

ADVANCE AI SECURITY MANAGEMENT

REFERENCE NOTES

EDITION ONE



KAMLESH SINGH

Advance AI Security Management Reference Notes

Edition One

Dec 2025

Sharing and Copyright Disclaimer

©2025 Kamlesh Singh

These notes have been created to support learners and professionals in growing together. The content has been collected, adapted, and summarized from multiple publicly available and credible sources.

You are welcome to **share and distribute** these notes **freely for learning and knowledge-sharing purposes**, provided they are **not used for sale or any commercial activity** and that **no modifications or edits** are made to the original content.

All original materials and ideas belong to their respective authors.

Table of Contents

Domain 1: AI Governance, Frameworks, and Regulations	4
1. Stakeholder consideration, Framework, and Regulations.....	4
2. AI Strategies, Policies, Procedures Implementation	10
3. AI Asset and Data Life Cycle Management	16
4. AI Security Program Development and Management	21
5. Business Continuity and Incident Response	25
Domain 2: AI Risk and Opportunity Management.....	30
1. AI Risk, Threat, Vulnerability, and Supply Chain Management.....	30
2. Threat and Vulnerability Management	34
3. AI Vendor and Supply Chain Management.....	38
Domain 3: AI Security Architecture and Life Cycle	42
1. AI Security Architecture and Design	42
2. AI Life Cycle (OECD Model): -.....	46
3. Data Management Controls	48
4. Privacy, Ethical, Trust, and Safety Controls	52
5. AI Security Controls and Monitoring	54

Domain 1: AI Governance, Frameworks, and Regulations

1. Stakeholder consideration, Framework, and Regulations

- **AI Governance:** - contains **senior oversight, policy, processes, standards, and safeguards** & ensures the **safe and ethical** use of AI systems and tools.
 - **Purpose:**
 - Direct research, development, and application to **promote safety, fairness, and human rights**.
 - Support **business objectives and drive value**
 - Protect assets, data, and business reputation.
 - Mitigate risk from the **human element** in AI (e.g., human biases/errors leading to discrimination).
 - **Risk Mitigation Approach via AI policy, regulation, and data governance.**
 - **Key Principles:**
 - **Responsible AI (RAI) Usage:** Essential for building trust.
 - **Transparent Decision-Making and Explainability:** Understanding *how* AI systems make decisions to ensure **fairness and accountability**.
 - **Current Trend:** Moving from standard **legal compliance** toward ensuring **AI's social responsibility** to safeguard against financial, legal, and reputational damage.
 - **AI Considerations for Enterprises:**
 - **Data Confidence:** Guaranteeing quality for data acquisition.
 - **Transparency/Explainability:** Understanding objectives, hypotheses, and rules in automated systems.
 - **Competitive Pressure:** Risk of not using AI.
 - **Bias/Errors:** Damage due to model decisions.
 - **Speed Disparity:** Gap between the speed of automated learning systems and human compliance controls.
 - **Workforce Impact:** Discrimination, redundancy, and loss of organizational knowledge vs. more time for human creativity.
 - **Reputation Impact:** Brand
 - **Risk Reduction:** AI complementing human performance on repetitive tasks.
- **AI Readiness:** - Assessment is imperative for AI adoption with adequate governance.
 - **Key Components for a Defined AI Governance System:**
 - **Mature I&T Governance Process:** Ensures **stakeholder needs and** balanced objectives and monitors performance/compliance.
 - **Mature Processes:** For innovation management, portfolio management, and external compliance management.
 - **Appropriate Governance Structures:** *Governance committee, board of directors, I&T strategy committee, and risk/audit committees.*
 - **Skills:** Ensuring the right resources with knowledge to execute informed AI decisions.
 - **Key Questions for Governance and Management**
 - **Performance/ROI:** How is success measured?
 - **Business Alignment:** Does AI support long-term strategic objectives?

- **Organizational Readiness:** Is the culture/operation ready? Is the right talent in place?
 - **Risk Management:** Is AI risk integrated into the enterprise risk register? Is ethical, security, and reputational risk managed?
 - **Regulatory Compliance:** Is the organization keeping up with evolving regulations?
 - **Data Privacy/Security:** Are processes in place to safeguard sensitive data?
 - **Bias Mitigation:** What steps are taken to prevent bias?
 - **Framework Adoption:** Does an AI governance framework exist and has it been adopted?
 - **Human Resources:** Are guidance, skilled staff, and training provided?
- **AI Roles and Responsibilities:** - Needed for Effective governance and require to ensure accountability, oversight, risk mitigation, and innovation.

Category	Focus	Common Examples
Leadership and Strategy	Define vision, objectives, and implementation roadmap.	Executive Managers (tone at the top, ultimate responsibility) Chief AI Officer (overall adoption), AI Steering Committee (cross-functional insight).
Developmental and Operational	Creation, implementation, and maintenance of AI solutions.	IT Development (data scientists, engineers, researchers) IT Operations (deployment, maintenance), Product Management (usability, bridging technical/user gaps).
Users	Actual users of AI solutions and supporting roles.	End Users (employees/consumers) Human Resources (job classification, training) Customer Service (provide support to customer)
Governance and Oversight	Ensure security, regulatory, and ethical adherence.	Governance Committee (overseeing program/policies), Risk Management (integrating AI into ERM), Information Security and Privacy (assets security) Internal Audit (Audit assessment)

- **Role of the Enterprise's Governing Body:** - responsible and accountable for all organizational activities, including AI consequences. This responsibility **cannot be delegated**.
 - **Core Responsibilities:**
 - Determine if the intended use of AI is in line with the **risk appetite**.
 - Inform themselves about the possibilities and **risk posed by AI systems**
 - Ensure practices are appropriate for the specific AI use cases.
 - **Key Actions:**
 - **Management:** Review/improve policies, strategies, ethical codes, resource allocation related to AI use.
 - **Supervision:** Evaluate the AI solution's value, risk appetite, monitoring, and decision-making capabilities.
 - **Evaluation:** Assess internal/external factors (threats/opportunities) and the effectiveness of governance mechanisms.
 - **Reporting:** Demonstrate effective governance to interested parties.
 - **Oversight/Final Approval:** Over change management processes.
- **Stakeholders:** - Internal and external stakeholder needs are essential for responsible AI deployment and organizational value.

- **Best Practices for Addressing Concerns:** Stakeholder mapping, prioritization, engagement/communication, ethical AI development, monitoring, and evaluation.
- **Internal Stakeholders:**
 - **Senior Leadership/Board:** Alignment with goals & strategy, risk management (implementation failure, reputational harm), compliance, and addressing job displacement.
 - **Employees:** Participation in training, ensuring AI complements workflows.
 - **IT and Technical Teams:** Infrastructure support, alignment with strategy, **data quality/quantity/accessibility**, and long-term support.
- **External Stakeholders:**
 - **Customers/End Users:** User-friendliness, **data privacy compliance (GDPR, CPRA)**, transparency on AI impact, preventing discrimination/bias.
 - **Third Parties (Vendors):** System integration, open communication, shared ethical AI practices.
 - **Regulators/Policymakers:** Adherence to regulations, dialogue to shape policies, proactive reporting.
 - **Society/Communities:** Socially responsible use, minimizing harm, inclusiveness, sustainability (energy consumption).
- **AI Charter:** - ensuring **clear governance, accountability, and strategic alignment** for an AI initiative.
 - **Key Components of AI charter**
 - Project name and description
 - Objectives and goals (specific, measurable outcomes)
 - Scope (in-scope and out-of-scope)
 - Stakeholders (internal and external)
 - **Governance structure** (steering committee, reporting, decision-making, **accountability for AI change management**)
 - Timeline and milestones
 - Resources and budget
 - **Risk management** (identifying risk, mitigation strategies)
 - Success metrics (ROI, accuracy, adoption rate)
 - Approval and authorization
- **AI Steering Committee:** - Provides **strategic oversight**, ensures alignment with organizational goals, and resolves major issues throughout the AI project life cycle.
 - **Key Responsibilities:** Maintain strategic alignment, make major decisions (budget, scope, risk), and ensure AI practices are **ethical and compliant**.
 - **AI steering committee composition (cross-functional):**
 - **Chairperson:** Senior leader (CIO, CTO, CDO).
 - **Project Sponsors:** Executives funding the project.
 - **Domain Experts:** Representatives from key business units (operations, HR, etc.).
 - **Technical Experts:** AI/data science leaders, security experts, developers.
 - **Compliance Officer/Legal/Ethical Advisors.**
- **Role of the Enterprise in AI (EU AI Act Definitions)**
 - **AI Provider:** Any person or entity that **develops** an AI solution/model and puts it on the market.
 - **AI Deployer:** Any person or entity that **uses** an AI system in a professional capacity.

- **AI Standards, Frameworks, and Regulations**

- **AI Standards and Frameworks:** - assist organizations with governance, compliance, and operational processes.

Standard/Framework	Description
ISO/IEC 42001	Voluntary Standard. Focuses on ethical use, transparency, and accountability in AI governance. Requirements for establishing, implementing, maintaining, and improving an AI management system.
IEEE 7000-2021	Standard Model Process for Addressing Ethical Concerns during System Design. Focuses on embedding human values into AI (privacy, transparency, fairness) during the life cycle.
NIST AI RMF	Assists organizations in identifying, assessing, and mitigating risk in AI systems.
ISO/IEC 23053	Framework for AI Systems Using Machine Learning (ML). Describes a generic AI system using ML technology.
OECD AI Principles	Key principles include transparency, inclusiveness, sustainability, and accountability.
COBIT	Offers a holistic approach to AI governance, aligning I&T with business objectives, managing risk, and ensuring optimal resource use.

