© July 2025| IJIRT | Volume 12 Issue 2 | ISSN: 2349-6002

Auditing AI – Extending SOC 2 Trust Criteria to Cover AI-Specific Risks

Prashant Magerde
National Law Institute University

Abstract- Artificial Intelligence (AI) is transforming the way businesses operate, enabling automation, predictive analytics, personalization, and decision-making at scale. However, the integration of AI into core business services introduces unique risks that are not fully addressed by traditional assurance frameworks such as SOC 2. This white paper explores how the existing SOC 2 Trust Services Criteria (TSC) can be extended to assess and mitigate AI-specific risks. It provides a structured methodology for embedding AI governance into the SOC 2 audit lifecycle and presents practical guidance for auditors, compliance professionals, and organizations adopting AI technologies.

INTRODUCTION

SOC 2, developed by the American Institute of Certified Public Accountants (AICPA), is a widely adopted framework for evaluating the effectiveness of an organization's controls related to security, availability, processing integrity, confidentiality, and privacy. However, as organizations increasingly deploy AI solutions, SOC 2 audits must evolve to accommodate new risk domains, including algorithmic bias, explain-ability, model drift, and data misuse. This paper aims to provide a comprehensive guide for extending SOC 2 practices to cover AI implementations, fostering trust and accountability in AI systems.

Understanding AI-Specific Risks

AI systems introduce unique technical, ethical, and operational risks that differ significantly from those in traditional IT environments. These risks can compromise trust, compliance, and fairness, especially when AI systems are used in decision-making processes that impact individuals or critical business outcomes.

- Algorithmic Bias and Discrimination AI models are trained on historical data, which may reflect existing biases and societal inequalities. If unchecked, this can result in discriminatory outcomes across gender, race, age, or socioeconomic groups. Bias may be introduced during data collection, labeling, feature selection, or model tuning, and can be exacerbated by imbalanced datasets.
- Lack of Explain-ability and Transparency Many AI models, particularly those based on deep learning, function as black boxes. This lack of interpretability limits stakeholders' ability to understand, verify, or challenge AI decisions. This creates compliance issues with regulations requiring transparency and undermines stakeholder trust.
- Data Poisoning and Model Manipulation Adversaries can intentionally inject malicious data into training sets (data poisoning) or craft inputs to manipulate outputs (e.g., adversarial examples). These attacks compromise model integrity and reliability, posing serious risks in applications like fraud detection, healthcare, or autonomous systems.
- Model Drift and Degradation AI models degrade over time as data patterns shift (e.g., due to changes in user behavior or market conditions). Without continuous monitoring and retraining, outdated models may yield inaccurate or misleading results, affecting decisions and outcomes.
- Security Risks in AI Pipelines AI development pipelines, which include data ingestion, model training, testing, and deployment stages, often involve multiple environments and dependencies. These pipelines may be vulnerable to supply chain attacks, misconfigurations, or unauthorized access, risking the integrity and confidentiality of the AI lifecycle.
- Privacy Violations and Data Leakage AI systems trained on large volumes of personal or sensitive

data may unintentionally memorize and expose individual records. Techniques such as model inversion, membership inference, or shadow model attacks can exploit models to reveal private information, creating privacy and legal concerns.

- Regulatory and Ethical Misalignment With global AI regulations evolving rapidly (e.g., EU AI Act, NIST AI RMF, ISO/IEC 42001), AI systems must be auditable, fair, and privacy-respecting. Failure to align with these frameworks can result in reputational damage, fines, or legal action. Ethical concerns, such as autonomy, human oversight, and value alignment, also need to be addressed.
- Operational and Organizational Risks AI deployment requires interdisciplinary collaboration between data scientists, engineers, risk managers, and compliance officers. Misalignment or lack of oversight in these areas can lead to poor model governance, lack of accountability, or failure to meet assurance requirements.

