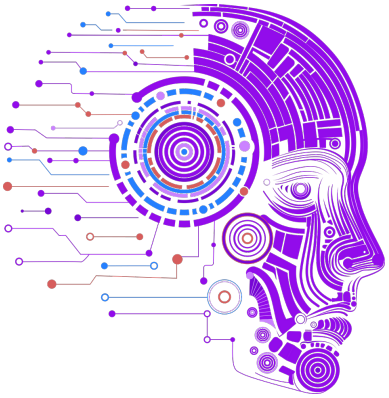


OWASP®

AI MATURITY ASSESSMENT

OWASP AI Maturity Assessment

Version V1.0 — August 11, 2025

Table of Contents

1 Preface

In recent years, innovation has moved faster than ever—but our ability to ensure that AI systems are secure, reliable, and aligned with human values hasn't kept pace.

Structured guidance isn't new to the software industry. Back in 2008, OWASP SAMM offered a practical maturity model that helped countless organisations embed security into their development processes. Today, with AI becoming a core part of products, infrastructure and everyday decisions, we're at a similar crossroads. The OWASP AI Maturity Assessment (AIMA) is our response.

As AI technologies become integral to products, services and critical infrastructure, the stakes are higher than ever. The industry is witnessing a surge in AI adoption, accompanied by heightened public scrutiny and evolving regulatory requirements. Organisations can no longer afford to be reactive when it comes to managing AI risks. Instead, there is a growing demand for actionable frameworks that empower teams to build AI systems responsibly, balancing innovation with oversight, agility with accountability, and technical excellence with ethical considerations.

AIMA adapts the foundational concepts of OWASP SAMM to the unique realities of AI lifecycle engineering. It extends traditional application security controls to encompass safeguards for data provenance, model robustness, privacy, fairness and transparency.

This document is intended for CISOs, AI/ML engineers, product leads, auditors and policymakers, helping them to translate high-level principles into day-to-day engineering decisions. Each maturity level is linked to tangible activities, artefacts, and metrics, enabling incremental improvement rather than disruptive transformation.

Version 1.0 introduces eight assessment areas, with detailed criteria and a worksheet for internal use or third-party evaluations. It's designed to help you spot gaps and manage risk across the entire AI lifecycle.

Of course, a model is only as strong as the community behind it. That's why we welcome your input—whether through GitHub Issues, pilot testing, or live discussions. Your feedback will shape future versions and help refine the scoring, guidance and coverage.

We would like to express our deepest gratitude to the OWASP Foundation, the SAMM core team, the early reviewers from academia and industry, and the volunteers who contributed test cases, glossary entries and real-world anecdotes.

We view this release as a living document. It will evolve in line with new research, regulatory changes and field experience, and we look forward to receiving your valuable input as we continue to develop it.

Onward,
Matteo Meucci & Philippe Schrettenbrunner
Project Co-Leads, OWASP AI Maturity Assessment

1.1 Authors

- Matteo Meucci
- Philippe Schrettenbrunner
- Arvinda Gangadhararao
- Sana Zia Hassan
- Abhinavdutt Singh
- Marco Denti
- Andrea Luigi Vitali
- Hubert Jackowski
- Keren Katz
- Montadhar Rekaya

We also want to thank everyone in the wider OWASP AIMA community, especially those in the Slack channel, who shared feedback, ideas, or encouragement along the way. Your input helped shape this project.

2 Introduction

The OWASP AI Maturity Assessment (AIMA) provides organizations with a structured approach for evaluating and improving the security, trustworthiness, and compliance of AI systems. Rooted in the principles of OWASP SAMM but tailored to the distinct challenges of AI, AIMA defines measurable pathways that guide responsible AI adoption across industries and organizational contexts.

AI systems introduce fundamentally new risks—ethical, operational, and technical—that require governance mechanisms beyond those used for traditional software. AIMA responds to this need with a risk-based model that integrates security, transparency, privacy, and lifecycle management into each phase of AI development and deployment.

2.1 The Need for AI Maturity Assessment Model

As organizations across the world accelerate the adoption of artificial intelligence (AI) technologies, the need for a structured AI Maturity Model has become increasingly important. AI systems present a unique set of challenges related to security, privacy, fairness, transparency, and accountability, which are not fully addressed by traditional IT governance models. A globally applicable AI maturity model enables organizations to evaluate their current capabilities, identify risk areas, and adopt best practices across the AI lifecycle. By integrating privacy-by-design and security-by-design principles from the very beginning—covering data collection, model training, deployment, and ongoing monitoring—organizations can foster trust, resilience, and ethical outcomes in AI applications.

Adopting an AI maturity framework is essential not only for reducing operational and regulatory risks, but also for building responsible, secure, and privacy-preserving AI systems that meet international expectations. These frameworks support consistency, interoperability, and compliance with evolving global standards, such as the EU AI Act,

OECD AI Principles, and emerging guidance from bodies like ISO and NIST. In an interconnected digital economy, maturity frameworks empower both public and private sector organizations to innovate confidently, safeguard individual rights, and reinforce trust in AI technologies that increasingly influence critical decisions across health, finance, education, and public services.

2.2 Why Existing Maturity Models Fall Short

Traditional maturity models like CMMI or OWASP SAMM provide proven methods for securing conventional software development, but they were not built with AI's unique properties in mind. AI-specific challenges include:

- **Non-deterministic behavior**: Model outputs change with data and context.
- **Opaque decision logic**: AI models often lack interpretability.
- **Data-centric vulnerabilities**: Adversarial attacks and data poisoning exploit training pipelines.
- **Dynamic risk surfaces**: AI systems evolve over time, requiring ongoing assurance.

Existing frameworks rarely address these issues comprehensively. Organizations attempting to apply them to AI are often left with policy-level principles and no actionable guidance.

2.3 What AIMA Adds

AIMA bridges the gap between principles and practice. It translates abstract goals such as fairness, robustness, and transparency into measurable activities and outcomes. It supports:

- **Contextual assessments**: Tailored to different levels of AI adoption and maturity.
- **Incremental improvement**: Maturity levels define a progression path without requiring immediate full compliance.
- **Cross-functional alignment**: Designed for technical teams, legal advisors, risk managers, and executive leadership.

Unlike some proprietary or closed maturity frameworks, AIMA is open-source and community-driven. It invites adaptation and evolution through real-world usage, feedback, and iteration.

2.4 The OWASP Ecosystem and AI-Specific Resources

AIMA builds on OWASP's broader commitment to AI security and privacy. Several sister projects provide complementary guidance:

- OWASP Top 10 for Large Language Model Applications: A curated list of the most critical security vulnerabilities in LLM-based systems.
- OWASP AI Security and Privacy Guide: Practical advice on building, testing, and procuring secure and privacy-preserving AI systems.
- OWASP AI Exchange: A comprehensive, community-driven repository of AI security and governance best practices.
- OWASP Machine Learning Security Top 10: A threat taxonomy for ML systems, including adversarial and infrastructure-level attacks.

Together, these resources form the backbone of AIMA's threat model, scope, and community approach.

3 The AIMA Model

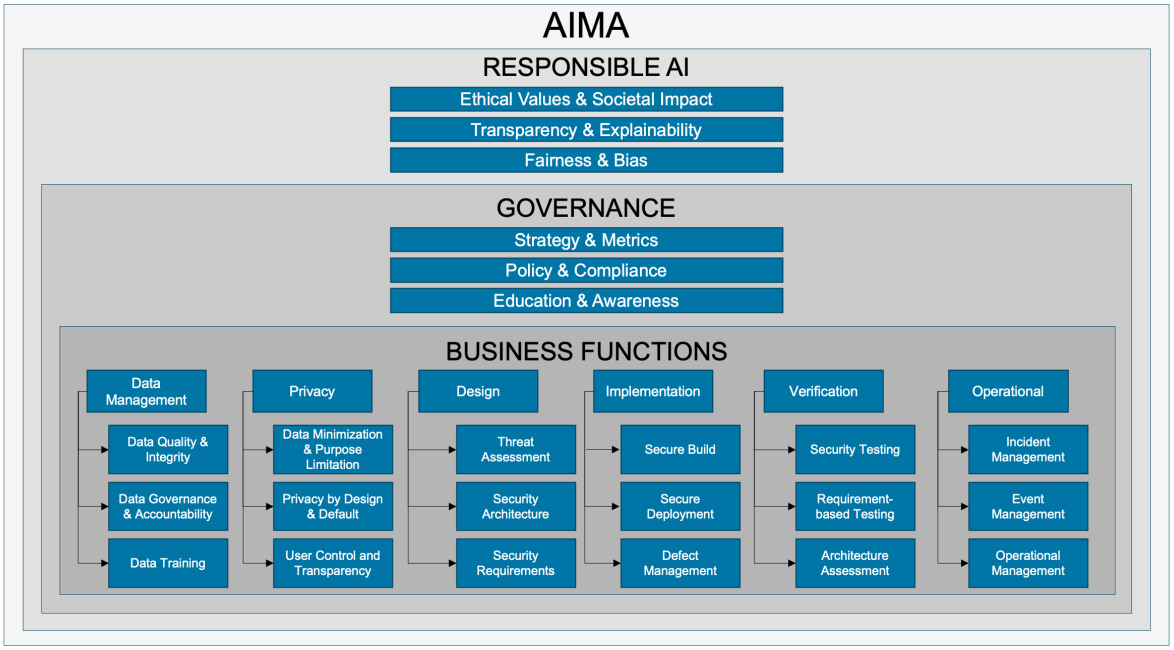
The Artificial Intelligence Maturity Assessment (AIMA) model emerges as a groundbreaking framework designed to guide organizations in designing, developing and deploying responsible and trustworthy AI systems. At its core, AIMA integrates robust Responsible AI (RAI) principles, strategic AI governance practices, and comprehensive alignment with critical business functions, creating a clear pathway towards AI maturity.

AIMA defines eight assessment domains that span the entire AI system lifecycle:

- **Responsible AI Principles**: Fairness, transparency, and societal impact.
- **Governance**: Strategy, policy, and education.
- **Data Management**: Quality, accountability, and training data practices.
- **Privacy**: Data minimization, privacy by design, and user control.
- **Design**: Threat modeling, security architecture, and requirements.
- **Implementation**: Secure build, deployment, and defect management.
- **Verification**: Testing and architecture validation.
- **Operations**: Monitoring, incident response, and system lifecycle management.

Each domain includes maturity criteria grouped into two complementary streams: *Create & Promote* (stream A) and *Measure & Improve* (stream B). Organizations can evaluate their posture, identify gaps, and prioritize improvements in a structured way.

This structured approach begins with foundational ethical values, emphasizing the societal impacts of AI, transparency and explainability of models, and proactively addressing fairness and bias. Building on these core values, AIMA incorporates detailed governance strategies to ensure organizations not only comply with regulatory demands but proactively shape internal policies, measurable strategies, and cultivate ongoing awareness and education among stakeholders.



Moreover, the AIMA model uniquely connects these overarching principles and governance practices directly with tangible business processes. By meticulously outlining critical areas—from Data Management and Privacy to Design, Implementation, Verification, and Operational practices—it empowers organizations to systematically evaluate and enhance their AI capabilities, ensuring secure, reliable, and ethical AI adoption. Through this integrated, holistic perspective, AIMA equips organizations to navigate the complexities of AI, promoting sustainable innovation and long-term success.

In the next chapters we will describe each of the Practices described in the model.

3.1 Responsible AI

The **Responsible AI** pillar addresses the distinct ethical and societal implications that arise specifically from artificial intelligence. While traditional software primarily focuses on functionality and security, AI systems introduce unique risks around fairness, transparency, and broader societal impacts. The opacity of AI decision-making processes, the potential for unintended biases in training data, and the wide-ranging effects of AI-driven outcomes on society necessitate dedicated governance.

To meet these challenges, this pillar focuses on three core practices:

1. **Ethical and Societal Impact** – Systematically assessing the broader ethical considerations and societal consequences of deploying AI systems, actively managing risks, and aligning AI deployment with organizational values and public expectations.
2. **Transparency and Explainability** – Providing meaningful insights into AI decisions, ensuring stakeholders can understand, trust, and appropriately challenge automated outcomes, facilitating accountability.
3. **Fairness and Bias** – Proactively identifying, mitigating, and continuously monitoring biases within AI systems to ensure fair treatment and equitable outcomes across diverse user groups.

Together, these practices foster responsible innovation, build user trust, reduce reputational risk, and ensure alignment with regulatory requirements and societal values.

3.1.1 Ethical Values and Societal Impact

Ethical values are fundamental in the design, development and implementation of Artificial Intelligence systems. As AI technologies increasingly influence various aspects of society and the world we live in, it is critical to ensure that these systems are in line with ethical values such as the centrality of the Human Being, the Well-Being of the Human Being, respect for Human Rights, compliance with the Law, respect for the Environment. Addressing these issues involves proactively identifying and mitigating potential negative impacts to human beings and the environment. By incorporating ethical considerations into AI practices, organizations can build stakeholder trust, comply with regulatory requirements, and contribute positively to the communities and world in which we live.

Objectives

1. **Promote Human-Centric**: AI must be conceived, designed, developed and used as a tool to serve the human being.
2. **Support Human Well-Being**: AI's ultimate goal is the improvement of Human well-being.
3. **Safeguard Fundamental Rights**: AI must always respect, protect and promote the rights of the human being, as a person and as a community.

4. **Ensure Rule of Law**: AI must always comply with existing and applicable laws and regulations.
5. **Secure Environmental Protection**: AI must promote environmental sustainability and the well-being of ecosystems throughout their life cycle.
6. **Encourage Stakeholder Engagement**: Involve diverse stakeholders in the design, development and evaluation of AI systems to reflect a broad range of perspectives and values.