- **COBIT Application to AI Life Cycle**

- **Design Phase (EDM - Evaluate, Direct, Monitor):** Ensures AI initiatives align with strategy, evaluates potential impact, directs resources, monitors performance (**ethical use, fairness**).
- **Development Phase (APO - Align, Plan, Organize):** Defines management structures, establishes processes for secure/trustworthy AI (data governance, security, compliance), and addresses bias/vulnerabilities.
- **Deployment Phase (BAI - Build, Acquire, Implement):** Ensures secure and efficient integration, proper testing/validation, and change management.
- **Operations Phase (DSS - Deliver, Service, Support):** Ensures efficient performance, service level maintenance, prompt issue resolution, and security/continuity.
- **Monitoring Phase (MEA - Monitor, Evaluate, Assess):** Continuous evaluation of AI systems against objectives, measuring performance, and ensuring compliance with evolving regulations.

- **Cloud Security Alliance (CSA) Four Pillars Framework:** Useful for AI projects.

- **Model Cards:** Transparent overview of an ML model (objectives, training data, capabilities, limitations, performance).
- **Data Sheets:** Detailed descriptions of the dataset used to train an ML model (creation process, composition, biases, ethical considerations).
- **Risk Cards:** Systematically identify, categorize, and analyze potential risk issues (observed risk, planned remediations, anticipated user behavior).
- **Scenario Planning:** Proposes hypothetical misuse/malfunction scenarios to identify unforeseen risk and develop mitigation strategies.

- **Laws and Regulations:** - Laws are relatively new, often in draft format, and untested from a litigation perspective. Noncompliance can result in fines/penalties.

- **EU AI Act (Key Example):**

- **Scope:** Providers and users of AI in the EU.
- **Risk Classification:** Four categories: **Prohibited, High, Limited, Minimal Risk.**
- **Penalties:** Up to €35M or 7% of global annual revenue.
- **Transparency:** Mandated for chatbots, deepfakes, and high-risk AI.
- **Human Oversight:** Mandated for high-risk AI.
- **Compliance Considerations:**
 - **Information Confidentiality:** Users can expose enterprise IP or trade secrets (violates regulations, reputational damage).
 - **Privacy:** Must be guaranteed (**Privacy by Design**), requires internal policy review, and a **Privacy Impact Assessment (PIA)** before production use.
 - **Intellectual Property (IP):** Debate on copyright for GenAI output. Enterprises must prohibit use of copyrighted material for training/content creation without legal guarantees.
 - **Employee Rights:** Review potential impact on employees/applicants; guarantee fair selection processes.
- **Gaps in Regulatory Coverage:**
 - **Lack of International Coordination:** Inconsistencies and lack of interoperability create challenges for multi-region organizations.
 - **Sectoral Gaps:** Healthcare is ahead; Finance and Education sectors lack comprehensive models for fairness, transparency, and bias.
 - **Enterprise Ownership Gap:** Unclear AI governance ownership within organizations, leading to lack of alignment with overall policies/strategy.
- **AI Use Cases**

AI systems leverage existing capabilities (networks, IoT, big data, cloud).

Machine Learning (ML) is a key technique giving computers the ability to "learn" without explicit programming.

 - **AI Characteristics Impacting Governance:**
 - Ability to **automate decisions**.
 - Use of **data analysis, knowledge, and learning** instead of logic.
 - Ability to **adapt** as the environment changes (not explicitly codified).
 - **Examples of AI Use Cases:**
 - automatic content creation (GenAI)
 - accelerated code writing, personalization
 - formatting/summarizing information
 - enhancing dataset quality
 - improved customer service (chatbots).
 - **Potential Business problem and solutions**

Problem	Type of Output	Examples
Classification	Discrete categories (tags)	Spam detection, image recognition, medical diagnosis.
Regression	Continuous value	Prediction of house price, energy consumption.
Clustering	Similar data groups	Customer segmentation, gene grouping.
Detection of Anomalies	Deviations from expected pattern	Fraud detection, predictive maintenance.

Recommendation Systems	Personalized recommendations	Product or movie suggestions.
------------------------	------------------------------	-------------------------------

- **AI Project Failure Factors:** Often attributed to people, process, and data issues.
 - Misunderstanding/miscommunicating the problem.
 - Lack of quality data.
 - Not solving the business problem.
 - Inadequate infrastructure.
 - Applying AI to problems too difficult for AI.
- **Limitations of AI Use**
 - **Unexpected Results:** AI output heavily influenced by user input (prompts must be precise).
 - **Inaccurate or Inappropriate Results (Hallucinations):** AI can produce results that do not align with real-world data; it is crucial to thoroughly verify GenAI results before further use or model training.
 - **Lack of Context:** AI lacks real-world knowledge and is limited to training data.
 - **Response Constraints:** Inherent limitations in architecture to prevent undesirable or unlawful content, which may also reduce usefulness for acceptable applications.
 - **Bias:** AI can deliver biased responses due to biased or censored training data.
 - **Token Limits:** Limits AI's ability to retain information from lengthy conversations.
- **Business Cases, Needs, Scope, and Objectives:** - crucial to justify the investment and ensure intentional implementation.
 - **Minimum Outline for a Business Case:**
 - The problem the AI solution solves.
 - How the proposed solution will solve the problem.
 - Expectations for **ROI** based on a cost-benefit analysis.
 - Identified risk and planned mitigations.
 - A detailed implementation plan.
- **Cost-Benefit Analysis:** - benefits related to the business use case.
 - **Unique Cost Considerations:**
 - **Implementation Costs:** Development, licensing, employee training.
 - **Operational Costs:** System maintenance, energy consumption, new AI-related employee salaries.
 - **Risk and Compliance Costs:** Privacy protection, reputation management.
 - **Common Benefits:**
 - **Increased Productivity:** Automation, allowing humans to focus on higher-value activities.
 - **Cost Reduction:** Lowering labor- and error-related business process costs.
- **Return on Investment (ROI)**
 - **Metrics:** Increased revenue or decreased costs are easiest to quantify. **Human factors** (employee satisfaction, customer experience improvements) should also be considered.
 - **Expectations:** Initial ROI may be low due to high upfront costs. It is important to set **reasonable expectations early** in the life cycle.

2. AI Strategies, Policies, Procedures Implementation

- **Implementation:** Implementing AI solutions is similar to implementing other software but has unique **operational and risk** differences. Organizations must leverage an established **change management process**.
- **AI Strategy:** An organization's approach to **adopting, developing, and governing** AI solutions.
 - Enables the realization of benefits while systematically managing risk.
- **AI Strategies**
 - AI strategies provide **guidance to applicable stakeholders**.
 - **Governments** focus on ethical considerations and societal values.
 - **Companies** focus on competitive advantages.
 - **Examples of AI Strategies:**
 - **National:** Aligning ethics/country values, guiding laws/regulations, promoting national security.
 - **Industry:** Promoting innovation, guiding standards/best practices, interpreting regulations.
 - **Corporate:** Identifying competitive advantages, ensuring alignment with applicable requirements.
 - **Guiding Principles for AI (Value-Based):** Generally, focus on common ethical concepts: **explainability, fairness, transparency, and human-centricity**.
 - **OECD AI Principles:** A significant example representing guidance for trustworthy AI.
 - **Five Values-Based Principles for Trustworthy, Human-Centric AI:**
 - ◆ Inclusive growth, sustainable development, and well-being.
 - ◆ Respect for the rule of law, human rights, and democratic values (including fairness and privacy).
 - ◆ **Transparency and explainability,**
 - ◆ **Robustness, security, and safety.**
 - ◆ **Accountability.**
 - **Five Recommendations to Governments for AI Ecosystems to Benefit Societies:** Investing in R&D, fostering an inclusive ecosystem, shaping an enabling governance/policy environment, building human capacity, and international cooperation.
 - **Strategy Creation:** Must be accepted by **senior management/officials** to be successful.
 - Should be created by a **cross-functional team** (e.g., IT, risk management, legal, operations, finance).
 - Expert AI knowledge is not required within the core team; outside experts or consultants can be leveraged.
 - Existing committees (IT steering, innovation) can be candidates to define the strategy.
- **Opportunities**
 - The AI strategy proactively identifies opportunities for leveraging AI.
 - **Benefits:** Drive **efficiency, innovation, competitive advantage, and long-term value creation**.
 - **Increased Innovation:** Develop unique products, services, or business models.
 - **Enriched Data Analysis:** Analyze vast data in real-time for accurate, actionable insights.

