as precision, recall, and overall predictive accuracy. In a financial investment use case, where decisions are crucial and can have significant financial implications, accuracy and reliability audits are indispensable. For instance, a wealth management AI undergoes rigorous scrutiny to ensure that investment recommendations are based on accurate predictions. By prioritizing regular audits and fine-tuning for accuracy in the financial domain, organizations mitigate the risk of relying on flawed AI outputs, safeguarding against potential financial losses and maintaining the trust of investors and stakeholders.

Transparency and Explainability

Transparency and explainability are crucial components of an AI audit, ensuring that the inner workings of the AI model are understandable to stakeholders. Auditors assess whether the decision-making process of the AI system can be explained in a clear and interpretable manner. Real-world examples include loan approval AI, where transparency is essential for users to comprehend why a loan application was accepted or rejected. By prioritizing transparency and explainability, organizations not only meet ethical standards but also build user trust by demystifying the often complex processes behind AI decision-making.

Intellectual Property and Confidential Data Protection

Intellectual property (IP) considerations in an AI audit involves two distinct aspects.

First, when organizations share valuable IP or confidential data, such as their code base for software, it is imperative to ensure that there is no IP leakage that could compromise the company's competitive edge. Auditors scrutinize the mechanisms in place to safeguard proprietary algorithms, training datasets, and other sensitive information, ensuring that the underlying system processes are secure and do not inadvertently expose critical intellectual property or confidential data.

Second, when organizations utilize publicly sourced data for training or leverage open-source foundation models, it is crucial to verify that they are not indirectly violating copyright laws or infringing on copyrighted content. This becomes especially pertinent in the context of retrieval-augmented generative AI systems. Auditors examine the organization's practices to ensure compliance with copyright regulations, verifying that the use of open-source components aligns with licensing agreements and does not introduce legal risks. By

addressing both aspects of intellectual property protection, organizations can foster innovation, collaboration, and responsible AI development while mitigating the potential legal consequences associated with IP infringement.

Ethics and Reputation

The ethics and reputation component of an AI audit involves evaluating the ethical implications of the AI system's behaviour and its potential impact on the organization's reputation. Auditors assess whether the AI aligns with ethical standards and societal norms. For instance, a social media platform might undergo an ethics audit to ensure that its recommendation algorithms promote responsible content. By incorporating ethical considerations into the audit process, organizations not only avoid negative publicity but also contribute to the responsible development and deployment of AI technologies.

Licensing, Legal and Regulatory Compliance

In the context of an AI audit, legal and regulatory compliance encompasses adherence to both overarching laws and industry-specific regulations. One crucial aspect of legal compliance involves scrutinizing licensing provisions, particularly concerning open-source models. This is especially pertinent when enterprises collaborate with AI vendors. The audit process is essential for confirming that models designated for non-commercial use are not repurposed for commercial applications, preventing potential legal repercussions. For instance, when integrating open-source foundation models into proprietary AI systems, it is imperative to respect the stipulations outlined in the licensing agreements. By instituting robust audit processes, organizations can navigate the complex landscape of legal and regulatory compliance, fostering responsible AI deployment and mitigating the risks associated with the misuse of open-source models for commercial purposes. This approach not only safeguards against legal challenges but also reinforces the lawful use of AI technologies in diverse contexts.

Policies, Processes and Controls

Finally, policies, processes and controls need to be in place to ensure that the above standards are upheld on an ongoing basis. For example:

Policies - Organizations need to establish policies on matters such as their stance for web scraped content (legal), whether they allow bringing in PII data into the enterprise data fabric (legal, privacy), how they plan to evaluate AI generated output (biases, accuracy) or what encryption or guardrails need to be in place (system security, data protection).

Processes and controls – The associated processes and controls will then need to be put in place to ensure that policies are adhered to. For example, vendor security assessment questionnaires should be updated to take into consideration the distinct and unique nature of AI system security; model guardrails will need to be built into the systems; independent teams will need to be assigned the responsibility for assessing model output etc.

AI Audit Implementation

Below, we provide a summarized outline of Aigos' AI Audit guidelines for implementation purposes.