Streams

Maturity Level	Stream A	Stream B
1 - Define and implement a structured approach to AI ethics and risk management, replacing informal, reactive handling with proactive policies aligned to your organization's business goals, values, and regulatory obligations.	- **Incident-Driven:** Ethical concerns addressed post-incident without consistent practices. - **Informal Accountability:** Ethical responsibilities assigned ad-hoc with minimal documentation. - **Limited Follow-Up:** Post-incident documentation with little structured learning or improvement.	- **Occasional Discussions:** Ethical topics addressed informally, typically driven by personal initiative. - **No Structured Training:** Ethical training is absent or ad-hoc, with no formal programs, onboarding content, or role-specific support provided by the organization.. - **Variable Awareness:** Ethics understanding varies across teams without shared organizational standards.
2 – Implement structured AI governance frameworks with formalized ethics and environmental policies, defined roles, and accountability mechanisms to ensure consistent oversight.	- **Defined Ethical and Environmental Policy:** Explicit policy outlines values, principles, and responsibilities. - **Ethics Governance:** Designated Ethics Officers or Committees oversee ethical practices and governance. - **Integrated Assessments:** Ethical and Environmental impact assessments systematically	- **Role-Specific Training:** Ethics training tailored to roles conducted regularly. - **Supported Discussions:** Encouraged open forums for ethical dilemmas and ongoing dialogue. - **Routine Reflection:** Ethical and Environmental considerations integrated into regular project activities.

Maturity Level	Stream A	Stream B
	embedded into planning and documentation.	
3 - Embed an Ethical AI Culture by continuously integrating ethical principles into AI development, monitoring outcomes, and reinforcing values across organizational practices.	- **Continuous Monitoring:** Ethical and environmental KPIs actively tracked and aligned with organizational performance metrics. - **Policy Evolution:** Regular updates based on stakeholder feedback and real-world insights. - **Automated Integration:** Ethics and environmental tools and processes embedded throughout all project lifecycle phases.	- **Rewarded Ethics:** Ethical and environmental behavior recognized in career progression and performance evaluations. - **Cultural Reinforcement:** Regular events and leadership modeling to reinforce proactive ethical and environmental behavior. - **Normalized Decision-Making:** Ethical and environmental considerations standard across all organizational decision-making levels.

3.1.2 Transparency and Explainability

Transparency and explainability are essential components of trustworthy AI systems. Transparency involves openly sharing information about how AI systems operate, including their design, data sources, and decision-making processes. Explainability focuses on providing clear and understandable reasons for AI decisions, enabling stakeholders to comprehend, trust, and effectively interact with these systems. Together, they ensure that AI systems are not "black boxes" but are instead accountable and aligned with ethical standards. This is particularly crucial in high-stakes domains like healthcare, finance, and criminal justice, where decisions can have significant impacts on individuals' lives. By fostering transparency and explainability, organizations can build user trust, facilitate compliance with regulations, and enable meaningful oversight and governance of AI technologies.

Objectives

1. **Enhance Trust and Accountability**: Ensure that stakeholders can understand and trust AI decisions, fostering confidence in AI systems.

2. **Facilitate Regulatory Compliance**: Meet legal and ethical standards by providing clear explanations for AI-driven outcomes.
3. **Enable Effective Oversight**: Allow for monitoring and auditing of AI systems to detect and correct errors or biases.
4. **Improve User Engagement**: Empower users to make informed decisions by understanding how AI systems arrive at specific outcomes.
5. **Support Continuous Improvement**: Use insights from explanations to refine AI models and processes, enhancing performance and fairness over time.

Streams

Maturity Level	Stream A	Stream B
1 - Establish a formal Transparency Model, where information is shared inconsistently and only in response to external demands.	- **Manual Documentation:** Documentation created reactively, usually after issues arise. - **Informal Roles:** Transparency responsibilities assigned ad-hoc, without formal definitions. - **Contextual Gaps:** Outputs frequently lack sufficient interpretability and context.	- **Informal Awareness:** Explainability discussed in informal settings; formal training absent. - **Voluntary Queries:** Encouraged, but not required, model explanation requests. - **Individual-Driven:** Transparency awareness driven by personal interest rather than institutional norms.
2 - Define Structured Implementation approach with formalized policies, tooling and clear responsibilities.	- **Defined Policy:** Established transparency and explainability policy guiding documentation and tool usage. - **Role Clarity:** Champions appointed to ensure explainability across teams. - **Standardized Tools:** SHAP, LIME, and model cards embedded into development pipelines.	- **Role-Based Training:** Targeted training on interpretability techniques provided regularly. - **Systematic Retrospectives:** Teams regularly review the clarity and impact of model explanations. - **Growing Consistency:** Transparency practices standardized and shared more broadly across the organization.
3 - Embedded Transparency Culture with Continuous	- **Automated Processes:** Explanation documentation automated and validated within CI/CD workflows.	- **Performance Integration:** Transparency effectiveness included in individual and team performance evaluations.

Maturity Level	Stream A	Stream B
measurement, automated transparency aligned with goals	- **Real-Time Metrics:** Transparency metrics continuously monitored via dashboards aligned to strategic KPIs. - **Automated Remediation:** Trigger automatic remediation workflows when explanation standards are unmet.	- **Cultural Innovation:** Organization-wide explainability events (e.g., hackathons) enhance innovation and accountability. - **Open Dialogue:** Institutional norms promote ongoing dialogue and continuous improvement in transparency and explainability.

3.1.3 Fairness and Bias

Fairness and bias practices make sure everyone involved in AI work - designers, developers, deployers, and the people making rules - actively tries to spot and fix unwanted biases in data, algorithms and results. Technical fixes like debiasing algorithms and fairness metrics matter but aren't enough on their own. We need organizational structures too - clear roles, policies and ways to learn - that turn ethical AI principles into everyday decisions. AI systems learn biases from historical data, algorithm choices, and human oversight. If we don't watch for these biases, they can keep discrimination going in hiring, lending, healthcare and more areas. A good fairness practice that grows over time combines awareness, defining requirements, using tools for assessment, and constant feedback to build responsible-AI thinking and systems that can handle problems.

Objectives

1. **Raise Organization-Wide Fairness Awareness**: Ensure that developers, data scientists, product managers, executives, and procurement teams understand common bias sources (data imbalance, label bias, proxy variables) and the business, legal, and reputational risks they pose.
2. **Define Role-Specific Fairness Requirements**: Document clear fairness criteria for each role: data stewards must perform representativeness checks; modelers must run statistical parity and equalized odds tests; product owners must evaluate user-impact scenarios and disparate outcomes.
3. **Embed Standardized Assessment Toolkits**: Roll out open-source frameworks and explainability libraries into CI/CD pipelines to automate bias detection and reporting at key development milestones.

4. **Establish Fairness Governance Forums**: Convene cross-functional councils—including legal, ethics, UX, and impacted community representatives—to review high-risk models, adjudicate trade-offs, and approve remediation plans before production deployment .
5. **Measure & Iterate on Fairness Outcomes**: Track fairness KPIs (e.g., demographic parity difference, false positive/negative rate gaps) through dashboards; conduct regular bias audits and red-teaming exercises to surface emergent issues, then refine policies, training, and tooling accordingly

Streams

Maturity Level	Stream A	Stream B
1 - Establish adhoc approach to respond to requests towards Fairness and bias	- **Ad Hoc Response:** Bias addressed inconsistently, primarily after complaints or incidents. - **Unclear Roles:** Responsibilities assigned informally, without defined roles or documented processes. - **Lack of Tools:** No standardized tools, checkpoints, or processes established for bias assessment.	- **Limited Awareness:** Cultural awareness driven by individual initiative without formal training. - **Informal Reporting:** Reporting of bias concerns voluntary and unstructured; insights not consistently acted upon. - **No Defined Metrics:** Absence of formal metrics or tracking methods for bias-related issues.
2 - Define structured Implementation with formalized policies and processes, but limited integration.	- **Defined Policies:** Formal policies, charters, and governance forums guide bias mitigation efforts. - **Tool Integration:** Fairness assessment tools and documentation used at key project milestones. - **Regular Assessments:** Regular bias evaluations conducted but not always tied explicitly to KPIs or business outcomes.	- **Role-Specific Training:** Regular fairness training tailored to specific roles. - **Feedback Mechanisms:** Project retrospectives and knowledge sharing occur regularly post-release. - **Partial Engagement:** Cultural engagement and fairness awareness present but inconsistently applied organization-wide.
3 - Embedded Fairness Culture with fully integrated	- **Automated Monitoring:** Continuous, automated bias detection tools trigger real-time	- **Incentivized Culture:** Fairness integrated into career growth, performance reviews, and

Maturity Level	Stream A	Stream B
into core processes, automated monitoring, continuous improvement.	remediation. - **Enterprise-Wide Metrics:** Fairness KPIs tracked organization-wide, integrated into business performance metrics and OKRs. - **Process Integration:** Fairness assessments enforced through automated CI/CD pipelines and ongoing production validations.	recognition programs. - **Proactive Exercises:** Regular red-team exercises and simulations strengthen organizational resilience to bias. - **Continuous Enhancement:** Active promotion of continuous improvement initiatives across all teams, regularly celebrated and incentivized.

3.2 Governance

Artificial-intelligence systems introduce **unique governance challenges** that go well beyond those of traditional software:

- **Non-deterministic behaviour** – Predictions can drift as data or context change, requiring continuous oversight.
- **Opaque decision logic** – Complex models make it hard to explain outcomes, raising regulatory, ethical, and reputational risks.
- **Data-centric attack surface** – Poisoned or biased data can silently compromise security and fairness before a single line of code is written.
- **Rapidly evolving regulation** – Frameworks such as the EU AI Act, NIST AI RMF, and ISO 42001 demand traceable risk controls throughout the model lifecycle.

The **Governance pillar** of the **OWASP AI Maturity Assessment (AIMA)** equips organisations to meet these challenges by embedding a closed-loop system of **direction, control, and enablement** across every AI initiative. It is built around three mutually reinforcing practices:

1. **Strategy & Metrics** – Define a forward-looking AI-security and Responsible-AI vision, map it to business value and risk appetite, and quantify progress with model-aware KPIs/KRIs (e.g., adversarial-robustness scores, bias indices, model-lifecycle coverage).

2. **Policy & Compliance** – Translate strategy into enforceable, AI-specific standards (data provenance, model transparency, human-in-the-loop thresholds) and continuously prove conformance to internal and external requirements.
3. **Education & Guidance** – Ensure every actor in the AI supply chain—from prompt engineer to board director—has the role-tailored knowledge, playbooks, and real-time guardrails needed to make secure, ethical decisions.

Together these practices create a **governance engine** that:

- **Connects intent to execution** – Policies, controls, and training all trace back to strategic AI-risk objectives.
- **Enforces "shift-left" accountability** – Requirements are embedded in data pipelines, model cards, and CI/CD gates, not bolted on after deployment.
- **Drives measurable improvement** – Continuous metrics expose blind spots, inform investment, and demonstrate due diligence to auditors and regulators.

By assessing maturity against the Governance pillar, organisations can chart a clear, incremental path from **ad-hoc experimentation** to **institutionalised, trustworthy AI**—unlocking innovation while safeguarding users, regulators, and the business itself.

3.2.1 Strategy and Metrics

The Strategy & Metrics practice sits at the foundation of AI-security governance. Its purpose is to make sure the organization knows where it is going with AI security (the strategy) and how it will know it is getting there (the metrics). By defining a clear, business-aligned AI-security strategy and pairing it with objective, repeatable measures of success, leaders gain the insight needed to allocate resources, manage risk, and demonstrate return on investment. Sound strategy keeps AI initiatives pointed toward the organization's risk appetite and regulatory obligations, while robust metrics turn vague goals into actionable data that drives continuous improvement across the AI life-cycle.

Objectives

1. **Strategic Alignment**: Align AI Security Strategy with organizational business goals and risk management.
2. **Metrics and Value Tracking**: Define and Track AI Security Metrics that guide improvements and demonstrate value.

3. **Continuous Improvement**: Ensure Continuous Improvement of AI security practices through iterative reviews.

Streams

Maturity Level	Stream A	Stream B
1 - Establish an AI Security and Responsible AI Strategy aligned with the organization's overall business goals, ethical standards, and risk profile.	- **Minimal Alignment:** AI security and RAI efforts are not consistently linked to business or ethical goals. - **Unclear Accountability:** No formal ownership for AI security or ethical governance; responsibilities may be scattered. - **Ad Hoc Processes:** AI security actions happen on-demand (e.g., after an incident), with no strategic roadmap.	- **No Formal Metrics:** AI security and RAI outcomes (e.g., incident counts, bias incidents, model validation) are not measured or measured informally. - **Incident-Driven Insights:** Data is gathered primarily after security or ethical incidents with no routine analysis. - **Lack of Standardization:** Reporting varies widely, making organization-wide comparisons difficult.
2 - Define and Track AI Security and RAI Metrics to measure effectiveness, maturity, fairness, transparency, and return on investment.	- **Documented Strategy:** A formal AI security and RAI strategy exists, referencing relevant enterprise risk, compliance, and ethical needs. - **Clear Governance:** Defined roles (AI Security Lead, AI Ethics Officer, AI Security Committee) ensure accountability, fairness, and decision-making. - **Planned Integration:** AI security and ethical oversight efforts included in project roadmaps, budgets, and organizational planning.	- **Established Metric Set:** KPIs/ KRIs (e.g., fairness metrics, model risk classification, explainability standards) tracked over time. - **Regular Collection & Reporting:** Metrics gathered at intervals and shared with stakeholders through dashboards/ reports. - **Action-Oriented Insights:** Metrics drive resource allocation, ethical policies, fairness improvements, and actions for regulatory compliance.
3 - Continuously Improve AI Security and RAI Posture through iterative learning,	- **Fully Embedded:** AI security and RAI strategy integrated into broader corporate governance and ethics frameworks, continuously updated.	- **Advanced Analytics & Monitoring:** Real-time monitoring of AI systems (data drift, adversarial attack detection, bias detection), automated alerts,

Maturity Level	Stream A	Stream B
adaptation, and ethical alignment.	- **Executive Sponsorship:** Senior leadership proactively supports AI security and responsible AI as strategic investments. - **Lifecycle Integration:** Mandatory AI security controls (model audits, fairness assessments, transparency measures, human oversight protocols) throughout all AI development and deployment phases.	and comprehensive audit trails. - **Predictive & Preventive Metrics:** Metrics forecast risks (ethical, security, compliance issues) proactively addressing concerns. - **Culture of Data-Driven and Ethical Governance:** Metrics feed strategic decision-making; clear processes for continuous feedback, fairness enhancements, transparency improvements, and regulatory compliance.

3.2.2 Policy and Compliance

The **Policy & Compliance** practice translates high-level AI-security and Responsible-AI principles into concrete rules and oversight mechanisms. By formalizing AI-specific policies—and continuously verifying that systems, data, and processes comply with internal standards and external regulations—organizations reduce legal exposure, safeguard user trust, and uphold ethical commitments. Effective policy frameworks provide clear guidance to data scientists and engineers, while structured compliance activities (risk assessments, audits, attestations) create documented evidence that AI initiatives operate within agreed security, privacy, and fairness boundaries.