- **Challenge:** Achieving benefits often requires **redesigning key business processes** and potentially new business models, going beyond mere technology adoption.
- **Vision and Mission**
 - A successful AI strategy requires understanding the organization's **mission** and having a clear **vision**.
 - **AI Vision:** Provides internal and external stakeholders with an understanding of the organization's intentions for AI adoption, development, and use to further its mission.
 - *Example: European Union's vision focuses on fostering excellence and uptake, ensuring AI works for people, and building strategic leadership.*
- **Value Alignment**
 - **Definition:** Designing AI systems that behave in ways consistent with **human values and ethical principles**.
 - **Variance:** Value alignment differs by organization and country; developers must consider the intended user base and their applicable values.
 - **Common Elements in AI Value Alignment:**
 - **Community:** Focus on practical solutions, adapt over time with feedback, strengthen interdisciplinary collaboration.
 - **Ethical Foundations:** Consider influences from different cultures, base ethics on comprehensive philosophical studies, respect universal human values.
 - **Legal Compliance:** Comply with specific legislation (e.g., GDPR, lèse-majesté laws).
 - **Operational Strategy:** Engage stakeholders, continuously monitor/refine, incorporate sustainability and fairness as KPIs, ensure no misinformation spread.
 - **Evaluation:** Must be evaluated through the **AI life cycle** and driven by the AI strategy, enabled by a strong **AI governance process**.
 - **AI Value Alignment Enablers (World Economic Forum):**
 - Frameworks and guidance
 - Human engagement
 - Organizational change
 - Audits and assessments
- **Build Vs. Buy**
 - **Adoption Options:** Purchase **commercial off-the-shelf (COTS)** software or develop software **in-house**.
 - **Unplanned Adoption:** Can occur with little oversight, especially with AI solutions included in existing products. Governing bodies must demonstrate **stewardship** and robust **governance/ethical practices**.
 - **Hosting Strategy (Internal vs. Cloud):**
 - **Internal (On-Premises):**
 - ♦ **Benefits:** More **control, customization**, better performance, greater ease in controlling data security/privacy.
 - ♦ **Limitations:** Additional computing resources/power/personnel required, **inability to scale**, higher resource consumption.
 - **Cloud:**
 - ♦ **Benefits:** Better **scalability**, less overhead for maintenance, less upfront costs.
 - ♦ **Limitations:** Less control, potential **vendor lock-in**, performance concerns (latency, shared resources), security/privacy risks for data leaving controlled environment.

- **Key Consideration for Cloud:** Physical location of the provider can be a concern if data is covered by privacy laws (e.g., GDPR).
- **Six Questions for Build vs. Buy Decision:**
 - How do **total costs** (implementation and operations) compare?
 - Does the enterprise have the **data, capability, and time** to build a similar or better solution?
 - What is the **risk** of buying versus building concerning regulations?
 - Which is a better fit for the enterprise's current **AI operating model**?
 - What are the **data privacy challenges**?
 - What is the risk of **vendor lock-in**?
- **Vendors**
 - Vendors must be properly **vett**ed during initial onboarding and throughout the life cycle.
 - **Initial Onboarding Assessment (Specific to AI):**
 - Alignment of **AI strategies and values**.
 - Adherence to ethical considerations (**transparency and explainability**).
 - **Data security and privacy** protection.
 - **Ongoing Oversight:**
 - Establish clear **Service Level Agreements (SLAs)**.
 - Review SLAs periodically based on risk.
 - Oversight personnel must understand the organization's **AI vision and strategies**.
 - Input from oversight functions should be leveraged for **ROI evaluations**.
- **Shared Responsibility (for Off-the-Shelf Solutions)**
 - **Risk Considerations for COTS:**
 - **Accountability and Ownership:** Unclear definition of responsibilities for maintenance and controls.
 - **Ethical Concerns:** Confusion over who is responsible when AI causes harm (incorrect results, bias).
 - **Data Security:** Sharing data with third parties increases oversight responsibilities and attack vectors.
 - **Responsibility:** Clearly defining responsibilities between the organization and the COTS provider **lies solely with the purchasing organization**.
 - **Definitions:** Clear definitions must identify components for which the organization is **fully responsible**, the vendor is **fully responsible**, and those that are **shared**.
- **AI Acceptable Use Policy (AUP)**
 - **Purpose:** Dictates what the enterprise decides is **permitted use** and **not permitted** (e.g., types of data allowed for training). A tool for **stewardship**.
 - **Common Steps Before Creating an AI AUP:**
 - **Understand AI Technologies:** Familiarize with technologies, applications, risks, and benefits (e.g., GenAI models like ChatGPT).
 - **Assess Organizational Needs:** Understand intended AI use to inform policy clauses; research legal and regulatory requirements.
 - **Conduct a Risk Assessment:** Identify potential technical (unintended outputs) and ethical risks.
 - **Confirm Purpose and Scope:** Define the policy's objective, scope (e.g., GenAI only?), and intent (ethical use, legal standards).

- **Examine Existing Policies:** Check for overlapping IT/information security policies to ensure consistency.
 - **Engage Stakeholders:** Crucial for comprehensive policy, legal compliance, operational feedback, and enforcement.
 - **Adopt a Governance Framework:** Creates a structured approach to decision-making, ownership, and assurance.
 - **Prepare for Communication:** Policies are living documents; communicate updates to employees, customers, and regulators.
- **AI Policy Development**
 - **Policy Role:** Establishes guidelines, facilitates decision-making, guarantees homogeneity of criteria, aligns employees, and communicates strategy/values.
 - **Need for AI Policy:** Necessary to implement corporate strategy and ethical values for AI use, establishing principles, criteria, and rules for employee behavior.
 - **Considerations for AI Policy Development:**
 - **Senior Management Support:** Essential for commitment to definition, communication, and implementation.
 - **Strategic Alignment:** Must align with enterprise objectives, regulatory concerns, and corporate culture/ideology.
 - **Clarity and Consistency:** Must follow best practices for policy development and be clear & concise.
 - **Feedback and Consensus:** Circulate drafts to stakeholders to ensure concerns are addressed for successful adoption.
 - **Communication and Training:** Communicate the approved policy to all stakeholders; establish a communication plan and associated training.
 - **Compliance:** Monitor compliance regularly.
 - **Periodic Review:** At least annually, or following significant changes/new regulations, to address evolving risk.
 - **Components of an AI Policy**
 - **Introduction:** Explains the reason for the policy, communicates senior management support, and strategic alignment.
 - **Glossary:** Clearly explains and defines AI terms and acronyms to reduce confusion.
 - **Scope of the Application:** Defines the framework of action and limits; specifies if it applies to all employees, vendors, suppliers, etc.
 - **Principles:** The main section establishes criteria, basic principles, and guidelines. Should include:
 - **Authorized uses (prohibited, allowed with/without authorization).**
 - **Responsible Use of AI (RAI).**
 - **Transparency of decision making.**
 - **Fairness and bias considerations.**
 - **Third-party considerations.**
 - **Governance and supervision.**
 - **Training and awareness.**
 - **Security considerations.**
 - **Compliance monitoring and Policy review/maintenance.**
 - **Responsible Use of AI Systems (RAI)**

- **Concern:** Using AI responsibly, ethically, and aligned with organizational standards, avoiding harm (reputational damage, litigation, fines, etc.).
- **Success Factors:** Organizational **culture** and appropriate **awareness and training** ensure employees understand the AI policy and AUP.
- **Elements of a Successful RAI Program:**
 - Set Tone, Empower Leadership, Establish Culture, establish a Baseline, Establish Tools and Process, Keep Humans in the Loop, Assess and Monitor.
- **RAI Maturity Matrix:**
 - **Baseline:** Little/no understanding, no documented policies, focus on data/privacy.
 - **Emerging:** Understanding of **RAI**, **foundational policies in place**, risk assessments for systems in scope.
 - **Developing:** Good understanding, **fairness, privacy, security, accuracy, oversight, transparency, and accountability** are considered; ongoing risk assessments are implemented.
 - **Realizing:** Thorough understanding with employee education, addressing all key principles, soliciting diverse stakeholder feedback, and implementing some best practices.
 - **Leading:** Thorough understanding with targeted education; addressing all key principles; implementing best practices, ongoing risk assessments, and diverse stakeholder feedback; **goal is continuous improvement**.
- **AI Procedures and Manuals**
 - **Purpose:** Standard operating procedures (SOPs), manuals, and playbooks are crucial for a robust program.
 - **Unique Qualities of AI:** Must not assume AI is just another IT system.
 - AI uses **vast amounts of data**; procedures for collecting, handling, protecting, processing, and sanitizing data for training/validation are essential.
 - **Alignment:** Procedures must align with the **AI AUP** and **RAI concepts**.
 - **Scope:** Must extend beyond the IT space to include **all enterprise uses of AI**.
 - **Adaptation:** IT security teams can adapt existing practices to guide the development of AI-specific tools. Policies must reflect the organization's **risk appetite** and **regulatory landscape**.
- **Using AI to Create Procedures and Manuals**
 - Organizations can use tools like **GenAI** and **NLP** to standardize and automate SOP creation.
 - **Pework for Using AI:**
 - **Determine processes needed:** Define the scope to tailor prompts and validate output.
 - **Use NLP to analyze existing documents:** extract key information, ensure consistent formatting/language, and pull relevant keywords/best practices.
 - **Oversight:** GenAI output requires **human oversight** and is only a starting point. The draft must be **reviewed and verified** using standard practices.
 - **Improvement:** AI can support continual monitoring and improvement by analyzing feedback and identifying bottlenecks (e.g., via **NLP/data analytics**).
- **Ethical Considerations**
 - AI solutions are subject to ethical considerations related to design and potential impact.