Our approach recognizes the different organizational context for firms using AI internally (e.g. an inhouse AI-based resume screening tool), versus that of AI service providers (e.g. a company offering AI-based solution to consumers/enterprises). We define base standards as those that applies to all organizational context, even when AI systems are only used internally. Enhanced Standards on the other hand applies primarily for firms offering AI systems/services externally and warrants a higher bar.

As this is an emerging field, our own assessment of the state of the market is that current ISO and SOC standards do not adequately address the breadth of AI Audit requirements. We believe however that (i) international standard setting bodies will eventually catch on in 2024/2025 and that (ii) local government agencies and industry bodies will also begin articulating requirements that are pertinent to the context of each region or industry.

Aigos' AI Audit guidelines should therefore be seen as an open-source work-in-progress that can be further built upon.

Type	Description	Base	Enhanced
System Security (specific to AI systems)			
Policy	Establish and document **security guidelines specific to AI system risks**	●	●
Policy	Maintain a **ledger of libraries, packages and models used**, including their sources.	●	●
Policy	Document **privileges and resources directly accessible** or callable by AI systems.	●	●
Controls	Implement remote system **backups for AI model weights.**	●	●
Controls	Implement remote system **backups for vector databases.**	●	●
Controls	Implement **logging** for user prompts and model outputs.	●	●
Controls	Implement guardrails for mitigating against prompt injection. Guardrails should adequately address range user input modalities.	●	●
Controls	Implement controls to **ring fence system resources** directly accessible by AI systems.	●	●
Controls	System design should ensure that **user inputs and upload** do not directly flow into any data ingestion pipelines.	●	●
Controls	System design should ensure that **publicly sourced data** are scanned before flowing into any data ingestion pipelines.	●	●
Controls	Regularly **scan models** for embedded backdoors to mitigate risk of triggering specific outputs.		●
Controls	Implement **encryption for vector embeddings** within databases.		●
Controls	Implement mechanism in to mitigate or deter against **shadow model reconstruction.**		●
Data Protection and Privacy Compliance			
Policy	Put in place policy for organizational **stance on ingestion and storage of copyrighted content** within AI models or vector databases	●	●
Policy	Put in place policy for organizational **stance on ingestion and storage of PII data** within AI models or vector databases	●	●
Policy	Put in place policy for organizational **stance on ingestion and storage of confidential data** within AI models or vector databases	●	●
Policy	Put in place policy for organizational stance on sending **PII or confidential data externally** to cloud-based AI services	●	●

Controls	Implement measures to **identify and handle user-generated content** that may contain sensitive or private information. Ensure that the AI system does not inadvertently generate or disclose personally identifiable information (PII) in its outputs.	●	●
Controls	Implement **masking** techniques during model training and/or inference to prevent the generation of specific sensitive information.	●	●
Controls	Implement guardrails to **identify, log and deter adversarial attacks** aimed at manipulating AI systems into generating unintended or privacy-compromising outputs.		●
Controls	Implement **continuous monitoring of generative AI outputs** to ensure consistency with privacy guidelines. Regularly assess and update the model based on emerging privacy concerns or changes in data protection regulations.		●
Controls	Integrate scheduled **human-in-the-loop safeguards** to allow human reviewers to assess, filter or remove sensitive / privacy-compromising data within AI models or vector databases.		●
Controls	Employ **privacy-preserving training techniques**, such as federated learning or homomorphic encryption, to train generative AI models without exposing sensitive training data to centralized systems		●
Fairness and Biases			
Policy	Implement policies to ensure that training **datasets are diverse and representative** of the target population. This should include guidelines for addressing under-representation to prevent biases in AI models or vector databases.	●	●
Policy	Establish policies for **bias-resistant feature engineering**, ensuring that the features used by AI models do not contribute to or amplify existing biases. Regularly review and update feature engineering practices to mitigate emerging biases.	●	●
Policy	Define and **track algorithmic fairness metrics** to quantitatively measure and monitor the fairness of the AI model across time. Regularly assess and update these metrics to align with evolving fairness standards and requirements.	●	●
Controls	Implement **techniques for detecting and mitigating biases** during both the training and deployment phases of the AI system. This includes exploring adversarial training methods and algorithmic interventions to reduce biased outcomes.	●	●
Controls	Implement a mechanism for collecting **user feedback on perceived biases** in AI system outputs. Use this feedback to iteratively improve the model's fairness and address any unintended biases identified by users.	●	●
Controls	Conduct inclusive user **testing with diverse groups** to identify potential biases in the user experience. Use feedback from testing to refine the AI system and improve its ability to cater to the needs of all users.		●
Controls	Engage external auditors or third-party organizations to conduct **independent assessments** on the AI system's fairness and biases.		●
Accuracy and Reliability			
Policy	Implement policies to ensure that the AI model's **performance metrics** align with the specific goals and requirements of its intended use case. Regularly review and update these metrics to reflect the evolving needs of the organization.	●	●
Policy	Put in place policies for **continuous monitoring** of the model's performance in real-world scenarios. Implement mechanisms to detect deviations from expected accuracy levels and trigger alerts for prompt investigation and intervention.	●	●