Objectives

1. **Define and Maintain AI-Specific Policies & Standards** that cover security, privacy, ethics, and quality across the AI lifecycle.
2. **Ensure and Demonstrate Compliance** with applicable laws, regulations, and internal governance requirements.
3. **Drive Continuous Policy Improvement** through regular reviews, monitoring, and automation of enforcement and assurance activities.

Streams

Maturity Level	Stream A	Stream B
1 – Establish Baseline AI Policies and Compliance Awareness to address foundational security, privacy, and ethical obligations.	- **Minimal AI-Specific Policies:** AI risks are loosely covered by general IT/security policies, if at all. - **Reactive Updates:** Policies change only after incidents or regulatory pressure. - **Limited Guidance:** Teams lack clear instructions for secure or responsible AI development.	- **Reactive Compliance:** Efforts focus on ad-hoc responses to audits or incidents. - **Limited Oversight:** No systematic tracking of AI-related regulations or risks. - **Informal Risk Assessment:** Assessments, when performed, are inconsistent and undocumented.
2 – Document and Enforce AI Policies, and Implement Structured Compliance Processes for security, privacy, and ethics.	- **Documented AI Policies & Standards:** Formal requirements cover data use, model validation, bias testing, explainability, etc. - **Periodic Reviews:** Policies reviewed on a defined schedule or when major changes occur. - **Consistent Application:** Projects follow standards; exceptions require documented approval.	- **Established Compliance Processes:** Regular reviews (privacy impact, bias audits) align with known regulations (e.g., GDPR, AI Act). - **Consistent Risk Framework:** A risk register tracks AI security and ethical posture across projects. - **Internal Audit & Reporting:** Findings are reported to governance bodies; remediation is tracked.
3 – Continuously Optimize Policies and Compliance Governance with proactive monitoring, benchmarking, and automation.	- **Integrated Policy Framework:** AI policies embedded in enterprise governance, risk, and ethics programs. - **Proactive Evolution:** Updates anticipate emerging threats and regulations, guided by continuous risk scanning and industry input. - **Automated Enforcement:** CI/CD gates, data-use controls, and policy-as-code tooling flag or block non-compliant artifacts automatically.	- **Holistic Compliance Integration:** Real-time regulatory watchlists inform automatic updates to controls and checklists. - **Advanced Risk Analytics:** Continuous monitoring detects drift, bias, or security anomalies that could trigger compliance breaches. - **Benchmarking & Certification:** The organization measures itself against leading frameworks and pursues external

Maturity Level	Stream A	Stream B
		attestations to demonstrate excellence.

3.2.3 Education and Guidance

The **Education & Guidance** practice ensures that everyone who influences an AI system—developers, data scientists, product managers, executives, risk officers, even procurement—understands their security, privacy, and ethical responsibilities. Good intentions alone do not create secure or trustworthy AI; people need the right knowledge, tools, and decision-making frameworks at the right moments. By establishing structured learning paths and feedback loops, organizations embed security-first and responsible-AI thinking into daily work, reducing the likelihood of errors and enabling fast, coordinated responses when new threats or regulations emerge.

Objectives

1. **Raise AI-Security and Responsible-AI Awareness** across all roles that design, build, deploy, or oversee AI.
2. **Provide Role-Specific Training and Resources** that are actionable, current, and aligned with policy and risk priorities.
3. **Measure Learning Effectiveness and Continuously Improve** curricula, guidance, and tooling based on feedback and emerging challenges.

Streams

Maturity Level	Stream A	Stream B
1 – Establish Baseline AI-Security and RAI Awareness for anyone touching AI initiatives.	- **Ad-Hoc Learning:** Security and ethics topics appear sporadically in general tech training or after incidents. - **Limited Reach:** Only core engineering teams receive any AI-security guidance; business and risk stakeholders rarely included. - **Informal Materials:** Slide	- **No Formal Measurement:** Completion rates, quiz scores, or adoption of guidance are not tracked. - **Reactive Improvements:** Content is updated only when major issues arise. - **Knowledge Gaps Unidentified:** The organization

Maturity Level	Stream A	Stream B
	decks or wiki pages exist but are not curated or kept up to date.	lacks insight into which roles need deeper AI-security skills.
2 – Provide Structured, Role-Based AI-Security and RAI Training aligned with policies and risk appetite.	- **Documented Curriculum:** Mandatory courses cover AI-specific threats, privacy, bias, and incident response; electives address deeper topics like adversarial ML or model interpretability. - **Role Tailoring:** Distinct learning paths for developers, data scientists, product owners, and executives. - **Guidance Library:** Curated playbooks, checklists, and coding examples are integrated into day-to-day tools (e.g., notebooks, IDE extensions).	- **Tracked Participation & Assessments:** Learning-management system tracks completion, scores, and certifications. - **Feedback Loops:** Learners rate relevance; course owners revise based on survey data and policy updates. - **Skill Gap Analysis:** Regular reviews map workforce skills to upcoming AI projects and risk areas.
3 – Embed Continuous, Data-Driven Learning Culture that adapts to evolving AI threats and regulations.	- **Just-In-Time Micro-Learning:** Contextual tips and secure-by-design snippets appear in pipelines, notebooks, and code reviews. - **Community & Mentorship:** Internal forums, guilds, and brown-bag sessions foster knowledge sharing; external conferences encouraged. - **Automated Guidance Updates:** New threat intel or policy changes automatically trigger content refresh and notification to affected roles.	- **Performance-Linked Metrics:** Training impact measured through defect density, incident trends, and model audit scores. - **Adaptive Curriculum:** AI identifies learning gaps and personalizes content sequences. - **Benchmarking & Recognition:** Organization compares learning maturity against industry, offers badges or career incentives, and publicly shares best practices to demonstrate leadership.

3.3 Data Management

The **Data Management** pillar addresses the unique and critical role that data plays in AI systems. Unlike traditional software, AI relies fundamentally on large datasets not only for operation but also for training and validation. Consequently, the security, quality, and governance of data significantly impact AI system reliability, fairness, and security.

To address these specialized challenges, the Data Management pillar includes three interconnected practices:

- **Data Quality and Integrity** – Ensuring data accuracy, completeness, reliability, and protection against tampering or corruption, which directly influences AI model performance and security.
- **Data Governance and Accountability** – Establishing clear ownership, responsibilities, and oversight mechanisms to manage data securely, ethically, and compliantly across its entire lifecycle.
- **Data Training** – Managing the collection, curation, and use of training datasets securely and responsibly, with explicit attention to bias prevention, data privacy, and traceability.

Together, these practices enable organizations to proactively manage data-driven risks, comply with evolving data regulations, and ensure AI systems operate securely, ethically, and effectively at scale.

3.3.1 Data Quality and Integrity

In the context of AI readiness, data quality and data integrity are foundational pillars that determine the effectiveness, reliability, and trustworthiness of AI systems. AI models are only as good as the data they are trained on. High-quality data ensures that models learn meaningful patterns, produce accurate predictions, and adapt well to changing environments. Integrity safeguards that data remains consistent, accurate, and trustworthy across its lifecycle i.e. from ingestion and processing to storage and consumption. Without robust data quality and integrity practices, organizations risk introducing bias, inaccuracies, compliance violations, and operational inefficiencies into their AI initiatives, ultimately undermining the business value AI is intended to deliver.

Definitions:

1. Data Quality for AI-ready data refers to the degree to which data is accurate, complete, consistent, timely, relevant, and fit for use in training, validating, and deploying AI models.
2. Data Integrity refers to the assurance that data remains authentic, reliable, and unaltered throughout its lifecycle, with robust mechanisms for traceability, auditability, and protection against corruption or unauthorized modifications.

Objectives

1. **Establish Data Fitness for Advanced AI Systems**: Ensure organizational data assets meet the stringent requirements for training, fine-tuning, and operating LLMs and agentic AI systems, including breadth, depth, freshness, and minimal noise or bias.
2. **Enable Scalable and Trustworthy AI Deployment**: Create a consistent foundation where AI models, including autonomous agents, can safely consume, reason over, and act upon data without human intervention compromising operational integrity.
3. **Identify and Prioritize Risk Areas**: Detect gaps in data sourcing, governance, and lifecycle management that could expose AI systems to risks such as model drift, data poisoning, decision errors, or compliance failures.

Streams

Maturity Level	Stream A	Stream B
1 - Establish an approach to identify and respond data quality or integrity issues reported.	- **Siloed Data:** Data fragmented, unstructured, lacking standardized definitions. - **Poor Quality:** High duplicates, missing values, and noise. - **No Validation:** Absence of accuracy or relevance validation rules.	- **No Traceability:** Data lineage or traceability non-existent. - **Manual Handling:** High risk of tampering or corruption due to manual updates. - **Poor Auditability:** Audit logs absent or unreliable.
2 - Define a formal approach with documented	- **Initial Cleansing:** Basic data profiling and cleansing processes implemented.	- **Partial Lineage:** Data lineage partially established across main systems.

Maturity Level	Stream A	Stream B
processes and initial rules for managing quality and integrity in data sets.	- **Early Standards:** Initial completeness and consistency rules applied. - **Metadata Tracking:** Early stages of data cataloging and metadata management.	- **Manual Change Tracking:** Some manual tracking of data changes, minimal automation. - **Inconsistent Controls:** Access controls in place but inconsistently enforced.
3 - Create a quality culture with fully integrated data management practices with robust, automated mechanisms for maintaining quality and integrity.	- **Standardized Metrics:** Defined metrics for accuracy, completeness, consistency, and timeliness systematically tracked. - **Active Quality Management:** Continuous data quality checks, real-time scoring, LLM-specific data filters (e.g., toxicity, hallucination-prone data). - **Curated Data:** Regular curation based on model feedback and bias tracking.	- **Full Traceability:** Comprehensive lineage and versioning across entire AI data pipeline. - **Automated Integrity Checks:** Real-time monitoring, immutable audit logs, automated anomaly detection for corruption, drift, or unauthorized changes. - **Proactive Compliance:** Integrated integrity checkpoints supporting rigorous compliance standards.

3.3.2 Data Governance and Accountability

As organizations accelerate their adoption of AI, especially advanced systems such as **Large Language Models (LLMs)** and **Agentic AI** — the need for robust **data governance** and **accountability** becomes paramount. Governance ensures that data used in AI systems is properly managed, secure, compliant, and fit for purpose while defined accountability ensures that stakeholders take responsibility for how data is used, models are built, and decisions are made. Without well-defined governance and accountability mechanisms, organizations risk model failures, compliance violations, reputational damage, and the unchecked deployment of biased or unsafe AI. In this context, **governance is the policy scaffolding**, and **accountability is the ethical compass** guiding responsible AI use.

Modern AI systems not only consume vast datasets but also make autonomous decisions and continuously evolve through feedback loops. This raises complex questions around data ownership, traceability, consent, transparency, and liability, all of which must be addressed through a structured and scalable governance and accountability framework.

Objectives

The objectives of assessing **Data Governance and Accountability** within AI maturity are to:

1. **Establish Policy Control:** Ensure the organization has formal, enforceable data and AI governance policies, including roles, rules, and standards aligned with enterprise risk posture and regulatory requirements.
2. **Enable Traceability and Oversight:** Provide full lineage and traceability across datasets, models, and decisions, enabling effective auditability and root cause analysis for AI-driven outcomes.
3. **Define Ownership and Stewardship:** Clearly assign data and model ownership across teams, including responsibilities for stewardship, model explainability, bias monitoring, and error handling.
4. **Scale Governance for Modern AI:** Evolve data governance models to accommodate dynamic, unstructured, and third-party data sources as well as continuous learning systems like LLMs and agentic frameworks.

Streams

Maturity Level	Stream A	Stream B
1 - Establish an approach for data governance or accountability.	- **No Formal Policies:** Absence of defined policies or standards for data governance. - **Undefined Roles:** Roles and responsibilities for data stewardship and governance are unclear. - **Unstructured Governance:** Lack of governance processes specifically for AI datasets.	- **Undefined Ownership:** Data and AI model ownership unclear or not assigned. - **Documentation Gaps:** Absence of consistent model documentation or reliable audit trails. - **No Accountability:** AI outcomes lack clear accountability, oversight, and responsibility mechanisms.
2 - Define a formal governance structures with well defined roles	- **Basic Governance Charter:** Initial governance framework defined, outlining basic roles and responsibilities. - **Initial Stewardship:** Basic data stewardship roles identified, with	- **Partial Ownership Assignment:** Data owners identified for select datasets and models. - **Preliminary Documentation:** Initial attempts at systematic

Maturity Level	Stream A	Stream B
and accountability assigned.	preliminary metadata management. - **Policy Development:** Early stages of formal data usage policies.	model documentation and traceability. - **Informal Ethical Concerns:** Ethical and bias concerns acknowledged, though informally managed.
3 - Continuously improve the implementation with robust, enterprise-wide governance and accountability matrix.	- **Comprehensive Framework:** Mature governance framework implemented enterprise-wide, regularly reviewed and updated. - **AI-Specific Policies:** Detailed governance explicitly addressing AI training datasets, LLMs, agentic systems, and external data integration. - **Dynamic Adaptability:** Governance practices dynamically scale with evolving AI technology needs.	- **Enforced Accountability:** Clearly enforced accountability with responsible AI review boards overseeing model and dataset use. - **Incident Management:** Comprehensive incident tracking, documentation, and continuous audits for responsible AI practices. - **Full Traceability:** End-to-end traceability from data sourcing to model decisions, with explicit, accountable roles.

3.3.3 Data Training

High-quality training data is the backbone of effective AI systems. As organizations scale the use of AI, especially with Large Language Models (LLMs) and Agentic AI, ensuring the accuracy, security, and compliance of training datasets becomes critical. Poorly managed data can introduce bias, hallucinations, or model drift, undermining model performance and trust. Data training readiness goes beyond initial dataset collection. It involves structured curation, ongoing validation, governance of data sourcing, and monitoring for ethical and legal compliance. This is particularly important when using third-party, user-generated, or web-scraped data, which may carry legal, reputational, or privacy risks.