- **Ethically Designed AI Solutions Must:**
 - **Minimize bias and ensure fairness.**
 - **Ensure transparency and be explainable.**
 - **Promote trust and ensure safety** (to humans and data).
- Ethically designed AI guards against unintended adverse effects (e.g., loss of privacy, IP concerns, environmental impacts).
- **Ethical Use**
 - **Principle:** Developing and applying AI in ways that align with core societal values, promote social good, and **minimize harm**.
 - **Review:** Ethical considerations should be reviewed **early** in development, integrated into the overall **change management process**, and part of a **risk assessment**.
 - **Ethical Impact Assessment (EIA) (UNESCO):** Evaluates the overall design, development, and deployment of AI systems, assessing risk *before and after* release.
- **Bias and Fairness**
 - **Goal:** AI solutions must be **free of bias** to ensure fairness and avoid unintended harmful effects on specific groups. AI can perpetuate or amplify biases present in training data.
 - **NIST SP 1270 Categories of AI Bias:**
 - **Systemic:** Results from procedures/practices of institutions that favor certain social groups (e.g., Amazon recruiting tool favoring men).
 - **Statistical/Computational:** Errors when a sample is **not representative** of the population (e.g., facial recognition high error rate for Black women due to dataset diversity issues).
 - **Human:** Errors that lead to unfair outcomes due to human choices in design/implementation.
 - **Fairness:** Ensuring AI solutions promote **equitable and accurate results** that do not harm vulnerable or marginalized groups.
 - **Mitigation Tools:**
 - **AI Fairness 360:** Open-source toolkit to examine, report, and mitigate discrimination/bias in ML models.
 - **What-If Tool:** Open-source application by Google to test performance in hypothetical situations and visualize behavior.
- **Transparency and Explainability**
 - **Need:** AI solutions must be **transparent and easily explainable** so users can understand how decisions are made and verify validity.
 - **Risk:** If AI is a "black box," untrustworthy outputs may lead to poor decisions or exploitation of vulnerabilities.
 - **Explainability:** Answering, "**How did the AI solution arrive at its result?**" (e.g., citing sources in AI outputs).
 - **Benefit:** Improves **trust and safety** of AI-related solutions.
- **Trust and Safety**
 - **Need:** AI must be **trustworthy** to perform as designed and not produce harmful outcomes.
 - **Goal:** Create sound AI solutions that operate **reliably, ethically, and securely**, minimizing risk to society.

- **Consequences of Failure:** Loss of trust, abandonment of system (high cost), potential litigation, and liability (especially for AI connected to the physical world, like autonomous cars).
- **Mitigation:** Risk must be adequately assessed, and appropriate controls established to reduce harm to humans or data.
- **Considerations**
 - **Major Concern:** Protection of **Intellectual Property (IP)**, protected, or copyrighted material from unauthorized use.
 - *Example:* Litigation over GenAI platforms using original works without licensing (e.g., *Andersen v. Stability AI*).
 - **Training Data:** Developers must **source data carefully and lawfully** (e.g., entering licensing agreements with data owners).
 - **Content Creation:** Uncertainty over who is the true "**owner**" of the **copyright** of generated material (user, developer, or data owners).
 - **Policy Requirement:** Organizations must ensure policies and procedures address IP-related concerns to avoid litigation.
- **Human Rights**
 - **Faults:** Bias and discrimination in AI can harm humans and are often related to **privacy, data protection, and equality** concerns.
 - **Consequences of Infringement:** Reputational damage, litigation, and financial risk.
 - **Fundamental Rights Impact Assessment (FRIA) (EU AI Act):** Required for high-risk AI solutions.
 - **Goal:** Assess the impact a high-risk AI could have on humans or society, considering the **EU Charter of Fundamental Rights**.
 - **Scope:** Broader than a Data Protection Impact Assessment (DPIA).
- **Environmental Impact**
 - **Issue:** AI solutions are **very resource-intensive**, leading to a need for more data centers.
 - High consumption of **electricity**
 - Increased need for **critical minerals/rare elements** (unsustainable mining).
 - High **water consumption** for cooling.
 - **Action Needed:** AI developers should **proactively consider environmental effects** and strive for more **efficient algorithms**.
 - **Long-Term:** Need for **standardized procedures to measure environmental impacts** to drive laws, regulations, and incentives.

3. AI Asset and Data Life Cycle Management

An effective AI security program must begin with comprehensive asset and data lifecycle management.

- **AI Asset Identification and Inventory**
 - **AI Solutions (AISs) vs. Traditional IT Assets:** - AIS is distinct from traditional IT assets (like an application) because it is a complex system composed of:
 - **Multiple components:** owners, models, and versions.
 - **Multiple requirements, including** workflow mappings.
 - **Multiple datasets:** E.g., training datasets and production data sources.

- **Multiple algorithms** and third-party tools (e.g., Open-Source AI, Kubeflow, SageMaker).
- **Multiple considerations:** SDLC, licensures, and legal/regulatory compliance.
- **AI Asset Identification Process:** Identifying all AIS assets (internal and external) is **critical** for decision-making, operations, and risk management.
 - **Structured Approach:** Requires combining governance, risk management, IT operations, and internal audit efforts.
 - **Leadership:** Should be undertaken by the department leading **AI and/or data management**, not Internal Audit.
 - **Starting Point:** An **AI usage policy** often facilitates initial asset identification.
 - **Maintenance:** The inventory/catalog should be updated at least **annually**.
 - **Model Catalog:** Enterprises that **develop their own AIS** products/services must maintain an AI model catalog.
- **AI Inventory Objectives and Procedures**
 - **Primary Purpose:** To **identify** current AIS assets, **not to penalize** non-compliant use.
 - **Communication:** Organization-wide communication must emphasize that disclosure of AI use **supports risk management and governance** and will not result in reprimand.
 - **IS Auditor Review:** The auditor should review change, deployment, release, and program/project management documents to validate the completeness and accuracy of the inventory process.
- **AI Inventory and Data Gathering Methods:** - A baseline inventory can be assembled using several methods:

Method	Description	Role/Value
Internal Validation	Check existing change, deployment, and project management documents.	Straightforward for organization-developed AISs.
Discovery Tools	Employed by IT/Security for systems and network discovery.	Tools include Dataflow Diagrams (DFDs), metadata analysis, and Active Directory checks.
Collaboration, Tools, Surveys, Interviews	Essential components for a comprehensive effort.	Used to establish the initial baseline.
Surveys	Efficient for collecting baseline data.	Anonymity encourages truthful responses; quantifiable questions ensure effectiveness.
Interviews	Use standardized questions with predetermined responses.	Minimally useful for baseline; more helpful for teams experimenting with AI tools for educational purposes.

Note: - The organization must establish a **standardized list of data fields** for quality and consistency, including: **AIS Name, Model Version, license, Cost, Deployment Method, Purpose, Frequency of use, Stakeholders, Accountable owner, and Third-party details.**

- **Documentation and Maintenance**
 - **Artifacts:** Multiple artifacts (e.g., AI policy, goals) must be available for the IS auditor to understand the organization's mission and AIS processes.

- **Integration:** Artifacts should be integrated with standard asset management platforms (e.g., ServiceNow, Jira).
 - **Shadow AI:** Maintenance differs from traditional IT asset management due to **Shadow AI** (decentralized, external deployment of AISs without IT/ security involvement), leading to decentralized ownership.
- **Data Inventory and Management:** - Data is **fundamental** to AIS development ("teaching" models) and use (input for response/action).
 - Many AI risks, like **hallucinations** in GenAI, are linked to **biased, incomplete, incorrectly labeled, or irrelevant training data**.

- **Data Inventory:** - list of data assets, built by IT, security, or data governance.

Concept	Description
Data Dictionary	Application-level inventory; contains info about each data element in an application/system. Must link to a governance-approved business term .
Business Glossary	A unique list of defined concepts that connects common data across applications under a shared meaning. Ideal place to formally link to data assets and manage policy/classification changes.
Data Catalog	The detailed inventory of all enterprise data assets. Correlates data inventories with business contexts (like the Glossary and Dictionary). Contains metadata about data, not the data itself.

- **Metadata:** Information that identifies, describes, explains, and provides content, context, structure, and classifications pertaining to data assets. Artifacts include data dictionaries, models, business glossaries, and diagrams.
 - **Change Management:** Protocols must be defined to align applications and procedures with policy changes, ensuring applications using the same data are coordinated.
- **Creating a Data Inventory (Four Steps)**
 - If a metadata repository is unavailable,
 - A data inventory must be created for a **Privacy Impact Assessment (PIA)**.
 - **Plan:** Define scope, stakeholders, and resources.
 - **Decide:** Determine what information to collect (metadata attributes, e.g., Description, Purpose, Owner, Location, Classification, Rights and Restrictions).
 - **Populate:** Collect information using techniques like:
 - **Asking data owners to populate.**
 - **Interviews, surveys, and questionnaires.**
 - **Automated processes (e.g., metadata form upon asset creation).**
 - **Checking existing data dictionaries/system documentation.**
 - **Publish:** Collate and publish the metadata (e.g., spreadsheet or database), considering data criticality and sensitivity for access control.
 - **Dataflow Diagrams (DFDs):** - DFDs and Usage/Activity diagrams are critical for understanding and managing data flow, quality, and privacy.

Diagram Type	Orientation	Purpose/Benefit
--------------	-------------	-----------------

Dataflow Diagram (DFD)	Horizontal (Source to Target)	Better for illustrating data flow at a high level of abstraction ; modeling data movement between systems across the enterprise.
Usage Diagram (Activity Diagram)	Vertical (Through Procedures)	Better for illustrating the details of a complex workflow ; implementing, monitoring, and controlling data according to quality and privacy policies. (Emphasis by GDPR).