Mapping AI Risks to SOC 2 Trust Criteria

Trust Criteria	AI-Specific Risks	Suggested Controls
Security	• Adversarial attacks on models • Unauthorized access to training/model APIs • Model theft/tampering	• Role-based access controls for model artifacts • Secure model versioning • Adversarial robustness testing • End-to-end pipeline encryption
Availability	• Inference downtime • Model latency issues • Retraining delays/failures	• Load-balanced, scalable infrastructure • Model uptime monitoring • Scheduled retraining and update SLAs
Processing Integrity	• Bias or unfair outputs • Invalid or unchecked inputs • Model drift over time	• Bias and fairness validation • Input/output sanity checks • Drift detection and retraining triggers • Model limitation documentation
Confidentiality	• Exposure of training data • Model inversion or membership inference attacks	• Differential privacy in training • Federated learning approaches • Access controls for model internals • Data anonymization techniques
Privacy	• Use of PII without consent • Risk of re-identifying individuals via outputs	• Consent management tracking • AI-inclusive DPIAs • Transparency logs (e.g., model cards) • Use of synthetic/masked data

Extending the SOC 2 Audit Lifecycle for AI

To effectively address AI-specific risks within the SOC 2 framework, organizations and auditors must adapt each phase of the audit lifecycle to reflect the unique characteristics and operational challenges of AI systems. Below is a detailed breakdown of how the SOC 2 audit lifecycle can be extended for AI-integrated environments.

Scoping and Planning

- Identify AI Components: Clearly define which systems utilize AI, including internally developed models, embedded third-party algorithms, and machine learning services.
- Define Business Context: Understand how the AI system supports or impacts the organization's services covered under the SOC 2 audit.
- Determine Risk Relevance: Evaluate the role AI plays in core functions (e.g., decision-making, automation) and the sensitivity of the data processed.
- Include Stakeholders: Involve technical (data science/engineering), risk, compliance, and business stakeholders early in the planning phase to ensure comprehensive understanding.

Risk Identification and Assessment

- Model-Centric Risk Evaluation: Analyze potential risks related to model bias, drift, explainability, data quality, and adversarial vulnerabilities.
- Data Flow Mapping: Understand data inputs and outputs, including how data is collected, preprocessed, used in training, and how outputs are consumed.

- Assess Impact of AI Decisions: Identify scenarios where AI may influence critical or regulated decisions (e.g., customer eligibility, hiring, pricing).

Control Design and Implementation

- AI-Specific Control Objectives: Extend traditional SOC 2 control objectives to address the AI lifecycle—from model development and training to deployment and monitoring.
  - Examples include:
    1. Policies for acceptable model behavior
    2. Model retraining protocols
    3. Secure data handling practices throughout the pipeline
- Document AI Processes: Maintain clear documentation around model design, intended use, limitations, versioning, and governance policies.
- Embed Governance into Development: Ensure controls are integrated early in the AI development lifecycle (e.g., during model design and data selection stages).

Control Testing and Evidence Collection

- Evidence Types: Collect evidence to support control effectiveness across AI components, such as:
  - Logs from model monitoring platforms
  - Evidence of review and approval workflows for model changes
  - Records of performance testing, fairness evaluations, or explain-ability assessments
- Testing Techniques:
  - Interviews and walkthroughs with data scientists or ML engineers
  - Review of AI documentation (e.g., model cards, logs, data lineage)
  - Observation of inference behavior in production or simulated environments

Monitoring and Ongoing Assurance

- Continuous Oversight: Establish mechanisms to monitor AI system behavior post-deployment, including:
  - Model accuracy and relevance
  - Detection of anomalies or drift
  - Real-time alerting for unexpected outputs or failures
- Change Management: Ensure updates to AI models follow a structured process with proper testing, validation, and audit logging.

Reporting and Communication

- Transparency in Audit Reports: Clearly describe the presence and scope of AI systems within the audited environment.
- Control Effectiveness Commentary: Provide context around how AI risks were addressed and whether controls operated effectively during the audit period.
- Exception Handling: Document any control weaknesses, unexplained behavior, or unaddressed risks, and provide recommendations for mitigation.

Sample Controls for SOC 2

TSC	AI Risk Domain	Sample AI-Specific Controls
Security	Unauthorized access, adversarial manipulation	• Implement RBAC/ABAC for access to training data, models, and APIs • Conduct adversarial robustness testing during model evaluation • Use cryptographic hashing and digital signatures to track model integrity • Secure ML pipelines (CI/CD for ML) with secure coding and peer reviews
Availability	Model downtime, retraining failures	• Establish SLA thresholds for model availability and response time • Implement automated retraining and deployment pipelines • Monitor inference endpoints for latency, error rates, and crash loops • Maintain fallback mechanisms or rule-based alternatives for critical use cases
Processing Integrity	Bias, model drift, data quality	• Validate model inputs and outputs for accuracy, completeness, and timeliness • Perform pre-deployment bias audits using fairness metrics (e.g., demographic parity, equal opportunity) • Set up model drift detection mechanisms with thresholds for retraining • Document model limitations, assumptions, and data provenance (aligned with NIST AI RMF: Function - Map & Measure)