As AI systems become more autonomous, training data pipelines must evolve to support continuous improvement, integrating feedback loops, surfacing blind spots, and retraining models safely and responsibly.

Objectives

The objectives of assessing **Data Training** within an AI maturity model are to:

1. **Ensure Fitness for Purpose:** Validate that training data is accurate, representative, labeled appropriately, and aligned with the use case and model type (e.g., LLMs, multi-agent systems).
2. **Mitigate Risk and Bias:** Monitor training data for bias, drift, imbalances, or toxic content, reducing the risk of unintended or unethical AI behavior.
3. **Enable Secure and Ethical Use:** Ensure all datasets comply with privacy laws, licensing terms, and internal ethical AI standards, especially when using third-party or user-sourced data.
4. **Support Continuous Improvement:** Establish feedback loops to evolve datasets based on real-world performance, model errors, and evolving domain needs.
5. **Maintain Transparency and Auditability:** Enable clear traceability of data origins, transformations, and usage throughout the training pipeline to support audits and regulatory inquiries.

Streams

Maturity Level	Stream A	Stream B
1 - Establish a Training data management structure with documented processes and standards.)	- **Unstructured Collection:** Data gathered without structured processes, inconsistent quality. - **No Labeling Standards:** Absence of formal labeling guidelines or dataset curation practices. - **Manual Validation:** Minimal or no validation; data quality highly variable.	- **No Compliance Checks:** Lack of monitoring for compliance, bias, or security. - **Unchecked Data Use:** Third-party or user-generated data integrated without licensing or consent verification. - **Security Risk:** High risk of privacy breaches or ethical violations due to unmonitored datasets.
2 - Defined formal governance structure with guidelines and initial	- **Guidelines Established:** Initial standards for dataset collection, labeling, and validation set. - **Partial Validation:** Manual	- **Initial Privacy Checks:** Basic privacy and security compliance checks introduced. - **Licensing Awareness:** Increased awareness and

Maturity Level	Stream A	Stream B
compliance awareness in place.	validation and checks performed on subsets of training data. - **Early-stage Curation:** Early stages of data quality management and documentation established.	preliminary adherence to licensing and regulatory obligations. - **Bias Awareness:** Emerging processes to identify obvious bias or harmful content, though inconsistently applied.
3 - Continuously improve with fully structured, automated, and compliant training data management.	- **Automated Pipelines:** Standardized, automated pipelines for data preparation, quality control, deduplication, and labeling accuracy checks fully operational. - **Continuous Validation:** Real-time or regular validation ensuring high-quality, representative, and reliable training data. - **Dynamic Curation:** Active dataset curation based on model feedback, performance metrics, and evolving requirements.	- **Systematic Compliance:** Routine compliance audits for security, licensing, ethical use, and bias mitigation. - **Verified Usage Rights:** Comprehensive vetting and documentation for third-party and sensitive data usage. - **Robust Security Measures:** Secure data handling protocols with regular drift and toxicity monitoring, maintaining regulatory readiness and ethical standards.

3.4 Privacy

The **Privacy** pillar addresses the critical need to safeguard personal and sensitive information used by AI systems. AI's dependence on extensive data sets amplifies privacy risks, as data can inadvertently reveal personal details, lead to unauthorized profiling, or be subject to misuse. Proactive privacy management becomes essential, not only for regulatory compliance but also to maintain user trust.

This pillar is structured around three focused practices:

1. **Data Minimization and Purpose Limitation** – Ensuring AI systems collect and process only the data strictly necessary for clearly defined purposes.
2. **Privacy by Design and Default** – Embedding privacy measures deeply within the AI system lifecycle to mitigate risks proactively.

3. **User Control and Transparency** – Providing clear, understandable communication about data use and offering users meaningful controls over their data.

3.4.1 Data Minimization and Purpose Limitation

The Privacy practice addresses the critical need to safeguard personal and sensitive information used by AI systems. AI's reliance on extensive datasets amplifies privacy risks, such as inadvertent disclosure of personal details, unauthorized profiling, and potential misuse of data. Proactive privacy management is essential not only for regulatory compliance but also for maintaining user trust.

Objectives

1. **Data Minimization**: Ensure data minimization by collecting and processing only necessary data.
2. **Purpose Limitation**: Limit data use strictly to clearly defined and communicated purposes.
3. **Proactive Privacy**: Embed privacy measures proactively throughout the AI lifecycle.

Streams

Maturity Level	Stream A	Stream B
1 - Establish Privacy Principles and Policies clearly defining the scope and limits of data use.	- **Informal Approach:** Limited documentation of data collection and processing purposes. - **Reactive Management:** Privacy actions taken primarily after incidents or upon request. - **Undefined Responsibilities:** Privacy responsibilities not clearly assigned or formalized.	- **No Formal Monitoring:** Privacy compliance and data usage not regularly monitored. - **Incident-Based Learning:** Privacy improvements largely triggered by privacy incidents. - **Lack of Metrics:** Privacy metrics or assessments are informal or absent.
2 - Implement Structured Privacy Controls for data	- **Documented Policies:** Clear and comprehensive policies defining data minimization and purpose limitations.	- **Routine Monitoring:** Regular audits and reviews of data practices and compliance with privacy policies.

Maturity Level	Stream A	Stream B
minimization and clear purpose limitations.	- **Defined Accountability:** Specific roles (Privacy Officer, Data Steward) established with clear responsibilities. - **Planned Compliance:** Proactive privacy reviews integrated into AI project planning and execution.	- **Basic Metrics:** Privacy metrics (e.g., incident counts, data usage audits) routinely collected and reported. - **Proactive Adjustments:** Metrics inform adjustments to practices, reducing privacy risks and improving compliance.
3 - Embed Continuous Privacy Improvement into organizational culture and processes.	- **Fully Integrated Practices:** Privacy principles and policies deeply embedded in organizational workflows and practices. - **Strategic Alignment:** Privacy practices explicitly aligned with business objectives, ethics, and regulatory frameworks. - **Lifecycle Integration:** Continuous privacy impact assessments and controls throughout AI system development and operation phases.	- **Advanced Analytics:** Real-time monitoring and analytics of data usage, access, and compliance. - **Predictive Privacy Management:** Proactive identification and mitigation of privacy risks through predictive analytics and automated controls. - **Culture of Privacy Excellence:** Metrics drive organizational strategies, support transparency, foster user trust, and ensure regulatory compliance.

3.4.2 Privacy by Design and Default

Privacy by Design and Default (PbD&D) is a foundational principle for building trustworthy and compliant digital systems. It helps organizations assess and evolve their privacy practices from reactive to proactive implementation. It emphasizes integrating privacy into both governance frameworks and technical workflows. The objective is to ensure privacy is not an afterthought but a core design feature across the entire lifecycle. By progressing through the levels, organizations can move toward automated, measurable, and scalable privacy practices.

Objectives

1. **Privacy by Design**: Embed privacy principles into system design and development from the outset, rather than addressing them post-deployment.

2. **Governance and Accountability**: Establish clear roles, policies, and accountability for privacy management across teams and functions.
3. **Enablement through Tools and Patterns**: Equip engineering and design teams with reusable tools, patterns, and frameworks to implement privacy by default.
4. **Pipeline Integration**: Integrate privacy assessments and controls into development pipelines, CI/CD workflows, and governance reviews.
5. **Continuous Privacy Improvement**: Continuously monitor, measure, and improve privacy effectiveness using KPIs and automation.

Streams

Maturity Level	Stream A	Stream B
1 – Establish a privacy program addressing privacy risks and user compliant.	- **Ad Hoc Practices:** Privacy risks are addressed post-deployment and handled case-by-case. - **Missing Standards:** No standardized processes for data minimization, DPIAs, or policy application. - **Manual Communication:** Privacy notices and consents are manually generated, often retroactively.	- **No Privacy Engineering:** Developers and designers operate without privacy design patterns or reusable components. - **Lack of Tools:** No standard tools for consent, purpose limitation, or data classification. - **Reliance on Individuals:** Teams depend on personal initiative rather than embedded technical safeguards.
2 – Define a formal privacy program with well defined privacy practices, policies and assigned responsibilities.	- **Policy Adoption:** A Privacy by Design policy is published and adopted organization-wide. - **Assigned Roles:** Privacy Officers or Data Stewards are appointed to oversee compliance. - **Integrated Processes:** DPIAs and privacy reviews are integrated into product development and procurement lifecycles.	- **Reusable Components:** Privacy design patterns and libraries (e.g., consent modules, data masking APIs) are made available. - **Process Guidance:** Templates and checklists guide teams through privacy requirements in design and development phases. - **Shared Tooling:** Teams use shared SDKs for compliant data handling and user control mechanisms.

Maturity Level	Stream A	Stream B
3 – Continuously improve the Privacy program with automation and monitoring of key metrics.	- **Automated Governance:** DPIAs and approvals are integrated into CI/CD with automated gates. - **Code-Level Enforcement:** Data retention, access controls, and minimization are enforced via code. - **Data-Driven Review:** Privacy KPIs are reviewed quarterly and linked to org-wide OKRs.	- **Embedded PETs:** Privacy-enhancing technologies (PETs) like differential privacy and synthetic data are provided by default. - **Integrated Safeguards:** Privacy controls are embedded into design systems and dev workflows. - **Continuous Metrics:** Metrics on privacy defaults and user control coverage are continuously monitored and improved.

3.4.3 User Control and Transparency

User Control and Transparency are essential pillars of responsible AI and digital product design. They ensure users understand how their data is used and have meaningful agency over that use. This maturity model guides organizations in progressing from minimal disclosure to embedded, proactive transparency and user empowerment. It emphasizes both governance structures and practical implementation in design and engineering workflows. The goal is to build trust, comply with regulatory expectations, and respect user autonomy across the product lifecycle.

Objectives

1. **User Information Transparency**: Provide users with clear, accessible, and timely information about how their data and AI-powered features operate.
2. **User Control and Consent**: Enable meaningful user control over data use, personalization, consent, and algorithmic decisions.
3. **Governance for User Engagement**: Establish governance policies that define standards for disclosure, consent, and user engagement.
4. **Design Enablement for Transparency**: Equip product and engineering teams with design patterns, APIs, and UI components to support transparency and control.
5. **Feedback-Driven Improvement**: Continuously evaluate and improve user-facing transparency mechanisms through metrics and user feedback.

Streams

Maturity Level	Stream A	Stream B
1 – Establish User control practices and transparency mechanisms that are legally required.	- **Opaque Communication:** Disclosures are written in legal terms with limited accessibility. - **Generic Consent:** Consent mechanisms are generic and often bundled. - **Unclear Ownership:** No clear ownership for transparency or user agency.	- **Inconsistent UI:** UI elements for control (e.g. toggles, preferences) are ad hoc and hard-coded. - **No Design Standards:** No reusable components or design guidelines for transparency. - **Limited User Access:** Users cannot access or manage their data effectively.
2 – Defined Policies and workflows to standardize user control and disclosure practices.	- **Policy Enforcement:** A user transparency and control policy is published and enforced. - **Assigned Roles:** Roles (e.g., UX Privacy Leads or Product Compliance Liaisons) are assigned. - **Reviewed Consent Flows:** User consent flows are aligned with legal bases and reviewed periodically.	- **Standardized Interfaces:** Common UI patterns are introduced for preferences, opt-ins/outs, and data visibility. - **Process Integration:** Consent and disclosure flows are reviewed in design and development phases. - **Consistent Access:** APIs are used to give users access, edit, and delete data consistently.
3 – Optimized transparency and user control processes are embedded by default and continuously improved through feedback and automation.	- **Measured Transparency:** User transparency KPIs (e.g. consent clarity, user opt-out rates) are tracked across products. - **Live Consent Tracking:** Real-time consent and preference tracking is integrated with systems. - **Contextual Explanations:** User-facing explanations are tailored based on context and usage.	- **Adaptive Components:** Dynamic UI components adapt transparency and control options based on user needs. - **Feedback-Driven Design:** Feedback loops inform design updates based on user behavior and satisfaction. - **Comprehensive Control Panels:** Privacy dashboards and granular controls are standard in all user-facing systems.

3.5 Design

The **Design** pillar focuses on proactively integrating security and ethical considerations into the fundamental architecture and conceptualization of AI systems. AI introduces new vulnerabilities such as adversarial attacks and requires careful threat assessment and secure architecture decisions from the earliest stages.

To manage these challenges, this pillar includes three core practices:

1. **Threat Assessment** – Systematically identifying and addressing AI-specific threats, including adversarial manipulation and data poisoning.
2. **Security Architecture** – Designing robust and resilient architectures tailored to protect AI systems and their data, leveraging proven security principles like defense-in-depth.
3. **Security Requirements** – Clearly defining explicit AI-related security requirements at the design phase, informed by threat modeling outcomes.

3.5.1 Threat Assessment

The **Threat Assessment** practice addresses unique security, ethical, and operational risks associated with Large Language Models. Given their dynamic nature and extensive interaction with end-users, LLMs introduce specific vulnerabilities such as prompt injection, data leakage, and harmful or unethical outputs. This practice aims to proactively identify, assess, and mitigate these threats systematically, ensuring LLMs are secure, trustworthy, and aligned with organizational values.

Objectives

1. **LLM-Specific Threat Mitigation**: Identify and mitigate threats unique to LLMs (e.g., OWASP Top 10 for LLM Applications).
2. **Mitigation Strategies**: Align mitigation strategies with organizational values and compliance goals.
3. **Security Governance**: Integrate threat insights into broader AI governance and oversight processes.