- **Data Lineage:** - Modeling data flow **beyond its system of record**.
 - **Includes:** Sources, transformations, storage, usage, outputs, and governance.
 - **Crucial for:** Identifying the **ultimate source** of data defects and monitoring privacy compliance.
 - **Management:** Difficult to maintain manually; requires **purpose-built tools** for automation.
 - **Best Practice:** Restrict scope to most important applications; work **backward (downstream to upstream)** in the context of a **business value stream**.
- **Data Classification:** - Delineating data into categories based on shared characteristics (can use multiple criteria simultaneously). It enriches the inventory and helps determine **sensitivity and criticality**.

Classification Type	Description & Examples
Critical Data Elements (CDEs)	Data highly effective to the enterprise. Identified via Business Impact Analyses (BIAs) . CDEs should be mapped to an approved Business Glossary entry.
Data Security Level	Tags based on intended domain of use. Categories often include: Public, Internal, Confidential (e.g., customer lists, PII/PHI governed by HIPAA/PCI DSS), and Restricted (Highly sensitive, e.g., proprietary research).
Data Sensitivity Type	Specialized handling required. Examples in the US: PII (Personally Identifiable Information) and PHI (Protected Health Information). GDPR also considers trade-union membership and religious affiliation as sensitive.

- **Standardization:** Classifications must adhere to a **common internal standard**, developed collaboratively with stakeholders (business, privacy, governance, security, compliance).
- **Control:** The classification should drive the specific **security controls** (e.g., encryption, access limits).
- **Data Collection for AI**
 - AI models require **large volumes of verifiable data**;
 - performance positively correlates with **data size, volume, and variety**.
 - Data is aggregated in **data warehouses, data lakes, or other solutions**.
 - **The Five Vs of Big Data**

Velocity	Speed of data creation and transmission.
Volume	Amount of available data.
Value	What an enterprise can do with the data.
Variety	Diversity of the data.
Veracity	Quality, accuracy, integrity, and credibility of data.

- **Data Types**
 - **Unstructured Data:** No predefined structure/format (e.g., MS Office files, text files, audio/video/image files, social media posts, metadata).
 - **Structured Data:** Organized in a predefined format, typically in a relational database with a schema and potentially a data dictionary.
- **Data Collection Risk**
 - **Consent:**
 - Using **customer data** for AI training often requires additional consent. Regulations like GDPR require consent for personal data processing.
 - The **EU AI Act** extends this to the testing of high-risk AI systems in real-world conditions.
 - Processes must be in place to **exclude** data if consent is denied or revoked.
 - **Fit for Purpose:**
 - The AI solution must be designed to accomplish a specific goal.
 - Signs it's **not fit for purpose** include data unavailability
 - Wrong granularity/volume/veracity, or restricted use case.
 - **Data Lag/Model Drift:** AI models trained on historical data may become less accurate over time due to new, un-trained data.
 - **Mitigation:**
 - ◆ **Retraining** with updated datasets or applying
 - ◆ **Retrieval Augmented Generation (RAG)** to ground LLM prompts with specific, up-to-date data.
- **Data Confidentiality Considerations:** - Confidentiality must be maintained throughout the AI development life cycle.

Data Location	Confidentiality Control Concerns
Data Source	Data may be restricted, obfuscated, or encrypted. Metadata/controls may not be extracted/carried forward .
Data Lake	Access controls are not always preserved .
Data Exploration/Training Platform	Data of different classifications may be stored in notebooks (e.g., Jupyter, SageMaker).
Vector Database	Stores data as mathematical representations (vectors) with metadata. New access control models (e.g., attribute/index or query-based) must be considered.
AI System Production	Live production data runs through prep pipelines for inference. Classification and handling protocols must be implemented in the production system.

- **Data Quality:** - Data quality is **paramount** to AI performance; the results are only as good as the data they were trained on.
- **Dimensions of Data Quality:** - Data profiling and cleansing are common practices.

Dimension	Definition
Accuracy	Data is free from errors and representative of real-world situations.
Completeness	Data contains all necessary fields and records.

Consistency	Data is uniform and standard across datasets (e.g., formats, lengths).
Timeliness	Data is up to date and available when needed (e.g., real-time GPS).
Validity	Data adheres to defined business and technical logic.
Uniqueness	Data has no duplicative or redundant records.

- **Data Balancing:** Data imbalance occurs when the training dataset has insufficient samples of a minority class, leading to **biased or inaccurate results**.
 - **Mitigation:** Address data balancing **early** (collection and preprocessing).
 - **Techniques:** **oversampling**, **undersampling**, or applying **cost-sensitive algorithms**.
- **Data Scarcity:** - The lack of sufficient **high-quality, relevant, fit-for-purpose, and consented/licensed data** for developing prioritized AI models.

Root Cause	Mitigation Strategies
Data quality issues, Unavailability of diverse/minority data, Lack of consented data, Data trapped in source systems, Unavailability of labeled data.	Augmentation: Collecting/procuring missing data, generating synthetic data , or imputing missing data.
	Model Selection: Choosing an AI model that works well with the available dataset to avoid overfitting.

- **Model Cards:** - Structured documents providing technical details, intended use, and limitations of an AI model to promote **transparency and governance**.
 - **Key Elements**
 - **Model Details:** Name, description, version, parameters, fairness constraints, license, responsible personnel.
 - **Use Case:** Intended uses, use cases, and users.
 - **Architecture Information:** Design, hardware, and underlying technologies.
 - **Training Information:** Summary of the dataset (what, how, where, and when obtained, statistical distribution, and may be limited for proprietary info).
 - **Performance Metrics:** How the model performs against a testing dataset.
 - **Limitations and Biases:** Known variables, limitations, or potential biases.
 - **Ethical Considerations:** Details on privacy, fairness, and human/societal impacts.

4. AI Security Program Development and Management

Developing an AI security program requires a **proactive approach** that factors in the unique characteristics and challenges of AI.

Continuous risk management, monitoring, security policies, and controls are essential to mitigate **data leakage** and **AI misuse**.

AI can also be used to enhance security by automating process reviews and non-critical decisions or by engaging managed service partners with specialized AI knowledge.

• AI Security Program Development

The following protocols and practices are key to establishing an effective AI security program:

- **Trust but Verify**

- **Issue:** AI models, especially public, third-party Large Language Models (LLMs), evolve continuously and pull from potentially **unverified data sources**.
 - **Vulnerability:** AI platforms are susceptible to hacking, infiltration, and "jailbreaks" (working around rules) to produce unwanted content or inject malware.
 - **Mitigation:** All AI output **must be perpetually validated**. Teams need to develop mechanisms to assess and approve AI-generated work.
- **Design Acceptable Use Policies (AUPs)**
 - Policies must be established for **approved and unauthorized uses** of AI.
 - AUPs must be verified, and employees must receive appropriate **training and education**.
- **Designate an AI Lead**
 - While a C-level role may eventually be needed, organizations should appoint an analyst or project manager to:
 - Track **AI evolution**.
 - Create a dedicated plan documenting the company's relationship with AI tools.
 - Collaborate with a **cross-functional and diverse group** of stakeholders (cybersecurity, data privacy, legal, procurement, risk, audit, etc.).
 - Document the enterprise's **history of AI use** for future decisions and lessons learned.
- **Perform a Cost-Benefit Analysis**
 - A thorough cost-benefit analysis (CBA) is critical for AI implementation (e.g., evaluating **build vs. buy**).
 - CBA must include the **cost of security controls**, alongside possible productivity gains and workforce optimization.
- **Adapt and Create Cybersecurity Programs**
 - **Security by Design** is a key principle when adapting or creating new programs.
 - **Key Considerations:**
 - **Access Control:** Implementing **permission-based access visibility**, firewalls, and application controls to prevent data leakage and control *where* and *how* employees use AI.
 - **Disaster Recovery and Incident Response.**
 - **Threat Intelligence:** Consulting sources like **OWASP** for the latest AI-related threats.
- **Mandate Audits and Traceability**
 - Auditing capabilities are essential to understand an AI tool's **data sourcing** and **decision-making processes** (traceability).
 - **Key Questions:** Source data origin, manipulation history (AI or human), and presence of **systemic bias**.
- **Develop a Set of AI Ethics**
 - The **ethical impact** of AI is a unique concern for IT security professionals.
 - Enterprises can refer to recommendations from bodies like **UNESCO** (Recommendations on the Ethics of AI) or principles adopted by organizations like IBM and the US Department of Defense.
- **Societal Adaptation**

Organizations influence AI's impact on the global economy. They may need to:

 - **Retrain employees** to transition into new roles due to job displacement.
 - Change assessment methodologies (e.g., academic professionals).
 - Educate employees and the public on the reality of **deepfakes, AI errors, and disinformation campaigns**.

- Reevaluate employee selection/placement processes using AI to **mitigate bias risk**.

- **Components of an AI Security Program**

An AI security program should follow the same practices as the existing security program but must address **AI-unique security gaps**.

- **Essential Elements (General):**

- Must align with and support **organizational objectives** (demonstrate execution of a well-developed strategy).
- Must be well-designed with **management and stakeholder cooperation**.
- Must use **effective metrics** (for design, implementation, and ongoing management).