Policy	Implement policies and procedures to ensure the quality and reliability of **input data**. Regularly assess and validate the data sources to prevent the introduction of inaccurate or misleading information that could compromise the model's accuracy.	●	●
Policy	Establish policies for documenting and regularly reviewing the **assumptions underlying the model's design and implementation**. Ensure that any changes to these assumptions are thoroughly documented to maintain transparency and reliability.	●	●
Controls	Establish **communication protocols** for informing stakeholders about the level of accuracy and reliability expected from the AI model. Clearly communicate the limitations and uncertainties associated with model predictions to manage expectations.	●	●
Controls	Conduct **regular reviews of the model's performance**, focusing on accuracy, precision, recall, and other relevant metrics. Establish clear criteria for acceptable performance levels, and conduct corrective actions if deviations are identified.	●	●
Controls	Develop and implement procedures for model calibration to fine-tune the model's predictions based on **real-world feedback**. Regularly recalibrate the model to account for changes in the underlying data distribution and user behaviour.	●	●
Controls	Implement **cross-validation** practices to assess the model's generalization performance and reliability across different subsets of data. This helps ensure that the model performs consistently and reliably in diverse scenarios.	●	●
Controls	Implement policies for **detecting and handling outliers** in both training and inference phases. Address outliers promptly to prevent them from disproportionately influencing the model's accuracy and reliability.	●	●
Controls	Conduct regular **benchmarking exercises** to compare the AI model's performance against industry standards and best practices. Use benchmarking results to identify areas for improvement	●	●
Controls	Conduct **scenario-based testing** to evaluate the model's accuracy in specific situations or contexts relevant to its application. This includes testing the model's reliability in handling edge cases or critical scenarios that may have significant implications.		●
Controls	Develop mechanisms for incorporating user feedback into model improvement processes. Use feedback from end-users to identify issues related to accuracy and reliability and prioritize enhancements accordingly.		●
Transparency and Explainability			
Policy	Implement policies to thoroughly **document the architecture** of AI systems, including: (i) models used, (ii) training datasets involved in the foundational models, (iii) training datasets used in model refinement or RAG pipelines, (iv) algorithms employed, and (v) key parameters. Ensure that this documentation is accessible to stakeholders for a clear understanding of the model's inner workings.	●	●
Policy	Develop policies to provide transparency regarding the **sources of training data**. Clearly communicate the types of data used to train the model, including information about diversity, representativeness, and potential biases.	●	●
Policy	Establish protocols for **communicating updates or changes** to the model's architecture or algorithms to stakeholders. Clearly explain how these updates may impact the model's performance and decision-making processes.	●	●
Policy	Develop **human-readable model descriptions** that provide non-technical stakeholders with a clear understanding of the model's	●	●