Streams

Maturity Level	Stream A	Stream B
1 – Establish threat assessment process with identification of LLM-specific Risks	- **High-Level Risks Identified:** Initial identification and acknowledgment of broad risks (e.g., data leakage, unethical or harmful outputs). - **Ad Hoc Documentation:** Risks are documented informally, without standardized structures or severity ratings. - **Limited Stakeholder Awareness:** General awareness among stakeholders regarding potential risks, but no systematic tracking.	- **Use of Basic Checklists:** Teams utilize basic threat checklists (e.g., OWASP Top 10 for LLM Applicationss) to identify common issues like prompt injection or sensitive data exposure. - **Informal Approach:** Threat identification relies primarily on manual, informal processes. - **Limited Coverage:** Threat assessments cover only selected or high-visibility LLM deployments.
2 – Define processes for Centralized and Standardized Risk Management.	- **Centralized Risk Inventory:** Established and maintained comprehensive risk inventory specific to LLM use cases, detailing vulnerabilities such as adversarial attacks, prompt manipulation, and ethical concerns. - **Severity Scores:** Risks assigned severity scores based on potential impact, likelihood, and organizational context. - **Regular Updates:** Risk inventories updated periodically or when significant changes in LLM use cases occur.	- **Standardized Threat Modeling Process:** Organization-wide standardized approach to threat modeling, clearly mapping adversarial attack vectors such as prompt injection, unauthorized data disclosure, and unethical content generation. - **Structured Documentation:** Threat models documented systematically and reviewed regularly. - **Integrated into Development:** Threat modeling integrated into the design phase of LLM projects.
3 – Continuously improve the process with automated and Proactive Risk Detection.	- **Automated Risk Monitoring:** Continuous, automated detection and monitoring of LLM outputs for potentially harmful content, data leakage, and security anomalies. - **Real-time Alerting:**	- **Full Automation of Threat Detection:** AI-driven tools automatically detect adversarial attempts, prompt injection attacks, and other security threats in real-time.

Maturity Level	Stream A	Stream B
	Automated alerts triggered by identified risks, facilitating immediate investigation and mitigation. - **Continuous Improvement:** Risks dynamically reassessed through continuous monitoring and real-time data analytics.	- **Integrated Alerts into Operational Tools:** Threat detection integrated into operational and incident response systems (e.g., SIEM, SOAR). - **Predictive Analytics:** AI-assisted predictive analytics anticipate new or evolving threats based on historical data and emerging trends.

3.5.2 Security Architecture

The **Security Architecture for AI** practice focuses on designing and implementing robust, secure infrastructure specifically tailored for deploying and monitoring artificial intelligence systems. AI systems, due to their complexity, evolving threat landscape, and dynamic operational characteristics, require tailored architectural safeguards to mitigate vulnerabilities, ensure reliable operation, and enable rapid incident response. By embedding secure deployment patterns and comprehensive monitoring within the infrastructure, organizations can proactively address threats and maintain continuous protection of their AI-based systems.

Objectives

1. **Secure Infrastructure and Monitoring**: Design secure infrastructure for AI deployment and continuous monitoring.
2. **Realtime threat detection**: Embed secure deployment patterns and monitoring mechanisms to enable real-time threat detection and rapid incident response.
3. **Secure Workloads**: Ensure infrastructure resilience and integrity by mitigating vulnerabilities and supporting continuous, secure operation of AI workloads.

Streams

Maturity Level	Stream A	Stream B
	- **Basic Isolation & Access Control:** Implement fundamental	- **Baseline Security Features:** Utilize frameworks, libraries, and

Maturity Level	Stream A	Stream B
1 – Initial Secure Practices	security measures such as authentication and rate-limiting to secure AI APIs, aligned with industry standards and best practices. - **Limited Runtime Protection:** Initial protections mainly focused on basic perimeter defenses and simple access restrictions.	platforms with built-in security functionalities and protections. - **Informal Selection Criteria:** Basic awareness in selecting technology stacks that provide foundational security capabilities.
2 – Standardized Deployment Safeguards	- **Runtime Guardrails:** Deploy comprehensive runtime guardrails including output sanitization and input validation to mitigate common vulnerabilities (e.g., OWASP Top 10 for LLM Applications). - **Structured Deployment Processes:** Standardize deployment procedures to ensure consistent application of security controls across all AI environments.	- **Standardized Monitoring & Observability:** Implement standardized monitoring tools that track performance, observability, and key security metrics, providing clear visibility into AI operational health. - **Regular Metrics Review:** Structured review processes established for ongoing monitoring and maintenance of technology stack security.
3 – Advanced and Proactive Defenses	- **AI-Driven Adversarial Detection:** Integrate advanced, AI-driven anomaly detection and adversarial monitoring capabilities into deployment environments, proactively identifying and addressing threats in real-time. - **Model Versioning & Rollback:** Implement model versioning with swift rollback mechanisms to enable rapid incident recovery and response, particularly relevant for private or fine-tuned deployments.	- **Automated Patch Management & Scanning:** Fully automate vulnerability scanning and patch management processes, regularly reviewing and securing all dependencies within the technology stack. - **Continuous Improvement Cycles:** Establish continuous review cycles, automatically adapting security practices in response to emerging threats and updated security intelligence.

3.5.3 Security Requirements

The **Security Requirements** practice ensures that AI systems are designed with clear, comprehensive guidelines addressing ethical, legal, and technical risks. Unlike traditional software, AI systems introduce unique challenges, including ethical considerations, compliance with emerging regulations, and vulnerabilities arising from complex data dependencies and model behaviors. Establishing explicit, documented security requirements early in the design phase helps mitigate these risks, supporting responsible innovation and secure AI deployments.

Objectives

1. **Risk-Based Requirement Definition**: Define explicit requirements to address ethical, legal, and technical AI risks.
2. **Secure by Design**: Integrate security considerations early in the AI system design lifecycle.
3. **Compliance with regulations**: Ensure compliance with emerging AI regulations and governance standards.

Streams

Maturity Level	Stream A	Stream B
1 – Baseline Documentation of Requirements	- **Baseline Ethical Guidelines:** Document foundational ethical guidelines addressing bias, fairness, transparency, and compliance standards (e.g., GDPR, EU AI Act). - **Basic Compliance Measures:** Initial strategies for meeting regulatory requirements (e.g., data privacy, user consent). - **General Awareness:** Stakeholders have basic awareness of ethical and compliance obligations.	- **Basic Data Provenance:** Document initial sources of training data and maintain basic data lineage records. - **Manual Tracking:** Data provenance records are manually created and updated, with limited standardization or automation. - **Limited Visibility:** Partial visibility into third-party data and model components.
	- **Standardized Bias & Fairness Tools:** Implement	- **Automated Quality Checks:** Automate validation processes for

Maturity Level	Stream A	Stream B
2 – Standardized Implementation and Validation	standardized tools for bias detection and fairness measurement within training pipelines and application outputs. - **Integrated Compliance Processes:** Consistent application of compliance controls (e.g., automated checks for GDPR compliance, consent verification). - **Structured Documentation:** Ethical and compliance measures systematically documented and regularly reviewed.	third-party datasets and AI models, including quality assurance and security assessments. - **Enhanced Provenance Records:** Automated maintenance of detailed data lineage and provenance documentation, ensuring traceability and accountability. - **Structured Validation:** Standardized criteria established for acceptance of third-party components.
3 – Automated and Continuous Compliance Assurance	- **Real-Time Compliance Monitoring:** Automated compliance checks integrated throughout AI system lifecycles, with real-time audit trails and immediate alerting mechanisms. - **Expert Human Oversight:** Complex compliance decisions trigger expert human review to balance automation with accountability. - **Predictive Compliance Management:** Utilize predictive analytics to proactively identify emerging compliance and ethical risks.	- **Real-Time Provenance Tracking:** Real-time capture and automated management of comprehensive data and model provenance across all lifecycle stages, from initial sourcing through deployment. - **Advanced Provenance Analytics:** Integrate analytics to proactively detect anomalies, unauthorized changes, or potential security risks within data and model workflows. - **Continuous Provenance Auditing:** Automatically generate detailed audit trails, enabling immediate and transparent reporting for governance, compliance, and incident response.

3.6 Implementation

The **Implementation** pillar ensures that secure, ethical, and resilient practices are embedded throughout the AI system development lifecycle.

Unlike traditional software, AI systems introduce additional complexities such as dynamic model behavior, data-driven vulnerabilities, and evolving threat landscapes. These systems often make autonomous decisions based on probabilistic outputs, increasing the potential for unpredictable or unintended behavior. Therefore, organizations must adopt tailored implementation practices that address these unique risks at every phase—from data handling and model training to deployment and post-deployment management. Furthermore, the integration of external AI services and pre-trained models demands heightened scrutiny regarding supply chain integrity and continuous performance monitoring.

The Implementation pillar is organized into three interconnected practices that reflect the critical stages where AI systems must be safeguarded:

1. **Secure Build**: Integrating secure development practices, data hygiene, supply chain validation, and responsible AI principles during model creation, selection, and integration. This includes ensuring that input/output behaviors are well understood and that components are sourced, configured, and validated securely.
2. **Secure Deployment**: Protecting AI models and associated data throughout the deployment process by ensuring operational resilience, maintaining confidentiality, safeguarding integrity, and monitoring system behavior under live conditions. This phase also emphasizes the need for governance, rollback planning, and compliance verification.
3. **Defect Management**: Establishing ongoing processes for the systematic identification, prioritization, and mitigation of vulnerabilities, performance issues, and ethical risks within AI systems. Defect management is crucial for maintaining long-term trustworthiness and ensuring that AI behaviors align with organizational values and regulatory expectations.

By adopting these practices, organizations can proactively mitigate risks related to model misuse, bias, adversarial manipulation, and data leakage. In doing so, they not only achieve compliance with evolving legal and ethical standards but also build durable trust with users, partners, and regulators across diverse application domains.

3.6.1 Secure Build

Secure Build practices form the foundation for trustworthy AI by ensuring that risks are addressed early—during model selection, development, and integration. Unlike traditional software builds, AI systems depend on data quality, third-party model provenance, and probabilistic behaviors that must be explicitly controlled. A secure build process incorporates defensible supply chain decisions, ethical considerations, and reproducible configurations. It mandates that all models—whether pre-trained or custom—are assessed not only for technical performance but also for licensing, robustness, and alignment with intended use. Effective implementation includes integrating automated security scans, adversarial robustness checks, and validation mechanisms into development pipelines. These controls help prevent the downstream amplification of vulnerabilities and reduce bias propagation. They also ensure the system remains observable, verifiable, and secure before deployment.

Objectives

- **Promote Secure Foundations**: Ensure responsible sourcing, secure coding, and defensible supply chain decisions are embedded in build practices.
- **Establish Model Accountability**: Verify licensing, purpose alignment, and robustness for all models before integration.
- **Automate Trust Checks**: Incorporate reproducibility, adversarial robustness, and validation into automated pipelines.

Streams

Maturity Level	Stream A	Stream B
1 – Establish awareness with Governance and controls for foundation framework implementation.	- **Ad hoc Model Selection**: Model sources selected without standard criteria. - **Lack of Inventory**: Inventory is informal or outdated. - **Missing Provenance**: Purpose and provenance of models are rarely documented.	- **Unchecked Licensing**: License terms and dependencies rarely verified. - **Vulnerability Gaps**: Known vulnerabilities not consistently scanned. - **No Tooling**: No formal toolchain for validation.
2 – Defined Practices with	- **Secure Guidelines**: Secure development guidelines include	- **I/O Controls**: Input/output sanitization in place.

Maturity Level	Stream A	Stream B
security and governance practices are being documented and implemented.	AI-specific considerations. - **Basic Model Review**: Model reviews include basic ethical and compliance checks. - **Inventory Control**: Inventory management is standardized but not automated.	- **Versioning**: Models and datasets version-controlled. - **Initial Validation**: Basic output validation initiated.
3 – Continuously Manage Risk with proactive governance and supply chain-level awareness.	- **Formal Risk Reviews**: Formal risk assessments conducted for third-party and internal models. - **Custody Controls**: Custody of AI assets is tracked and managed. - **Supplier Assurance**: Attestations and compliance documents are requested from providers.	- **Adversarial Testing**: Adversarial testing is routinely performed. - **CI/CD Integration**: AI checks are integrated into CI/CD pipelines. - **Edge Case Validation**: Behavior under edge cases is validated.

3.6.2 Secure Deployment

Secure deployment of AI requires organizations to recognize that models do not behave as static software artifacts but evolve over time. Therefore, deployment strategies must account for model drift, environmental changes, and real-world adversarial conditions. Continuous operational monitoring and governance are vital to ensure that deployed AI systems remain effective, fair, and secure throughout their lifecycle.

Objectives

1. **Establish clear documentation and traceability** for AI model configurations, environments, and runtime dependencies to support operational visibility.
2. **Implement structured approval workflows and rollback procedures** to ensure safe, auditable, and accountable AI deployments.
3. **Secure deployed AI assets** by enforcing access controls, logging access events, and encrypting sensitive model data.
4. **Integrate legal and regulatory compliance into deployment pipelines** through automated checks and continuous documentation readiness.

5. **Enhance resilience and reliability** by detecting model drift, monitoring for hallucinations, and triggering real-time alerts during high-risk behavior.

Streams

Maturity Level	Stream A	Stream B
1 – Foundational Deployment Practices with focus on basic documentation and monitoring.	**Environment Capture:** Document deployment configurations and runtime environments. **Dependency Logging:** Record libraries, dependencies, and versions. **Manual Tracking:** Maintain basic records without automation.	**Basic Monitoring:** Track model performance over time. **I/O Logging:** Log inputs and outputs for traceability. **Usage Metrics:** Collect simple metrics (e.g., invocation count, latency).
2 – Structured Deployment Governance with deployment governed by formal processes and access protections.	**Approval Workflows:** Define clear steps for review and sign-off before deployment. **Rollback Plans:** Establish mechanisms to revert to a prior version safely. **Audit Trails:** Log and store deployment decisions for traceability.	**Access Restrictions:** Implement role-based access control for deployed models. **Access Logging:** Record and monitor access to model endpoints. **Encryption:** Secure model artifacts and sensitive data at rest.
3 – Proactive and Compliant Operations with continuous compliance and resilience mechanisms integrated into operations.	**Compliance Checks:** Regularly assess for legal, regulatory, and policy alignment. **Automation:** Integrate compliance checks into CI/CD workflows. **Audit Readiness:** Maintain documentation for regulatory or internal audits.	**Resilience Design:** Build fallbacks or safe shutdown options. **Drift Detection:** Monitor models for data or performance drift. **Alerting Systems:** Trigger real-time alerts for hallucinations or anomalies.

3.6.3 Defect Management

Effective defect management in AI systems demands more than traditional bug tracking. It must extend to monitoring for ethical failures (e.g., biased outputs), reliability issues

(e.g., hallucinations), and security vulnerabilities (e.g., model inversion attacks). Additionally, organizations should aim for proactive identification of emerging risks through user feedback, anomaly detection, and adaptive testing strategies that evolve alongside the deployed AI solutions.