- **Alignment with Organizational Objectives (COBIT Framework)**

Tailoring the **COBIT** governance and management objectives helps manage AI risk and ensures alignment with elements of **trustworthy AI**:

- **BAI06 (Managed IT Changes)**: Supports structured **change management** for frequent AI system updates (models/algorithms) to maintain resiliency and coordination.
- **EDM01 (Ensured Governance Framework Setting and Maintenance)**: Aligns the ethical development and use of AI with enterprise direction, goals, and societal values (fostering **transparency, fairness, accountability, and inclusivity**).
- **APO01 (Managed I&T Management Framework) & EDM02 (Ensured Benefit Delivery)**: Mandate that AI strategies adhere to **ethical standards** and monitor outcomes to align with ethical use cases.
- **DSS05 (Managed Security Services) & BAI09 (Managed Assets)**: Establish robust security controls for AI systems (protecting against adversarial attacks) and securing the infrastructure housing AI systems (data storage, model training).
- **APO14 (Managed Data)**: Outlines roles and responsibilities for managing metadata and data quality, which is critical for protecting **IP** and sensitive data in AI systems (supporting **Privacy**).

- **AI Security vs. Securing AI**

A robust program must consider both approaches across the enterprise and the entire AI life cycle.

Concept	Focus	Example
Securing AI	Security of the AI systems themselves (protecting models, training data, and ensuring models function as intended).	Implementing controls to protect an AI model against prompt injection attacks .
AI for Security	Incorporating AI technologies to support traditional security and cybersecurity processes (enhancing detection, prevention, and response).	Cybersecurity team using an AI algorithm to aid in analyzing network traffic patterns for threats.

- **Key Insight:** Security considerations must be applied at points previously not considered, such as securing **training data** to ensure fairness or prevent data leakage.
- **Senior Management and Stakeholder Support**
 - Senior leadership and stakeholders must provide **support and direction** for the AI security program's success.
 - To gain support, address workplace changes like **job displacement, reskilling/upskilling, and revision of existing roles**.
- **Program Metrics**

- AI governance requires defined program metrics to monitor and evaluate effectiveness.
- Metrics often align with **AI ethical principles**: bias, fairness, transparency, and explainability.
- **Challenge**: Lack of consensus on robust and verifiable measurement methods for risk and trustworthiness due to:
 - Lack of standard guidance.
 - Variability and rapid pace of AI solutions.
 - Oversimplification of measurement approaches.
 - Institutional ethical concerns.
- **Priority**: When creating metrics, consider the **potential harm to humans first**, followed by business factors (employee/customer satisfaction, ROI realization).
- **AI Key Performance Indicators (KPIs) and Key Risk Indicators (KRIs)**:
 - **KPIs (Performance Monitoring)**: Accuracy, precision, recall, compliance/governance adherence, data quality/volume, failure rate/downtime, impact on business KPIs, user adoption rate, ROI, and training/update frequency.
 - **KRIs (Risk/Bias Mitigation)**:
 - ◆ **Disparity Metrics**: Measure differences in outputs across population groups to identify bias.
 - ◆ **Fairness Metrics**: Ensure equitable benefits and equal proportions of positive outcomes across demographics.
 - ◆ **Frequency Metrics**: Track user reports of bias or compare expected vs. actual positive outcomes.

• AI-Enabled Security

AI plays a significant role in enhancing information security and cybersecurity. Best practices for vetting, development, and implementation must be followed.

▪ Artificial Intelligence and Network Security

AI helps security practitioners **find patterns and vulnerabilities** that might otherwise be missed.

▪ Common Uses:

- **Security Information and Event Management (SIEM)**: AI assists in the detection and correlation of log entries and events, providing context and suggesting courses of action.
- **Threat Detection/Prevention**: AI increases the difficulty for adversaries to bypass systems by analyzing complex, non-linear combinations of patterns.
- **Behavioral AI (BAI)**: Focuses on **patterns of human behavior** to detect anomalies indicating threats. BAI with ML speeds up trend analysis, strengthening internal cybersecurity controls.
- **Threat Intelligence**: GenAI tools provide new insights and detect previously missed threat alerts (e.g., Google Cloud Security AI Workbench).
- **Phishing Detection and Spam Filtering**: AI trained on acceptable/suspicious email patterns to block/quarantine threats.

• AI and Decision-Making

AI improves productivity by shaping business decisions, from automating manual tasks to complex threat analysis.

▪ Uses in Business Decision Making:

- **Project Management:** Automating manual tasks and making "small" business decisions to free up managers for strategic issues.
- **Vulnerability Analysis:** GenAI sifting through information to **prioritize patches** from thousands of options or implementing **zero-trust policies**.
- **Threat Detection:** LLMs based on previous breaches can quickly detect future breaches and develop counter-responses based on attacker actions.
- **Information Sharing:** GenAI creates **synthetic data** to simulate novel attack patterns for training other ML defense systems.
- **AI-Enabled Supply Chain and Vendor Management:** - AI supports data-driven vendor selection, contract automation, and performance monitoring, enhancing operational effectiveness.
 - **Risk:** Heightened risk of **data breaches** and reputational damage due to the large amount of data used in AI supply chain solutions.
 - **Risk Management Strategies:**
 - Conducting **due diligence** when reviewing vendors.
 - **Monitoring vendor performance** regularly.
 - Conducting **thorough and regular risk assessments**.
 - Implementing **strong cybersecurity controls**.
 - Conducting **employee awareness training**.
 - **Monitoring the supply chain for risk**.

5. Business Continuity and Incident Response

- **AI Incident Response:** Incident response must be structured to handle unique AI threats while ensuring **system reliability, security, and ethical compliance**.
 - **Goal of IR:** Identify events that deviate from normal operations, contain disruptions, and resume operations at acceptable levels quickly to reduce impact and cost.
 - **Challenges of AI in Incident Management:** Reliance on ML models, large datasets, and real-time decision-making.
 - **AI-Related Incidents:**
 - **Adversarial Attacks:** **Manipulating input data to deceive** AI models (e.g., misleading facial recognition).
 - **Data Poisoning:** Injecting **harmful/biased data into training** datasets to corrupt model outputs.
 - **Bias Exploitation:** Using inherent model biases to **generate unethical or unfair outcomes**.
 - **Model Drift:** Degradation of **AI model performance over time due to changing** data patterns.
 - **Unauthorized Access/Exploitation:** Hacking into AI systems to manipulate or extract sensitive information.
 - **Automation Failures:** Errors in AI-driven automation leading to incorrect or dangerous outcomes.
- **ISO 27035-1 Process for Handling Security Incidents:** - The phases of AI incident response follow a structured approach consistent with traditional plans but tailored to AI.
 1. **Prepare**

2. **Identify and report**
 3. **Assess**
 4. **Respond**
 5. **Lessons learned**
- **Prepare** for recovery and long-term resiliency.
 - An AI-ready IRP must address the unique, additive risk and the impact an AI incident has on the organization and human welfare.
 - **Policies, Procedures, and Model Documentation**
 - **IR Policy:** Must address AI-specific threats, detection/response techniques, and **ethical considerations**.
 - Decision-making over AI use should occur **outside** of active IR efforts.
 - **Model Documentation:** Maintaining thorough documentation (e.g., **model cards**) helps investigators interpret the model's architecture, training data, and decision-making.
 - **Communication Plan:** Include stakeholder escalation protocols to inform proper internal/external teams and law enforcement.
 - **Incident Response Team**
 - The team requires key stakeholder representation and subject matter experts beyond conventional IT and security responders.
 - **Data Stewards/Owners:** Experts on the **training dataset**.
 - **Data Engineers and Scientists:** Resources involved in data preprocessing and training; help interpret anomalous output.
 - **Privacy Experts:** Guide IR on impact to personal data, data subject rights, and regulatory compliance.
 - **AI Ethicists:** Responsible for the safe and ethical use of AI.
 - **Tabletop Exercises**
 - Train the AI IR team on SOPs and required resources (e.g., audit logs, forensic tools, and AI model cards).
 - **Adaptation:** Exercises must be tailored to AI-specific activities, such as
 - ◆ Dataset exploration and data input/output analysis.
 - ◆ Requires collaboration between IR and data engineering/science teams (e.g., to analyze a data lake for poisoning).
 - **Identify and Report**

The probabilistic nature of AI (it makes mistakes by design) makes detection challenging.

 - **Challenge:** Defining the organization's threshold for categorizing an **incorrect AI prediction** as an incident.
 - **Focus:** Categorize an incident based on its ability to cause harm, not just technical performance.
 - **Detection:** Some AI incidents are detected automatically; others require a **human-in-the-loop** for subjective judgments.
 - **Data Poisoning:** Harder to detect than traditional data loss because malicious data can be injected at many stages of the data pipeline.
 - **Detection Techniques for Common AI Attacks:**
 - **Prompt Injection:** Evaluate/sanitize data input; monitor inputs for atypical patterns (e.g., code structure, excessive special characters).
 - **Data Poisoning:** Monitor **access/change logs in the data supply chain**; review data preprocessing scripts for unauthorized changes; review dataset versioning and hashes.