Confidentiality	Data leakage, model inversion	• Apply differential privacy during model training (e.g., DP-SGD) • Use federated learning to prevent raw data aggregation • Encrypt model parameters at rest and in transit - Restrict access to sensitive training datasets with audit logging and zero trust architecture
Privacy	Use of PII, consent, data subject rights	• Embed consent management into data ingestion workflows • Conduct privacy-preserving training (e.g., synthetic data, homomorphic encryption) • Maintain data minimization by excluding non-essential PII • Generate AI transparency reports and model cards (as suggested by EU AI Act and NIST AI RMF Function: Govern & Communicate)

Recommendations for Service Organizations and Auditors

As AI becomes embedded in core business services, both service organizations and SOC 2 auditors must evolve their practices to ensure that AI systems meet the Trust Services Criteria (TSC). Below are targeted recommendations to support responsible AI assurance and audit readiness.

For Service Organizations

1. Establish AI Governance Frameworks
- Form an AI risk and ethics committee comprising legal, compliance, data science, IT, and business leadership.
- Define clear roles and responsibilities for AI model owners, developers, and approvers.

2. Develop and Maintain Model Documentation
- Implement “model cards” for each AI system, detailing intended use, limitations, performance metrics, bias testing results, and retraining cycles.
- Maintain version history, data lineage, and audit logs for all training and inference pipelines.

3. Integrate Risk Management into the AI Lifecycle
- Embed risk reviews into model development, deployment, and monitoring phases.
- Perform regular risk assessments focused on explain-ability, fairness, robustness, and compliance with privacy regulations.

4. Ensure Model Transparency and Fairness
- Use interpretable models when possible or deploy explain-ability tools (e.g., SHAP, LIME) for black-box models.
- Run fairness and bias checks across sensitive attributes and make mitigation part of model approval criteria.

5. Strengthen Technical Controls
- Implement security controls across the AI pipeline, such as input validation, access control, encryption, and adversarial testing.
- Use privacy-enhancing technologies like differential privacy, federated learning, or homomorphic encryption.

6. Monitor and Validate Model Performance Continuously
- Automate model monitoring for drift, anomaly detection, and performance degradation.
- Establish thresholds that trigger retraining or rollback actions.

7. Align with Global AI Standards and Regulations
- Stay updated on regulatory developments (e.g., EU AI Act, NIST AI RMF, ISO/IEC 42001).
- Incorporate these guidelines into internal policies, AI system design, and control frameworks.

For SOC 2 Auditors

1. Expand Audit Scope to Include AI-Specific Risks
- Understand the role of AI systems in the service organization’s control environment and determine whether they impact TSCs.
- Adjust audit planning and procedures to address AI-specific components, including data sources, model behavior, and automated decision-making.

2. Adapt Control Testing for AI Environments
- Validate not only traditional IT controls but also those related to AI governance, model development, and monitoring.
- Examine the existence and effectiveness of model documentation, fairness assessments, versioning, and approval workflows.

3. Enhance Evidence Collection Practices

- Collect artifacts such as model cards, explainability reports, performance monitoring logs, and training dataset descriptions.
- Evaluate whether controls over data quality, access control, and bias mitigation are documented and enforced.

4. Train Audit Teams on AI Fundamentals

- Ensure auditors have foundational knowledge of AI concepts, risk types (e.g., bias, drift, adversarial attacks), and technical terminology.
- Encourage certifications or workshops on emerging AI assurance frameworks (NIST AI RMF, ISO/IEC 42001).

5. Report Clearly and Transparently

- Disclose in the SOC 2 report when AI systems were in scope, how AI-related risks were addressed, and whether any control gaps were identified.
- Recommend remediation strategies where AI systems failed to meet trust criteria or posed unexplained risks.

By following these recommendations, service organizations can strengthen the trustworthiness of their AI systems, and auditors can perform more relevant, informed, and future-proof SOC 2 engagements.