Objectives

1. **Defect and Failure Tracking**: Create systematic approaches to identify, categorize, and track AI-specific defects and failure modes across the entire model lifecycle.
2. **Maturity-Based Capability Building**: Develop organizational capabilities from basic monitoring to advanced automated response systems through structured maturity levels.
3. **Human-Automation Balance**: Balance human-driven process improvements with automated technical solutions for comprehensive AI quality coverage.
4. **Predictive Quality Assurance**: Transition from reactive issue handling to predictive quality assurance through advanced analytics and automated monitoring.
5. **Closed-Loop Learning Systems**: Establish closed-loop systems that learn from defects and automatically enhance model performance and quality processes.

Streams

Maturity Level	Stream A	Stream B
1 - Establish Foundational Quality Practices that enable consistent defect tracking, basic monitoring, and awareness of model reliability risks.	**Defect Taxonomy** Define and adopt a standard taxonomy for AI defects and failure modes. **Basic Tracking** Begin tracking model behavior issues and performance degradation. **Initial Documentation** Log known issues and defects manually for future reference.	**User Feedback Monitoring** Deploy basic systems to capture user-reported issues. **Regression Testing** Perform regression tests after model updates. **Alerting for Failures** Create simple alerting for obvious or repeated model errors.
2 – Integrate AI Defect	**Defect Prioritization** Score defects based on impact and	**Advanced Testing** Implement targeted tests for edge

Maturity Level	Stream A	Stream B
Prioritization and Testing into QA processes to optimize quality insights, fairness, and model reliability.	severity. **Workflow Integration** Embed defect tracking into QA and release processes. **Defect Analytics** Analyze trends and patterns across logged AI defects.	cases, fairness, and bias. **Scheduled Reevaluation** Routinely test model behavior in varied deployment contexts. **Controlled Experiments** Use A/B testing to validate model improvements.
3 – Achieve Advanced AI Quality Assurance through automation, root cause analysis, and adaptive learning systems.	**Root Cause Analysis** Investigate failures at data, training, and architecture levels. **Knowledge Sharing** Document and share lessons learned in a knowledge base. **Cross-Functional Review** Form teams across roles to analyze complex failures.	**Automated Pipelines** Deploy retraining and rollback pipelines for rapid response. **Real-Time Monitoring** Implement anomaly detection for live model performance. **Closed-Loop Learning** Enable self-correcting systems that learn from defect signals.

3.7 Verification

The **Verification** pillar addresses the unique challenges of validating and testing AI systems to ensure their security, functionality, and ethical compliance. AI's complex, dynamic behaviors demand specialized verification approaches, ongoing assessments, and rigorous testing processes.

This pillar is structured around three critical practices:

1. **Security Testing** - Conducting specialized assessments of AI systems, including adversarial attacks, robustness testing, and model poisoning resilience.
2. **Requirement-based Testing** - Verifying systematically that AI systems meet defined functional, security, and ethical requirements throughout their lifecycle.
3. **Architecture Assessment** - Regularly reviewing AI system architectures to ensure they conform to best practices, industry standards, and emerging security guidelines.

3.7.1 Security Testing

The Security Testing practice is essential for proactively identifying and mitigating security vulnerabilities in AI systems. By systematically performing specialized security assessments tailored for AI, organizations can strengthen their defense against evolving threats such as adversarial attacks, model poisoning, and robustness failures. These security tests ensure AI systems operate safely, securely, and ethically.

Objectives

1. **Adversarial Testing**: Identify vulnerabilities through targeted adversarial testing.
2. **Model Resilience**: Ensure resilience against model poisoning and robustness failures.
3. **Security-Driven Improvement**: Foster continuous improvement in AI security practices based on security insights.

Streams

Maturity Level	Stream A	Stream B
1 - Identify the need for establishing a framework of basic security testing	- Ad hoc security tests with no systematic approach. - Reactive security activities triggered mainly by incidents. - Limited understanding of AI-specific threats.	- No formal security metrics defined or tracked. - Security insights derived primarily from incident response. - Inconsistent or irregular reporting.
2 - Define a proper framework with defined policies, processes and procedures	- Structured AI security testing approach established (adversarial tests, robustness evaluations). - Defined responsibilities for conducting regular AI security assessments. - AI security activities integrated into broader security testing efforts.	- Defined security metrics (incident frequency, robustness indicators, resilience scores). - Regularly collected and reported security metrics for stakeholder visibility. - Metrics guide security improvements and resource allocation.
3 - Continuously Optimize the	- Comprehensive security testing integrated throughout the AI	- Real-time monitoring and advanced analytics detecting AI-

Maturity Level	Stream A	Stream B
processes with monitoring and metrics reporting	lifecycle. - Advanced threat simulations (continuous adversarial testing, proactive poisoning resistance evaluation). - Dedicated AI security team actively adapting to emerging threats.	specific security threats. - Predictive metrics forecasting vulnerabilities and proactively addressing them. - Robust, continuous feedback loop driving strategic security enhancements and resource decisions.

3.7.2 Requirement-based Testing

The AIMA Requirement-based Testing practice ensures AI systems consistently align with defined functional, security, and ethical requirements throughout their lifecycle. By systematically verifying these requirements, organizations ensure reliability, compliance, and ethical integrity of AI deployments, enabling informed and responsible AI use.

Objectives

1. **Standards Compliance Verification**: Verify compliance with functional, security, and ethical standards.
2. **Requirements Traceability**: Maintain clear traceability between requirements and testing activities.
3. **Insight-Driven Improvement**: Continuously improve based on requirement-driven insights.

Streams

Maturity Level	Stream A	Stream B
1 - Foundational Testing Practices Emerging with siloed practices	- Testing is informal or inconsistently linked to requirements. - Requirement traceability is limited or non-existent. - Testing often reactive rather than planned.	- Minimal or no metrics related to requirement testing. - Testing results documented irregularly. - Limited stakeholder visibility into testing outcomes.

Maturity Level	Stream A	Stream B
2 - Define a documented process with documented guidelines.	- Formal testing process established with clear links to defined requirements. - Responsibility for requirement-based testing clearly assigned. - Regular execution of testing aligned with the AI lifecycle.	- Defined metrics (coverage, requirement compliance rates, defect rates). - Regular reporting of metrics to stakeholders. - Metrics inform decisions and drive continuous improvement.
3 - Coninously improvement focused approach with monitoring and metric reporting.	- Requirement-based testing fully integrated into continuous development and deployment processes. - Automated and continuous verification against requirements. - Active use of feedback to refine testing and requirement definitions.	- Advanced analytics to continuously track and analyze requirement compliance. - Predictive metrics anticipate issues proactively. - Strong culture of accountability and continuous enhancement driven by detailed, actionable metrics insights.

3.7.3 Architecture Assessment

The AIMA Architecture Assessment practice ensures AI system architectures consistently adhere to best practices, industry standards, and emerging security guidelines. Regular assessments help identify architectural weaknesses early, fostering robust, secure, and ethically compliant AI solutions.

Objectives

1. **Standards-Based Architecture Validation**: Regularly validate AI architectures against evolving standards and best practices.
2. **Architectural Vulnerability Remediation**: Identify and remediate architectural vulnerabilities.
3. **Architecture Refinement through Assessment**: Continuously refine AI system architectures based on assessment insights.

Streams

Maturity Level	Stream A	Stream B
1 - Initial Steps Toward AI Architecture Governance with basic processes and practices in place	- Architecture reviews informal or ad hoc. - Limited awareness of AI-specific architecture standards. - Reactive to architecture-related incidents rather than proactive assessments.	- Few or no metrics related to architectural quality or security. - Irregular documentation of assessment outcomes. - Limited stakeholder engagement or reporting.
2 - Structured and Integrated AI Architecture Governance with defined processes and guidelines	- Defined architecture review process integrated into AI projects. - Clearly assigned responsibilities for architecture assessments. - Regular architecture evaluations aligned with lifecycle milestones.	- Established metrics (compliance with architectural guidelines, identified vulnerabilities, remediation rates). - Routine reporting to stakeholders. - Metrics actively guide architecture improvements.
3 - Continuous and Adaptive AI Architecture Excellence with monitoring and metrics reporting	- Comprehensive and continuous architecture assessment embedded in the AI lifecycle. - Proactive identification and remediation of architectural vulnerabilities. - Active adaptation to emerging AI architectural best practices and guidelines.	- Advanced metrics and analytics for real-time architectural monitoring. - Predictive analytics proactively identifying potential architectural weaknesses. - Strong organizational commitment to continuous architectural refinement driven by actionable metrics insights.

3.8 Operations

The **Operations** pillar focuses on securely maintaining AI systems after deployment. Given the complexity and continuous nature of AI operations—including model updates, real-time inference, and dynamic data flows—organizations require dedicated operational practices to manage risks and ensure ongoing reliability, security, and ethical integrity.

This pillar includes three primary practices:

1. **Incident Management** – Establishing rapid-response protocols and capabilities to effectively handle and mitigate security incidents involving AI systems, such as data breaches or model manipulation.
2. **Event Management** – Continuously monitoring AI systems for anomalous behavior, performance degradation, or unexpected outcomes to proactively identify issues.
3. **Operational Management** – Managing and maintaining AI deployments securely, responsibly, and sustainably, ensuring they operate effectively and ethically over time.

3.8.1 Incident Management

Incident Management within AI systems is essential for promptly addressing security breaches, unexpected behaviors, or system failures. AI systems, including machine learning (ML) and large language models (LLMs), introduce unique challenges like adversarial attacks, model manipulation, and emergent biases. Traditional incident management approaches may not sufficiently address these complexities. Effective Incident Management ensures security, reliability, and ethical alignment throughout the AI system lifecycle.

Objective

1. **Swift Detection and Containment:** Rapidly identify and contain vulnerabilities and incidents.
2. **Root Cause Analysis and Mitigation:** Investigate incidents thoroughly and implement measures to prevent recurrence.
3. **Transparent Communication and Reporting:** Maintain clear and timely communication with stakeholders, regulators, and users.
4. **Continuous Improvement:** Use insights from incidents to enhance security protocols and improve response strategies.

Streams

Maturity Level	Stream A	Stream B
1: Establish Initial AI Incident Detection and Basic Response Capabilities	**Reactive Detection** - Basic incident detection with reactive responses. **Ad Hoc Containment** - Limited formal processes for incident containment. **Minimal Analysis** - Initial triage without deep forensic investigation.	**Informal Reporting** - Incident reporting is informal with minimal stakeholder communication. **Limited Communication** - Stakeholder engagement is minimal or ad-hoc. **Sparse Post-Incident Review** - Post-incident reviews are limited or informal.
2: Manage and Standardize AI Incident Handling and Post-Incident Evaluation	**Standardized Protocols** - Established protocols for detection, containment, and initial analysis. **Defined Roles** - Clear roles and responsibilities in incident response teams. **Consistent Workflows** - Repeatable incident handling processes.	**Structured Communication** - Formal communication protocols with key stakeholders. **Regular Reviews** - Scheduled post-incident reviews with documented outcomes. **Tracked Improvements** - Outcomes and lessons learned are documented and tracked.
3: Advance to Real-Time Detection and Continuous Learning from AI Incidents	**Automated Detection** - Automated detection systems leveraging real-time analytics. **Integrated Forensics** - Comprehensive forensic analysis integrated into workflows. **Adaptive Response** - Incident response evolves based on root cause and threat intelligence.	**Proactive Notifications** - Automated and timely notifications to stakeholders. **Detailed Reporting** - Full incident reports including impact and response evaluations. **Continuous Improvement** - Systematic improvements driven by incident data and emerging threats.

3.8.2 Event Management

Event Management involves structured detection, response, and learning processes for anomalies, failures, and deviations in AI systems. Effective Event Management ensures

that incidents related to model behavior, infrastructure, or data flow are rapidly identified, thoroughly investigated, transparently documented, and leveraged for continuous improvement.

Objective

1. **Real-Time Detection:** Quickly identify model drift, unfair predictions, infrastructure failures, and compliance violations.
2. **Structured Response:** Implement standardized procedures for incident handling and resolution.
3. **Transparency and Accountability:** Maintain thorough records and audit trails for all AI-related incidents.
4. **Continuous Learning:** Utilize incident insights to enhance models, processes, and deployment safeguards.
5. **Minimize Impact:** Limit business disruptions and reputational risks resulting from AI system issues.

Streams

Maturity Level	Stream A	Stream B
1: Establish Basic Monitoringand Ad Hoc Response Capabilities to detect issues manually and react without formal processes.	**Manual Detection** - Events are identified manually, often after impact is observed. **No Anomaly Detection** - No structured methods for identifying drift, outliers, or degradation. **Reactive Approach** - Monitoring is not proactive or automated.	**Ad Hoc Management** - Incidents are handled reactively without structured processes. **No Documentation** - Incidents are rarely logged or reviewed systematically. **Lack of Learning** - No mechanisms in place for organizational learning from incidents.
2: Develop Structured Monitoring and Initial Learning Mechanisms to track key metrics, detect	**Basic Monitoring** - Latency, availability, and accuracy metrics are tracked. **Initial Anomaly Detection** - Basic drift and outlier detection introduced. **Alerting Setup**	**Incident Logging** - Incidents are logged and tracked manually. **Occasional RCA** - Root cause analysis is performed inconsistently. **Partial Documentation**

Maturity Level	Stream A	Stream B
anomalies, and analyze incidents.	- Manual or threshold-based alerting in place.	- Some lessons learned are captured but not systematized.
3: Achieve Proactive, Intelligent Monitoring and Continuous Learning by automating detection and integrating improvements from incident insights.	**Real-Time Monitoring** - Continuous monitoring with dashboards and alerting tools. **ML-Driven Detection** - Advanced analytics and machine learning detect anomalies and drift proactively. **Proactive Alerts** - Intelligent alerting reduces false positives and accelerates response.	**Comprehensive Workflows** - Formal incident response workflows across teams. **Systematic RCA** - Structured root cause analysis feeds into quality improvements. **Continuous Learning Loop** - Learnings are documented, shared, and integrated into system design.

3.8.3 Operational Management

Operational Management encompasses processes and practices aimed at maintaining consistent, reliable, and secure AI system operations. Effective Operational Management ensures that AI systems perform optimally, securely, and continuously meet compliance and business objectives.