- **Adversarial Inference:** Analyze API call logs; scan for anomalous patterns indicating systematic adversarial testing.
- **Assess:** Collect facts to establish the **timeline, scope, and impact** without damaging the integrity of the investigation.
 - **Model documentation** is key.
 - **Key Assessment Questions:**
 - Has the incident stopped?
 - Who was impacted, and what harm occurred?
 - When was the incident discovered (for breach notification)?
 - What exactly happened, which AI systems/data were impacted, and what attack tactics were used?
 - What is still unknown to return to a safe operational state?
- **Respond:** - Conventional **containment, eradication, and recovery** strategies are relevant but less effective due to the complexity and probabilistic nature of AI.
 - **Containment:** Immediate action is needed if there are **threats to human life and safety**. Containment can be difficult due to AI environment complexities.
 - **Containment Techniques:**
 - **Prompt Injection:** Deploy **data input/output validation** and **prompt templates** to screen and sanitize abuse prompts, shielding the AI model.
 - **Data Poisoning: Revoke access** to datasets and code/scripts that execute data preprocessing functions in the data pipeline.
 - **Adversarial Inference:** Implement additional **data input throttling, validation, and sanitization** to reduce systematic inference attacks.
 - **Eradication:** - More challenging than patching conventional software. The technique depends on the attack.
 - **Eradication Techniques:**
 - **Prompt Injection:** The underlying AI model may need to be **retrained or fine-tuned** (can be costly). **Data input validation and prompt templates** are effective long-term controls.
 - **Data Poisoning: Isolate and remove** poisoned data, then **retrain** a new model version (costly/time-consuming).
 - Short-term: data output sanitization.
 - Long-term: **selective retraining** on clean data to force the model to "forget" poisoned data.
 - **Adversarial Inference:** Use **regularization techniques** and defensive distillation to make the AI model more resilient.
- **Recovery:** - Ensure the threat is contained and eradicated.
 - **Resolution:** Reduce the attack surface, actively monitor residual threats, and continuously improve models for **robustness and resilience**.
 - **Action:** Implement stricter access controls, new data input validations, and other guardrails.
 - **Final Step: Post-incident validation** of new control measures and stakeholder approval before reactivating the AI solution.
- **Post-incident review:** -
 - Conduct a postmortem to identify areas for improvement: data preprocessing, security controls, sufficiency of adversarial testing, input controls, fairness of outputs, and the AI provider's processes.

- The goal is to move from a reactive to a **proactive readiness posture**.
- **Traditional Incident Response vs. AI-powered Incident Response:** - AI- powered IR enhances efficiency, scalability, and decision-making.

Aspect	Traditional Incident Response	AI-Powered Incident Response
Data Analysis	Manual analysis of logs and alerts	Automated with AI models analyzing large datasets
Incident Triage	Based on predefined rules and human evaluation	Automated using AI systems
Response Time	Slower due to manual processing	Faster, real-time responses due to automation
Scalability	Limited by human capacity	Highly scalable with AI systems
Root Cause Analysis	Time-consuming manual investigation	Swift and automated root cause identification
Decision-making	Relies on human expertise and predefined procedures	Enhanced with AI driven insights and predictive analytics

- **AI-Enabled Incident Response Mechanisms:** - AI systems continuously learn and enhance accuracy and response times through:
 - **Data Ingestion and Normalization:** Collect and normalize data from multiple sources (logs, network traffic, threat intelligence).
 - **Anomaly Detection:** Use pattern recognition and ML to identify deviations from normal behavior for **early detection**.
 - **Event Correlation:** Correlate data across sources to detect complex, multistage attacks.
 - **Automated Incident Triage:** Categorize incidents based on severity to reduce burden and ensure timely response.
 - **Root Cause Analysis:** Accelerate identification of the problem source to reduce resolution time.
 - **Response Automation:** Automatically execute predefined actions (e.g., isolating systems, blocking IP addresses).
 - **Continuous Improvement:** AI models continuously learn from past incidents to refine detection and response.
- **Use Cases of AI-Driven Incident Response**
 - **Detection and Alerting:** Continuous monitoring of network traffic, logs, and user behavior for real-time anomaly detection.
 - **Root Cause Analysis:** Quickly analyzing and correlating events across systems for faster, more accurate remediation.
 - **Incident Resolution and Automation:** Automating actions (isolating systems, blocking traffic, applying patches) to reduce human intervention during critical phases.
 - **Post-incident Analysis and Learning:** Continuous analysis of incident data to identify patterns and gaps, contributing to continuous improvement of IR strategies.
- **Benefits of AI-Powered Incident Response**
 - **Enhanced Detection/Response:** Analyze large volumes of security data in real time, catching threats missed by traditional methods.
 - **Faster Response Times:** React instantly to security incidents, minimizing damage.
 - **Accelerated root cause analysis** and resolution.

- **Improved Accuracy:** Minimize false positives/negatives.
- **Scalability:** Handle large data volumes and growing incidents without increasing headcount.
- **Challenges of AI-Powered Incident Response**
 - **Accuracy and Trust Issues:** AI can produce inaccurate results or false positives, leading to distrust. Requires **continuous monitoring and refinement**.
 - **Human Oversight:** Critical to ensure AI-driven actions align with business objectives and ethical standards, avoiding unintended consequences.
 - **Evolving Threat Landscape:** AI models must be kept updated with the latest threat intelligence to remain effective.
 - **AI Biases:** Biased training data leads to skewed results, potentially overlooking threats or causing disproportionate responses.
- **Incident Response Automation Best Practices**
 - **Seamless Integration:** Integrate AI with existing tools (e.g., SIEM) to create a unified response system.
 - **Adaptive Learning Algorithms:** AI should continuously evolve to refine responses based on previous incidents.
 - **Customizable Automation Playbooks:** Implement flexible playbooks tailored to company needs.
 - **Strategy Updates:** Regularly update automation strategies to align with emerging threats.
 - **Incident Prioritization:** AI dynamically prioritizes incidents based on severity and potential impact.
 - **Collaboration Tools:** Facilitate smooth AI-human collaboration.
 - **Compliance Automation:** Integrate compliance checks into workflows (e.g., with the EU AI Act, GDPR).
 - **Continuous Learning:** AI reviews performance after each incident to identify improvements.

Domain 2: AI Risk and Opportunity Management

1. AI Risk, Threat, Vulnerability, and Supply Chain Management

- **AI in Risk Management and Cybersecurity:** - AI is increasingly being leveraged in IT and cybersecurity for risk management tasks like **threat modeling**. Tools like **STRIDE GPT** and **Irius Risk** use AI to generate risk and threat scenarios based on common frameworks.
- The combination of **omnipresent internet connectivity and AI** is fundamentally reshaping global society and enterprise operations, presenting a dual nature:
 - **Benefits:** Enhanced social interaction, creation of new art/content, efficiencies in tasks like candidate screening.
 - **Risks:** Deepened isolation, dilution of traditional expression, perpetuation of discriminatory practices, accelerating job displacement.
- Information security traditionally focuses on **Confidentiality, Integrity, Availability (CIA), and Privacy**.
- AI adoption expands threats and vulnerabilities to encompass both **technical** and **nontechnical** risks to the enterprise and society.
- AI Risk Assessment Frameworks, Acceptable Limits, and Responses: - Artificial Intelligence (AI) introduces significant, complex risks (security, privacy, safety, well-being) that require a structured risk management approach to identify, classify, and manage potential harm.
 - **Core AI Risk and Program Elements:** - The primary **Goal** of an AI risk program is to manage AI risk in alignment with **strategic goals** and balance the needs of affected stakeholders (e.g., assessing an AI fraud system for reliability, bias, and compliance using frameworks like **NIST AI RMF**).
 - **Common AI Risks**

Area	Examples of Risk
Development	Inaccurate modeling, Biased (discrimination)
Data	Data quality issues, Poisoning
Operations	Excessive Agency, Security, Safety, Skill Set/Competencies
LLM/AI	Unbound Consumption, Injection, Extraction
Trust	Opaque, Privacy, Misinformation, Manipulation
Compliance	Disclosures, Unethical, Environmental

- **Elements of an AI Security Program:** - A robust AI security program should address risk through practical elements such as:
 - **Verify Then Trust** and **Consider Societal Adaption**.
 - Define **AI Acceptable Use Policies** and **AI Ethics**.
 - Identify an **AI Czar** and **Ensure Traceability**.
 - Perform **Cost-Benefit Analysis** and **Adapt Cybersecurity Program**.
- **AI Trust:** - AI systems can learn, adapt, and evolve in **unpredictable ways**, making accurate risk accounting problematic. Risk extends beyond technical considerations due to AI mimicking human behavior. Therefore, AI systems must be **trustworthy at their core**.
 - **Enterprise Risk Pyramid:** - AI risk layers on top of existing risk categories, including:

- **AI Risk:** Ethics and Biases, Data Privacy, Data Security, Operational, Societal, Costs, Transparency, Explainability, Strategic.
 - **AI Trust Attributes key attributes for trustworthy AI:**
 - Safe
 - Secure and Resilient
 - Explainable and Interpretable
 - Privacy Enhanced
 - Fair With Harmful Bias Managed
 - Accountable and Transparent
 - Valid and Reliable
 - **AI Risk Identification: Risk is contextual** and requires understanding the full scope of the AI application (e.g., in HR, supply chain).
 - **The Risk Management Life Cycle remains relevant:**
 - IT Risk Identification
 - IT Risk Assessment
 - Risk Response and Mitigation
 - Risk and Control Monitoring and Reporting
 - **Identification Process:** - The process involves
 - **Inventorying** AI use within the organization.
 - Creating and reviewing AI-related loss event scenarios by: mapping missions, determining services, identifying vital assets, establishing an AI value chain, and assessing AI impact.
- **OWASP Top 10 for LLM Applications:** - This list highlights critical security risks specific to Large Language Models (LLMs):

Attack Name	Attack Description
Prompt Injection	Attacker uses prompts to nefariously alter LLM behavior.
Sensitive Information Disclosure	Sensitive data (PII) is exposed due to use in the LLM.
Data and Model Poisoning	Manipulating training/validation data to introduce vulnerabilities or biases.
Excessive Agency	LLM gains ability to make decisions/perform actions via extensions based on prompts.
Misinformation	LLMs produce false, credible information (often due to hallucinations).
Unbounded Consumption	End users draw excessive inferences, causing DoS attacks or financial loss.