	functionality and limitations. Avoid overly technical language and ensure that explanations are accessible to a diverse audience.		
Policy	**Document ethical considerations** related to the AI model's decision-making processes. Clearly articulate any ethical guidelines or principles adhered to during model development, training, and deployment.	●	●
Controls	Conduct user education initiatives to **enhance awareness and understanding of the AI system's capabilities and limitations**. Provide resources and materials to help users interpret and trust the decisions made by the AI model.	●	●
Controls	Implement **mechanism for providing explanations** in cases where the AI model encounters errors or makes incorrect predictions. Clearly communicate the reasons behind errors and outline steps taken to address and mitigate such instances.		●
Intellectual Property and Confidential Data Protection			
Policy	Implement organization-wide policies and mechanisms to prevent sensitive/confidential data leakage during the **sharing of documents, emails, or when employees use public AI tools.**	●	●
Policy	Implement organization-wide policies and mechanisms to prevent IP leakage during the **sharing of code base, algorithms, or training datasets.**	●	●
Policy	Put in place policies to secure the deployment of AI models, ensuring that proprietary algorithms and confidential data are not inadvertently exposed. Employ **encryption, access controls, and secure deployment practices** to mitigate the risk of unauthorized access.	●	●
Policy	Establish procedures for reviewing intellectual property considerations when collaborating on AI model development. Clearly **define ownership and usage rights** of jointly developed models to avoid disputes over IP in collaborative projects.	●	●
Controls	Establish audit trails and **logging** mechanisms to monitor and track the inclusion of proprietary algorithms or confidential data in (i) user prompts for inhouse/public AI systems or (ii) training data. Regularly review these audit trails to detect and respond to any unauthorized sharing of information.	●	●
Controls	Develop an **incident response plan** specifically for addressing IP breaches. Clearly outline the steps to be taken in the event of a breach, including communication protocols, legal actions, and remediation efforts to protect intellectual property.	●	●
Controls	Put in place **contractual protections** in agreements with third parties to safeguard intellectual property. Clearly define the ownership and usage rights of IP-related assets, including algorithms and datasets, to mitigate the risk of unauthorized use or disclosure.	●	●
Controls	When collaborating with external entities, implement policies for **anonymizing data shared** to protect confidential information. Ensure that shared data does not contain identifiable elements that could compromise the organization's proprietary knowledge.	●	●
Controls	**Validate the sources of training data** to ensure they comply with intellectual property laws. Avoid using training data that may infringe on copyrights or proprietary rights, and document the steps taken to verify the legality of the data sources.	●	●
Controls	Implement **employee training** programs to educate staff on the importance of protecting intellectual property and boundaries when using public AI tools. Ensure that employees are aware of best practices for handling proprietary information and the potential legal consequences of IP infringement.	●	●
Ethics and Reputation			

Policy	Establish an **ethical AI framework** that guides the development, deployment, and use of AI systems. Define principles and values that align with ethical standards and societal norms, ensuring responsible behaviour in AI applications.	●	●
Policy	Implement an ethics review board or committee responsible for evaluating the ethical implications of AI systems. Ensure diverse representation on the board to consider various perspectives and potential impacts on different user groups.	●	●
Policy	Put in place policies to ensure that the behaviour of the AI system aligns with the **organization's values and mission**. Regularly review and update these policies to reflect evolving ethical considerations and organizational priorities.	●	●
Policy	Develop a **crisis response plan** specifically for addressing ethical concerns related to AI. Define communication strategies, corrective actions, and transparency measures to manage and mitigate any negative impact on the organization's reputation.	●	●
Policy	Implement policies for **ethical data sourcing**, ensuring that datasets used for training and testing AI models do not perpetuate biases, discrimination, or unfair practices. Regularly audit data sources for compliance with ethical standards.	●	●
Policy	Implement policies to assess and address potential **human rights considerations** associated with the use of AI. Avoid applications that may contribute to discrimination, surveillance, or other human rights violations.	●	●
Policy	Establish guidelines for **responsible content generation and dissemination**, especially in platforms where AI algorithms make recommendations. Ensure that the AI system promotes content that adheres to ethical standards and avoids misinformation or harmful content.	●	●
Controls	Conduct algorithmic **impact assessments** to evaluate the potential social, economic, and cultural impacts of the AI system. Assess how the system's outputs may affect different communities and demographics.	●	●
Controls	Develop and implement strategies for **mitigating biases** in AI systems. Regularly assess and update these strategies to address emerging biases and ensure that the AI system behaves ethically across diverse user groups.	●	●
Controls	Implement **training programs** for AI developers and stakeholders that focus on ethical considerations in AI development. Ensure that those involved in AI projects are aware of the ethical implications and potential impacts on reputation.	●	●
Controls	Establish mechanisms for **public accountability**, such as regular reporting on AI ethics efforts and outcomes. Keep stakeholders informed about the organization's commitment to ethical AI practices and its contributions to societal well-being.	●	●
Controls	Prioritize **user empowerment and control** in AI systems by providing transparent options for users to customize and influence the behaviour of AI systems. Allow users to understand and adjust AI-driven features according to their preferences.	●	●
Controls	Engage third-party organizations to conduct **independent ethical audits** of AI systems. External audits can provide an impartial assessment of the ethical considerations and reputation risk associated with AI applications.	●	●
Controls	Implement **continuous monitoring** mechanisms to assess the ethical impact of AI systems over time. Regularly review and update ethical guidelines and practices based on evolving ethical standards and societal expectations.	●	●