Objective

1. **Operational Stability:** Maintain reliable and consistent performance of AI systems.
2. **Security Compliance:** Ensure operations adhere to established security standards and compliance requirements.
3. **Effective Resource Management:** Optimize resource allocation and usage to sustain system performance and efficiency.
4. **Proactive Maintenance:** Implement regular maintenance schedules to prevent operational disruptions and maintain system health.
5. **Adaptability and Scalability:** Manage operations flexibly to accommodate changes, updates, and scalability requirements.

Streams

Maturity Level	Stream A	Stream B
1- Early-stage capabilities with initial monitoring, basic controls, and emerging awareness.	**Manual Monitoring** - Monitoring is ad hoc or manual, lacking structured visibility into system health. **Reactive Maintenance** - Maintenance occurs only after failures or disruptions. **Limited Coverage** - No proactive checks or resource planning in place.	**Basic Compliance Awareness** - Security and compliance concepts are understood but not formalized. **Ad Hoc Checks** - Security reviews are sporadic and undocumented. **Minimal Documentation** - Few written procedures or audit trails exist.
2-Developing structured processes and growing automation foster improved reliability and accountability.	**Scheduled Monitoring** - Regular system health checks and performance metrics are collected. **Preventive Maintenance** - Maintenance activities are performed on a routine schedule. **Improved Stability** - Operational disruptions are reduced due to consistent upkeep.	**Standardized Security Practices** - Security controls are documented and partially automated. **Regular Audits** - Periodic audits and compliance checks are initiated. **Policy Alignment** - Processes begin aligning with regulatory and organizational requirements.
3-Mature advanced systems driving resilience, compliance, and continuous performance improvement.	**Automated Monitoring** - Real-time, automated alerts with predictive performance and failure analysis. **Continuous Optimization** - Systems are tuned continuously for uptime and efficiency. **Proactive Resource Management** - Resource scaling and tuning are managed through automated tools.	**Automated Compliance Enforcement** - Continuous compliance monitoring is fully automated. **Integrated Security Audits** - Routine, detailed security audits with full traceability. **Proactive Threat Mitigation** - Threat detection and response are integrated into daily operations.

4 Applying the Model

Using the AIMA Maturity Levels

The OWASP AI Maturity Assessment model builds on OWASP SAMM's approach, defining three **maturity levels** for each practice in the 8 Business Functions domains with detailed components. The AIMA Business Functions are described in the chapter above: Responsible AI Principles, Governance, Data Management, Privacy, Design, Implementation, Verification and Operation.

Each Security Practices inside at each of the 8 Business Functions is described in terms of the following components:

- **Objective:** A general goal statement that captures the assurance aim of achieving that level for the AI governance practice. As levels increase, the objectives become more sophisticated in terms of establishing robust, responsible AI practices within the organization. In other words, each level's objective articulates the higher-level outcome (e.g. basic AI governance established at Level 1, a comprehensive AI strategy with metrics by Level 2, continuous improvement and optimization by Level 3).

- **Activities:** The core requisites or activities that must be performed to attain the level. Some activities are organization-wide (e.g. establishing an AI oversight committee or policy), while others may be project-specific (e.g. conducting risk assessments for each AI project). These activities capture the essential *AI governance functions*, but teams have flexibility in how they implement them. For example, a Level 1 activity in **Strategy & Metrics** might be *defining an initial AI strategy document*, whereas by Level 2 activities could include *implementing metrics tracking for AI system performance and risk*.

- **Results:** The capabilities or deliverables obtained by achieving the level. Results could be concrete artifacts or processes (such as an AI risk register, compliance reports, training programs) or qualitative improvements in capacity. For instance,

a Level 1 result in **Policy & Compliance** may simply be *awareness of legal requirements*, while a Level 2 result might be *established AI policies and periodic compliance audits*, and Level 3 could yield *fully integrated compliance evidence management and continuous monitoring*.

In the next chapter we will describe how to perform the assessment using the AIMA Model in order to evaluate the maturity of the target Company.

Conducting AIMA Assessments

Assessing an organization using AIMA is very similar to the SAMM assessment methodology, but focused on AI systems and practices. By measuring the organization against AIMA's defined Governance practices, one can create an overall picture of the **AI governance and assurance activities** in place. This helps in understanding the current breadth of responsible AI measures and in planning a roadmap for improvement. As with SAMM, there are two recommended styles for conducting an AIMA assessment:

- **Lightweight Assessment:** Use the AIMA assessment worksheets for each practice (e.g. Strategy & Metrics, Policy & Compliance, Education & Awareness for the Governance Business Function) to answer a series of yes/no questions. Each practice's worksheet covers key activities or criteria at each maturity level. Based on the responses, assign a provisional maturity level score for each practice. This lightweight approach is usually sufficient for an organization looking to quickly map its existing AI governance efforts onto the AIMA model and get a high-level view of where they stand. For example, a team might answer the questions for **Strategy & Metrics** and determine that they meet all Level 1 criteria and some Level 2 – giving them an initial score of 1+ (as explained below). The lightweight assessment is quick and can often be done via interviews and document reviews, without deep verification.

- **Detailed Assessment:** This goes a step further by incorporating verification and evidence gathering after the initial questionnaire. Once the worksheets are completed, the assessors perform additional **audit** activities to confirm that the prescribed AIMA activities at each level are truly in place (not just "paper compliance"). For instance, if a Level 2 activity requires *regular AI model risk assessments*, a detailed assessment might involve reviewing a sample of project documents or interviewing staff to ensure those risk assessments are happening with

the intended quality. Moreover, AIMA (like SAMM) provides **Success Metrics** for each practice, so a detailed assessment will also involve collecting data on those metrics to see if performance meets expectations. In short, the detailed approach validates the answers given in the lightweight step and requires evidence (e.g. policy documents, training records, model evaluation reports), giving a higher confidence in the accuracy of the maturity rating.

Scoring: Scoring in AIMA follows the SAMM scoring model. After answering the yes/no questions in a practice's worksheet, you determine the maturity level achieved for that practice. In general, if an organization answers "Yes" to all questions up to a certain level's marker, it achieves that level. For example, if all Level 1 criteria for **Policy & Compliance** are met, the organization is at least Level 1 in that practice. If it also meets some (but not all) Level 2 criteria, we denote that as **"Level 1+"**. The "+" indicates partial progress toward the next level. This is important because organizations don't always neatly fit into exact levels – they might be between levels, doing some advanced activities without having fully completed all prior maturity criteria. The plus designation captures that extra assurance in place beyond the base level obtained.

Scores for each practice can thus be **0, 1, 2, 3** or a **"+" variant** (e.g. 1+, 2+, 3+). A score of 0 means no appreciable activity in that area yet. A 3 (or 3+) is the highest, indicating the organization performs all defined activities (and possibly even beyond what AIMA defines) for that practice. Once each practice is scored, the organization can visualize its overall Governance maturity (often using a radar chart or scorecard) and identify which areas to target for improvement.

It's also wise to consider the **scope** of the assessment – e.g. whether you are assessing the entire organization's AI program or just one business unit or project. If the scope is narrower, some activities might be handled outside your scope (for example, a centralized AI governance function at corporate level), and the assessment should note those cases rather than simply marking "No" (similar to SAMM's guidance on not prematurely labeling things *Not Applicable*). In any case, conducting an AIMA assessment – whether lightweight or detailed – enables a structured evaluation of AI maturity and a fact-based discussion on how to advance responsible AI governance in alignment with business goals.

4.1 Responsible AI Assessment Worksheet

Below is a sample **Responsible AI domain assessment worksheet**. It is organized by practice (Ethical Value& Societal Impact, Fairness & Bias, Transparency & Explainability) and grouped by maturity level. For each question, an assessor would mark **Yes or No** to indicate whether the organization currently fulfills that criterion. Achieving all "Yes" answers in Level 1 for a given practice indicates maturity Level 1 is attained; all Level 1 and Level 2 "Yes" indicates Level 2, and so on. (Partial "Yes" in the next level would be noted as a "+" as discussed.)

Ethical Values and Societal Impact – Assessing and managing broader impacts of AI systems.

Maturity Level	Stream A: Impact Assessment Procedures	Stream B: Ethical Decision-Making Framework
Level 1	Is there informal awareness of the potential ethical and societal impacts of AI systems?	Are ethical considerations occasionally discussed in an informal manner?
Level 2	Have formal processes been established to assess AI's ethical and societal impacts?	Is there an established framework guiding ethical decision-making for AI systems?
Level 3	Are impact assessments systematically integrated into all AI projects, continuously reviewed, and updated?	Is ethical decision-making fully embedded in organizational processes, consistently guiding AI development and deployment?

Transparency & Explainability – Providing understandable explanations of AI systems.

Maturity Level	Stream A: Explainability Mechanisms	Stream B: Transparency Reporting and Communication
Level 1	Are there informal efforts to explain AI outputs or decisions when requested?	Is communication about AI systems' workings sporadic or reactive?

Maturity Level	Stream A: Explainability Mechanisms	Stream B: Transparency Reporting and Communication
Level 2	Are formal explainability mechanisms in place for critical AI models or systems?	Are transparency and explanations regularly documented and shared internally?
Level 3	Are advanced, comprehensive explainability techniques consistently applied across all AI systems?	Is there proactive external reporting and open communication regarding AI transparency?

Fairness & Bias – Ensuring AI systems operate without unfair discrimination.

Maturity Level	Stream A: Bias Identification and Assessment	Stream B: Bias Mitigation Strategies
Level 1	Is there initial awareness and informal identification of potential biases in AI systems?	Are any informal or ad hoc bias mitigation steps currently in place?
Level 2	Are systematic procedures established to regularly identify and assess biases in AI models?	Are defined mitigation strategies implemented and periodically reviewed?
Level 3	Is bias assessment integrated systematically across all AI lifecycle stages and audited regularly?	Are proactive mitigation practices continuously monitored and refined across AI deployments?

4.2 Governance Assessment Worksheet

Below is a sample **Governance domain assessment worksheet**, modeled after the OWASP SAMM Governance worksheet. It is organized by practice (Strategy & Metrics, Policy & Compliance, Education & Awareness) and grouped by maturity level. For each question, an assessor would mark **Yes or No** to indicate whether the organization currently fulfills that criterion. Achieving all "Yes" answers in Level 1 for a given practice

indicates maturity Level 1 is attained; all Level 1 and Level 2 “Yes” indicates Level 2, and so on. (Partial “Yes” in the next level would be noted as a “+” as discussed.)

Strategy & Metrics – Aligning AI initiatives with business strategy and measuring AI program effectiveness.

Maturity Level	Stream A: Strategy Alignment	Stream B: Metric Management
Level 1	Is there an initial AI strategy documented, even informally?	Are there any metrics informally tracked related to AI initiatives?
Level 2	Has the AI strategy been formally defined and communicated to stakeholders?	Are defined metrics regularly reviewed and communicated within the organization?
Level 3	Is the AI strategy integrated into the organization's broader business strategy and iteratively refined?	Are metrics systematically analyzed to drive improvements and decision-making processes?

Policy and Compliance – Establishing AI policies and meeting legal/ethical requirements.

Maturity Level	Stream A: Policy Development	Stream B: Compliance Management
Level 1	Is there an awareness or initial informal policy for AI usage within the organization?	Is there basic awareness of compliance needs relevant to AI (e.g., GDPR, ethical guidelines)?
Level 2	Has a formal AI policy been established and clearly communicated to all relevant stakeholders?	Are compliance requirements identified, documented, and regularly reviewed to ensure alignment with AI-specific regulations?
Level 3	Is the AI policy consistently enforced and reviewed regularly for relevance, accuracy, and alignment with organizational goals and external standards?	Is compliance management applied through formal mechanisms into daily operations, with proactive management of

Maturity Level	Stream A: Policy Development	Stream B: Compliance Management
		compliance risks and regular audits?

Education and Awareness – Training and guiding personnel on secure and ethical AI.

Maturity Level	Stream A: AI Security Training	Stream B: Awareness and Communication
Level 1	Is there initial informal training or general awareness about AI security risks within the organization?	Is communication about AI security risks sporadic or ad hoc?
Level 2	Are formal training programs on AI security established, targeting key stakeholders and teams?	Is there regular communication about AI security best practices and updates across the organization?
Level 3	Are AI security training programs regularly updated, mandatory, and effectively tailored for different roles and responsibilities?	Is there an established culture of proactive communication, continuous awareness, and engagement around AI security throughout the organization?

Each section above corresponds to one of the Governance practices in AIMA. An assessor would review each question with stakeholders, mark Yes/No, and then determine the maturity level achieved. For instance, if **Strategy & Metrics** has all Level 1 and Level 2 questions answered "Yes," but not all of Level 3, the score would be **Level 2+** for that practice. This worksheet format ensures a **structured yet flexible** assessment: it captures granular practices through yes/no checklists while mirroring the SAMM style that stakeholders may already be familiar with, now applied to the domain of AI.

4.3 Data Management Assessment Worksheet

Below is a sample **Data Management domain assessment worksheet**. It is organized by practice (Data Quality & Integrity, Data Governance & Accountability, Data Training) and grouped by maturity level. For each question, an assessor would mark **Yes or No** to indicate whether the organization currently fulfills that criterion. Achieving all "Yes" answers in Level 1 for a given practice indicates maturity Level 1 is attained; all Level 1 and Level 2 "Yes" indicates Level 2, and so on. (Partial "Yes" in the next level would be noted as a "+" as discussed.)

Data Quality & Integrity – Ensuring AI data is accurate, reliable, and consistent.

Maturity Level	Stream A: Quality Assurance Procedures	Stream B: Data Integrity Controls
Level 1	Are there informal or ad hoc processes to ensure basic data quality?	Are initial integrity checks occasionally performed on data?
Level 2	Are formalized data quality procedures defined and regularly executed?	Are consistent data integrity controls systematically applied and reviewed?
Level 3	Is data quality management embedded throughout the data lifecycle and continuously improved?	Are advanced integrity controls proactively monitored and refined across all datasets?

Data Governance & Accountability – Managing data responsibly and transparently.

Maturity Level	Stream A: Data Governance Structures	Stream B: Accountability and Compliance Monitoring
Level 1	Is there initial awareness or informal processes in place for data governance?	Are basic accountability measures occasionally discussed informally?

Maturity Level	Stream A: Data Governance Structures	Stream B: Accountability and Compliance Monitoring
Level 2	Are formal governance structures and responsibilities clearly defined and communicated?	Are accountability and compliance regularly reviewed through structured assessments?
Level 3	Is data governance systematically integrated into organizational operations, continuously reviewed, and optimized?	Is comprehensive accountability proactively managed, regularly audited, and documented?