Future Outlook

As artificial intelligence continues to evolve and scale across industries, so too will the expectations of stakeholders, regulators, and auditors. Key developments shaping the future of AI assurance include:

- Global Regulatory Convergence: Jurisdictions such as the European Union, United States, and India are moving toward comprehensive AI regulations. The EU AI Act, NIST AI RMF, and ISO/IEC 42001 are setting global benchmarks. Organizations will soon be required not only to comply with these frameworks but also to demonstrate auditable governance practices for high-risk AI use cases.
- Standardized AI Control Catalogs: Efforts are underway to define industry-standard AI control sets, enabling consistent assurance approaches. These will complement existing SOC 2 controls, facilitating interoperability between privacy, security, and ethical compliance audits.
- AI Assurance as a Service: Third-party providers may begin offering specialized AI audits or attestations, much like SOC 2 reports today. This opens up new pathways for vendor trust, especially in sectors like fin-tech, health-tech, and HR tech.
- Rise of Explainable and Responsible AI: Explainability, fairness, and human oversight will become foundational requirements for AI systems—especially in regulated industries. Organizations that embed these values by design will gain competitive trust advantages.
- AI and Continuous Compliance: With real-time model updates and data drift, AI systems demand continuous monitoring. This will necessitate AI-integrated GRC tools, automated evidence collection, and dynamic control testing aligned with audit cycles.

In this landscape, extending SOC 2 to include AI-specific risk domains positions organizations ahead of both regulatory and reputational risk curves.

CONCLUSION

As AI becomes central to digital transformation, ensuring that it operates securely, ethically, and transparently is critical to sustaining stakeholder trust. The SOC 2 framework—long established as a gold standard for assurance—can and should be extended to address the unique challenges posed by AI systems. This white paper outlined a practical roadmap for mapping AI risks to SOC 2 Trust Services Criteria, adapting audit procedures, and implementing targeted controls aligned with global AI governance standards. By proactively integrating these practices into their risk and compliance functions, service organizations can not only meet evolving expectations but also lead the way in trustworthy AI adoption.

For auditors, this represents an opportunity to evolve their methodology, broaden their skill sets, and deliver greater value in an AI-powered world.

Ultimately, the convergence of AI assurance and SOC 2 provides a pathway to operationalize trust in AI—ensuring innovation does not outpace accountability.

Appendices

Appendix A: Sample AI Audit Questionnaire

1. What datasets were used to train the AI models, and how were they sourced?
2. Is any personally identifiable information (PII) used or inferred during training or inference?
3. How is bias in training data and model outputs detected and mitigated?
4. Are the AI system's decisions explainable to internal and external stakeholders?
5. How is the model performance monitored over time (e.g., accuracy, drift)?
6. What security controls are in place to protect AI training data, models, and APIs?
7. Are privacy-preserving techniques (e.g., differential privacy, data masking) implemented?
8. How is model versioning and audit logging managed across deployments?
9. What fallback mechanisms exist if the AI system fails or behaves unexpectedly?
10. Has the organization aligned its AI development practices with frameworks such as NIST AI RMF or EU AI Act?

Appendix B: AI Risk-Impact Matrix Template

AI Risk	Likelihood	Impact	Recommended Controls
Bias in decision making	Medium	High	Bias audits, data balancing, fairness testing
Model drift	High	Medium	Drift monitoring, retraining triggers
PII leakage	Low	High	Differential privacy, access controls
Adversarial attacks	Medium	High	Robustness testing, API security
Lack of explain-ability	High	Medium	Model cards, explain-ability tools

Appendix C: Glossary of AI & Assurance Terms

Model Drift: The degradation of model performance over time as input data patterns change.

Model Card: A standardized documentation format that provides transparency on an AI model's purpose, usage, and limitations.

Explain-ability: The ability to interpret and understand how an AI model arrives at a given decision.

Differential Privacy: A mathematical technique to ensure that individual data points cannot be inferred from aggregated datasets.

Federated Learning: A decentralized approach to training models where data remains on local devices.

Adversarial Example: Intentionally crafted input designed to mislead an AI model.

Bias Audit: An assessment process to detect and address unfair treatment of specific groups by AI systems.

Data Provenance: The history and origin of data used in training and evaluating AI models.

AI Governance: The framework of policies, procedures, and oversight applied to the development and deployment of AI systems.

Drift Detection: Techniques used to identify when a model's input or output distributions deviate significantly from expectations.

Appendix D: References

[1] AICPA Trust Services Criteria for Security, Availability, Processing Integrity, Confidentiality, and Privacy.
[2] NIST AI Risk Management Framework (AI RMF), Version 1.0, 2023.
[3] EU Artificial Intelligence Act – European Commission, 2024.
[4] ISO/IEC 42001:2023 – Artificial Intelligence Management System Standard.
[5] ENISA Guidelines: Adverse Impacts of AI, 2023.