Licensing, Legal and Regulatory Compliance			
Policy	Develop processes for **verifying and documenting license agreements** associated with open-source components or publicly sourced data used in AI development. Regularly audit and ensure compliance with the terms and conditions specified in these agreements.	●	●
Policy	Implement policies to ensure legal compliance when utilizing **open-source or public components/information**, particularly in the context of retrieval-augmented generative AI systems. Verify that the use of open-source materials aligns with licensing agreements and does not infringe on copyright laws.	●	●
Policy	Put in place policies to prevent the **repurposing of open-source models** (e.g. derivative models) designated for non-commercial use for commercial applications. Clearly communicate guidelines to AI developers and collaborators to avoid legal repercussions.	●	●
Policy	Establish **legal protocols for collaboration** with AI vendors, including clear agreements on licensing, usage, and ownership rights. Verify that collaborations align with both overarching laws and industry-specific regulations to avoid legal challenges.	●	●
Policy	If applicable, establish protocols to ensure compliance with regulations related to **cross-border data transfer**. Verify that the AI system adheres to data protection laws and safeguards data appropriately when transferred across jurisdictions.	●	●
Policy	Implement policies to ensure **adherence to data privacy regulations**, especially when dealing with sensitive user information. Verify that the AI system's data handling practices align with legal requirements to mitigate the risk of legal challenges.	●	●
Controls	Prior to integrating open-source foundation models into proprietary AI systems, conduct **legal reviews** to ensure alignment with licensing agreements. Verify that the integration complies with legal requirements and usage restrictions specified in the licenses.	●	●
Controls	Conduct **regular audits on licensing compliance**. Ensure that AI models, especially those incorporating open-source components, adhere to licensing agreements, and promptly address any deviations to prevent legal risks.	●	●
Controls	Conduct assessments to ensure compliance with **overarching laws and industry-specific regulations** governing AI systems. Regularly review and update compliance measures based on evolving legal frameworks and regulatory requirements.	●	●
Controls	Document the organization's efforts to achieve and maintain regulatory compliance. This includes **records** of compliance assessments, legal reviews, and actions taken to address any identified non-compliance issues.	●	●

Conclusion

AI audit has existed for years, but recent technological enhancements have triggered a new wave of AI adoption across industries and organizations. As the field is still evolving, our set of guidelines should be seen as a foundational, open-source work-in-progress that can be further built upon but also immediately adopted by companies to frame existing efforts.

Organizations should view AI audits as a means to upkeep standards, maintain compliance and mitigate risk. The necessary policies, controls or process implementations would ultimately depend on where organizations are in their AI adoption journey. Regardless, these are aspects that will need to be driven top-down.

V1.01 (24 Dec 2023)

Aigos – Securing AI Foundation

Speak with us