Data Training – Managing and Monitoring AI training datasets.

Maturity Level	Stream A: Dataset Management (Accuracy, Consistency, Curation)	Stream B: Monitoring & Compliance (Security, Licensing, Ethical Use)
Level 1	Is training data gathered informally, with minimal consistency or curation standards?	Are there minimal or no compliance checks for third-party data usage?
Level 2	Are standardized processes for dataset collection and labeling formally defined?	Are compliance and ethical standards regularly reviewed for external datasets?
Level 3	Is data preparation fully automated, consistently maintained, and regularly evaluated and adjusted?	Is monitoring of datasets for security, licensing, and ethical use systematically implemented and regularly audited?

4.4 Privacy Assessment Worksheet

Below is a sample **Privacy domain assessment worksheet**. It is organized by practice (Data Minimization & Purpose Limitation, Privacy by Design & Default, User Control & Transparency) and grouped by maturity level. For each question, an assessor would mark **Yes or No** to indicate whether the organization currently fulfills that criterion. Achieving all "Yes" answers in Level 1 for a given practice indicates maturity Level 1 is

attained; all Level 1 and Level 2 “Yes” indicates Level 2, and so on. (Partial “Yes” in the next level would be noted as a “+” as discussed.)

Data Minimization & Purpose Limitation – Ensuring AI collects only necessary data for explicit purposes.

Maturity Level	Stream A: Data Minimization Practices	Stream B: Purpose Limitation Controls
Level 1	Is there basic awareness and informal processes around data minimization?	Are data collection purposes informally discussed or inconsistently documented?
Level 2	Are formal procedures established to regularly review and minimize data collection?	Are explicit purposes clearly defined, communicated, and regularly reviewed?
Level 3	Is data minimization proactively embedded into data collection practices across all operations?	Are stringent purpose limitation controls systematically enforced and audited?

Privacy by Design & Default – Integrating privacy considerations throughout AI system development.

Maturity Level	Stream A: Privacy by Design Procedures	Stream B: Default Privacy Settings and Controls
Level 1	Is there initial awareness or informal consideration of privacy aspects during AI design?	Are default privacy settings informally considered in AI systems?
Level 2	Are formal privacy by design procedures integrated into AI development processes?	Are default privacy controls systematically implemented and documented?
Level 3	Is privacy by design fully embedded and iteratively refined across the entire AI lifecycle?	Are comprehensive default privacy settings proactively managed and regularly audited?

User Control & Transparency – Empowering users with clear information and control over their data.

Maturity Level	Stream A: User Transparency Mechanisms	Stream B: User Control Mechanisms
Level 1	Is there basic, informal communication to users regarding data use and AI operations?	Are informal processes in place to occasionally respond to user data control requests?
Level 2	Are clear, formal transparency practices regularly provided to users regarding AI data usage?	Are structured mechanisms in place to facilitate user control over personal data?
Level 3	Is comprehensive transparency proactively maintained, with ongoing user communication and updates?	Are advanced user control mechanisms fully integrated, continuously improved, and audited for effectiveness?

4.5 Design Assessment Worksheet

Below is a sample **Design domain assessment worksheet**. It is organized by practice (Threat Assessment, Security Architecture, Security Requirements) and grouped by maturity level. For each question, an assessor would mark **Yes or No** to indicate whether the organization currently fulfills that criterion. Achieving all "Yes" answers in Level 1 for a given practice indicates maturity Level 1 is attained; all Level 1 and Level 2 "Yes" indicates Level 2, and so on. (Partial "Yes" in the next level would be noted as a "+" as discussed.)

Threat Assessment – Identifying and managing security threats specific to AI systems.

Maturity Level	Stream A: Threat Identification	Stream B: Threat Mitigation
Level 1	Is there basic awareness or informal identification of threats specific to AI systems?	Are informal threat mitigation strategies occasionally discussed or implemented?

Maturity Level	Stream A: Threat Identification	Stream B: Threat Mitigation
Level 2	Are threats systematically identified and documented for AI systems?	Are documented mitigation strategies developed and periodically reviewed?
Level 3	Is comprehensive threat assessment consistently performed and integrated across AI lifecycle?	Are proactive and comprehensive mitigation strategies continuously implemented and refined?

Security Architecture – Designing secure AI system architectures.

Maturity Level	Stream A: Secure Model Deployment	Stream B: Architectural Compliance
Level 1	Is initial security awareness or informal consideration present in AI deployment?	Are informal checks occasionally performed to ensure architectural compliance?
Level 2	Are formal procedures established for secure AI model deployment?	Are regular architectural compliance reviews systematically conducted?
Level 3	Is secure deployment consistently enforced, continuously refined, and fully integrated?	Is comprehensive architectural compliance proactively managed and regularly audited?

Security Requirements – Defining clear and actionable security criteria for AI systems.

Maturity Level	Stream A: Requirements Definition	Stream B: Requirements Verification
Level 1	Are security requirements informally identified or sporadically documented?	Are informal verification processes occasionally applied to security requirements?
Level 2	Are security requirements formally documented, clearly defined, and consistently communicated?	Are systematic verification procedures regularly conducted to ensure requirements are met?

Maturity Level	Stream A: Requirements Definition	Stream B: Requirements Verification
Level 3	Are security requirements continuously improved and fully integrated across AI projects?	Are comprehensive and proactive verification mechanisms consistently enforced and audited?

4.6 Implementation Assessment Worksheet

Below is a sample **Implementation domain assessment worksheet**. It is organized by practice (Secure Build, Secure Deployment, Defect Management) and grouped by maturity level. For each question, an assessor would mark **Yes or No** to indicate whether the organization currently fulfills that criterion. Achieving all "Yes" answers in Level 1 for a given practice indicates maturity Level 1 is attained; all Level 1 and Level 2 "Yes" indicates Level 2, and so on. (Partial "Yes" in the next level would be noted as a "+" as discussed.)

Secure Build – Ensuring AI systems are securely built.

Maturity Level	Stream A: Process-Oriented	Stream B: Technical Controls
Level 1	Are there basic informal practices for secure building of AI systems?	Is security tooling or automation occasionally used in the build process?
Level 2	Are formal, systematic build security procedures documented and consistently applied?	Is security tooling regularly integrated into the build pipeline?
Level 3	Is secure build methodology fully integrated, continuously monitored, and regularly improved?	Are advanced tooling and automation fully embedded and continuously enhanced in the build process?

Secure Deployment – Deploying AI systems securely.

Maturity Level	Stream A: Process-Oriented	Stream B: Technical Controls
Level 1	Are there informal or ad hoc processes for securely deploying AI systems?	Are basic technical controls occasionally implemented during deployment?
Level 2	Are formal processes defined and consistently followed for secure deployment of AI systems?	Are standard technical controls systematically implemented and regularly reviewed?
Level 3	Is secure deployment methodology fully integrated, continuously monitored, and regularly improved?	Are advanced technical controls proactively managed and audited during deployment?

Defect Management – Identifying, tracking, and resolving defects in AI systems.

Maturity Level	Stream A: Process-Oriented	Stream B: Technical Controls
Level 1	Are defect tracking processes informally applied or inconsistently documented?	Are basic technical methods occasionally used to identify and resolve defects?
Level 2	Are defect tracking processes systematically implemented and regularly documented?	Are technical methods consistently applied and regularly reviewed to manage defects?
Level 3	Are defect tracking processes fully integrated, proactively managed, and continuously refined?	Are advanced technical controls fully embedded and continuously enhanced in defect management?

4.7 Verification Assessment Worksheet

Below is a sample **Verification domain assessment worksheet**. It is organized by practice (Security Testing, Requirement-Based Testing, Architecture Assessment) and grouped by maturity level. For each question, an assessor would mark **Yes or No** to in-

dicate whether the organization currently fulfills that criterion. Achieving all "Yes" answers in Level 1 for a given practice indicates maturity Level 1 is attained; all Level 1 and Level 2 "Yes" indicates Level 2, and so on. (Partial "Yes" in the next level would be noted as a "+" as discussed.)

Security Testing – Validating AI systems through systematic security testing.

Maturity Level	Stream A: Conduct Security Assessments	Stream B: Measure and Improve Security
Level 1	Are basic security assessments occasionally conducted informally on AI systems?	Is there informal measurement and basic improvement of security practices?
Level 2	Is there a systematic approach documented for conducting regular security assessments on AI systems?	Are security practices measured consistently, with improvements periodically implemented?
Level 3	Are security assessments fully integrated, regularly performed, and subject to ongoing refinement?	Are security metrics comprehensively used to drive continuous improvement and regularly audited?

Requirement-Based Testing – Ensuring AI systems meet defined functional and security requirements.

Maturity Level	Stream A: Define and Execute Testing	Stream B: Measure and Improve Testing
Level 1	Are basic requirement-based tests occasionally conducted informally?	Is requirement verification informally performed with occasional improvements?
Level 2	Is there a systematic, documented approach for requirement-based testing regularly applied?	Is the effectiveness of requirements verification regularly measured and improved?
Level 3	Is requirement-based testing fully integrated, regularly executed, and continuously refined?	Is requirements verification proactively validated, improved, and consistently audited?

Architecture Assessment – Assessing AI system architectures for security and compliance.

Maturity Level	Stream A: Conduct Architecture Reviews	Stream B: Measure and Improve Architecture
Level 1	Are basic architecture reviews occasionally conducted informally on AI systems?	Is architecture improvement informally measured and occasionally addressed?
Level 2	Is there a systematic and documented approach for conducting regular architecture reviews?	Are architectural effectiveness and compliance regularly measured and improvements implemented?
Level 3	Are architecture reviews fully integrated, regularly executed, and continuously refined?	Is architectural effectiveness proactively managed, continuously measured, and regularly audited?

4.8 Operations Assessment Worksheet

Below is a sample **Operations domain assessment worksheet**. It is organized by practice (Incident Management, Event Management, Operational Management) and grouped by maturity level. For each question, an assessor would mark **Yes or No** to indicate whether the organization currently fulfills that criterion. Achieving all "Yes" answers in Level 1 for a given practice indicates maturity Level 1 is attained; all Level 1 and Level 2 "Yes" indicates Level 2, and so on. (Partial "Yes" in the next level would be noted as a "+" as discussed.)

Incident Management – Handling and resolving AI system incidents effectively.

Maturity Level	Stream A: Incident Response Preparedness	Stream B: Continuous Improvement and Reporting
Level 1	Are there basic informal procedures or ad hoc responses for managing AI incidents?	Are incidents informally documented and occasionally resolved?

Maturity Level	Stream A: Incident Response Preparedness	Stream B: Continuous Improvement and Reporting
Level 2	Is there a documented and consistently applied incident response procedure for AI systems?	Are incidents systematically managed, documented, and regularly reviewed?
Level 3	Are incident response processes fully integrated, continuously improved, and regularly exercised?	Are incident handling and resolution proactively managed, optimized, and regularly audited?

Event Management – Monitoring and managing AI system events.

Maturity Level	Stream A: Detection & Alerting	Stream B: Response & Continuous Learning
Level 1	Is there informal or occasional monitoring and detection of events in AI systems?	Are event responses informally conducted and sporadically documented?
Level 2	Are events systematically monitored and consistently detected through defined processes?	Are event responses systematically executed, documented, and regularly reviewed?
Level 3	Is event monitoring continuously refined, comprehensively managed, and fully automated?	Is event response proactively managed, iteratively refined, and regularly audited?

Operational Management – Ensuring secure and efficient operational management of AI systems.

Maturity Level	Stream A: System Monitoring & Maintenance	Stream B: Security & Compliance Management
Level 1	Are operational management procedures occasionally applied informally to AI systems?	Is operational effectiveness informally monitored and occasionally addressed?
Level 2		

Maturity Level	Stream A: System Monitoring & Maintenance	Stream B: Security & Compliance Management
	Are systematic operational procedures clearly defined, documented, and consistently applied?	Is operational effectiveness regularly assessed with improvements systematically implemented?
Level 3	Are operational processes fully integrated, consistently managed, and continuously refined?	Is operational effectiveness proactively managed, comprehensively optimized, and regularly audited?

5 Appendix

5.1 Glossary

- **Adversarial Attacks**: Malicious attempts to manipulate AI model inputs to produce incorrect or harmful outputs,
  exploiting vulnerabilities in data or model behavior.
- **Agentic AI**: Advanced AI systems capable of autonomous decision-making and continuous evolution through feedback
  loops, requiring robust governance.
- **Bias**: Unintended or unfair preferences in AI outputs, often arising from skewed training data, algorithms, or
  human oversight, potentially perpetuating discrimination.
- **Data Drift**: A gradual shift in the statistical properties of input data over time, which can reduce the accuracy, relevance, or stability of AI model predictions.
- **Data Poisoning**: The intentional insertion or manipulation of malicious data into a training dataset to degrade model performance or alter its outputs for malicious purposes.
- **Explainability**: The ability to provide clear, understandable reasons for AI decisions, enabling stakeholders to
  comprehend and trust outcomes.
- **Fairness**: The principle of ensuring that AI systems produce equitable outcomes across different user groups, mitigating systemic bias in data, design, or deployment.
- **Hallucinations**: Incorrect or fabricated outputs generated by AI models, often due to poor training data or model
  limitations.
- **Large Language Models (LLMs)**: Advanced AI models trained on vast text datasets to generate human-like language,
  requiring specific governance for risks like prompt injection.

- **Model Drift**: The degradation of an AI model's performance due to changes in data context, usage patterns, or external environments after deployment.
- **Model Poisoning**: A specific type of attack where malicious data or modifications are introduced during model
  training to alter its behavior.
- **Non-deterministic Behavior**: AI model outputs that vary with changes in data or context, making predictability and
  assurance challenging.
- **Opaque Decision Logic**: The lack of interpretability in AI models, where the reasoning behind outputs is difficult
  to understand or explain.
- **Prompt Injection**: A type of adversarial attack targeting LLMs by crafting malicious inputs to manipulate or bypass
  model behavior.
- **Responsible AI (RAI)**: A strategic approach to designing, developing, and deploying AI systems in alignment with ethical values, fairness, accountability, and legal compliance.
- **Transparency**: The practice of making AI system processes, data sources, decision logic, and risks openly available and understandable to stakeholders, supporting oversight and accountability.