- **AI Risk Frameworks:** - Focused, tailored frameworks are essential for managing AI risk.
 - **NIST - AI RMF**
 - **Type: Voluntary** framework providing best practices.
 - **Recommended for:** Organizations considering the design, development, deployment, or use of AI systems.
 - **Four Key Functions:**
 1. **Govern:** Establish risk policies, accountability, and governance.
 2. **Map:** Identify AI system characteristics, risk areas, and potential impact zones.
 3. **Measure:** Develop metrics and monitoring to manage identified risk.
 4. **Manage:** Implement risk response strategies and compliance checks.

- **Classification:** Follows a flexible, **industry-agnostic** approach; does not prescribe rigid risk classification thresholds.
- **European Union Artificial Intelligence Act (EU AI Act)**
 - **Type: Regulatory/Mandatory** framework establishing legally binding obligations and penalties. Applies extraterritorially.
 - **Goal:** Establish a risk-based framework for AI governance.
 - **Four Primary Risk Categories:**
 - **Unacceptable Risk:** Prohibited AI systems (e.g., biometrics/social scoring).
 - **High Risk:** Requires strict oversight, conformity assessments, and extensive documentation (e.g., critical infrastructure, law enforcement).
 - **Limited Risk:** Requires transparency (e.g., chatbots).
 - **Minimal Risk:** Little or no regulatory oversight required (e.g., games, spam filters).
 - Organizations deploying High-Risk AI must conduct **Conformity Assessments**.
- **Key Differences (EU AI Act vs. NIST AI RMF)**

Key Difference	EU AI Act (Regulatory)	NIST AI RMF (Voluntary)
Risk Classification	Detailed, specific categories (Prohibited, High-Risk) with mandatory measures.	Flexible, industry-agnostic approach. No fixed categories or mandated thresholds.
Roles/Responsibilities	Differentiated Stakeholders (Providers, Deployers) with tailored obligations .	Broad Stakeholder Definition ("AI actors"). General guidance, leaves role delineation flexible.

- **AI Risk Classification and Acceptable Limits:** - Organizations must establish AI risk categories and define **acceptable risk limits** (risk appetite and tolerance).
 - **Risk Classification Views**
 - **Qualitative View (EU AI Act):** Unacceptable, High, Limited, Low/No risk.
 - **Quantitative View (FAIR-AIR Approach):** Translates AI-related risk into **financial terms** for better communication to senior leadership.
 - **Process:** Contextualize Scope Quantify (financial loss), Prioritize/Treat, Decision Making.
 - **Essential Assessments for Compliance and Trust**
 - **Fundamental Rights Impact Assessment (FRIA):** **Mandated by EU AI Act** for High-Risk AI deployed by public bodies. Ensures the AI system does not negatively affect individual **fundamental rights** (e.g., privacy, equality).
 - **Conformity Assessments:** **Mandated by EU AI Act** for all High-Risk AI systems to demonstrate compliance with requirements.
 - **Privacy Impact Assessment (PIA):** Analysis of how personal information is collected, used, shared, and maintained. Essential for addressing **AI-driven privacy risk** like algorithmic bias and inference-based privacy risk.
 - **Acceptable Risk Limits**
 - **Responsibility:** The organization's governing body is responsible for establishing **risk appetite and tolerance**.
 - **Factors to Consider:** Cost, Benefits, Compliance Requirements (e.g., EU AI Act), Culture, Uncertainty, Policies, and Value.

- **AI Risk Response Strategies:** - **Appropriate risk** is necessary to pursue objectives; **inappropriate risk** exceeds appetite and can cause harm.

Strategy	Context/Goal	AI Example
Accept	Risk is within acceptable limits; cost of mitigation outweighs benefit.	Financial institution accepts a known 5% false positive rate in its AI fraud detection.
Avoid	Informed decision to <i>not pursue</i> an action because risk cannot be brought within limits (opportunity loss).	Corporation avoids deploying an AI-driven hiring system due to high potential for discriminatory practices and fines.
Mitigate	Lower the probability or resulting impact of a risk through efficient controls.	Healthcare provider improves training data/algorithm and adds Human -In - The Loop (HITL) oversight to an AI diagnosis system.
Transfer/Share	Impact of risk is partially moved to an external third party (e.g., insurance).	Car manufacturer transfers liability for autonomous self-driving car accidents to insurers.

- **AI Threat Modeling:** - Threat modeling is a key part of risk analysis, identifying actors, goals, methods, and required controls to create a clearer picture of **end-to-end risk**.
 - **Threat Model Methods**

Model	Traditional Focus	AI Considerations/Lacks
STRIDE	Spoofing, Tampering, Repudiation, Info disclosure, DoS, Elevation (Mature)	Foundation for basic vulnerabilities, but lacks ability to address AI-unique challenges (adversarial ML, data poisoning).
LINDDUN	Privacy concerns (Linkability, Identifiability, Non-repudiation)	Useful for data security, but lacks ability to model AI-specific threats , including autonomous decision making.
MAESTRO	Proposed model building on existing ones.	Includes focus on AI-specific categories (multi-agents, environmental) and considers adversarial machine learning and AI autonomy.
Trike	Risk Management and Threat Mitigation Prioritization across the overall threat environment. Integrates risk assessment and prioritization systems for completeness.	Difficulty addressing internal systems and related threats. Has vague, insufficient documentation .
VAST	Agile Risk Management and Mitigation Prioritization for iterative development. Focuses on scalability, consistency, and automation.	Lacks specificity for AI-specific threats and vulnerabilities.
OCTAVE	Critical Asset Identification (including information and data) and risk evaluation within an operational context .	Lacks the ability to specialize to AI-specific threats and challenges, such as attacks on data and models. Relies on strong risk management systems to be effective.

	Creates practice-based protection strategies and risk mitigation plans.	
--	---	--

Adaptation is Key: Existing techniques must be adapted to include considerations for **ethics, privacy, and human rights**.

2. Threat and Vulnerability Management

- AI introduces a unique and rapidly evolving threat landscape, a **fusion of technical and human-centric vulnerabilities**.
- These threats arise from **model development, training, operation, interaction** with other systems, and real-world outcomes.
- Understanding these categories is vital for proactive risk management.
- **The AI Threat Landscape (Key Categories)**
 - **AI Regulatory Risk:** Threats from fragmented or inconsistent AI regulations across different regions.
 - **Adversarial Attacks:** Specific attacks designed to **trick the AI system** (e.g., misinterpreting a stop sign).
 - **Data Poisoning:** Adversaries manipulating the **training data** to cause biased, inaccurate, or malicious outputs.
 - **Data Privacy and Security:** Risk of breaches or misuse of sensitive data processed by AI systems, leading to legal and reputational harm.
 - **Insider Threats:** Malicious or negligent insiders stealing proprietary data, model theft, or sabotaging models.
 - **Lack of Explainability/Transparency (Black Box):** Inability to share details of the AI model, making it hard to identify errors, biases, demonstrate regulatory compliance, and build trust.
 - **Model Theft/Reverse Engineering:** Stealing an enterprise's proprietary algorithms and parameters to replicate AI-driven services.
- **Systemic Risks (External Dependencies):** - Systemic risks introduced through external dependencies:
 - **Supply Chain Vulnerabilities:** Vulnerabilities in any component relied upon by the AI system (e.g., data sources, GPUs, software libraries) can compromise the entire system. Security managers must consider **hardware alterations** and **compromised datasets/pipelines** beyond just software.
 - **Vendor Lock-in/Dependency:** Reliance on a single AI provider, making the organization vulnerable if the provider experiences an incident (outage, breach).
- **Corresponding Enterprise Risk Categories:** - AI threats correspond to familiar enterprise risk categories:
 - **Competitive:** Loss of market share or stakeholder confidence due to data/model theft or adversarial attacks.

- **Fines/Penalties:** Legal action and financial penalties for noncompliant AI operations or privacy violations.
- **Productivity:** Operational downtime due to system failures.
- **Reputational:** Loss of public trust from security incidents or failures.
- **Response:** Costs associated with recovering from breaches or service disruptions.

- **Examples of AI Threat Actors and Scenarios**

Threat Actor	Description	AI Threat Scenarios
Insider threats	Employees causing intentional or unintentional harm.	Model sabotage, data exfiltration.
Nation states	Using AI to advance strategic interests (espionage, warfare).	Using AI for surveillance, censorship, or monitoring citizens.
Cybercriminals	Leveraging AI to make attacks more effective and harder to detect.	Using AI to create convincing phishing emails or advanced malware.
AI developers	Creating AI solutions misaligned with ethical standards.	Designing a biased algorithm for a credit rating agency leading to unjustified denial of credit.

The **MITRE Adversarial Threat Landscape for Artificial-Intelligence System (ATLAS)** is a knowledge base for security teams to address AI threats.

Reference (<https://atlas.mitre.org/>)

- **Technical Threats and Vulnerabilities:** - AI systems share threats with traditional software, but models, algorithms, and data introduce unique **technical threats**.

Threat Category	Shared Threats	Artificial Intelligence Threat Examples
Advanced threats	Zero-day exploits	Model backdoor attack
Availability/Reliability	Infrastructure failures	Model drift
Data security	Data breaches	Model inversion
Denial of Service (DoS)	DoS attacks	Adversarial attacks
Emerging AI-specific issues	N/A	Hallucination attacks, Prompt injection attacks
Human factors/insider	Malicious insiders	Unintentional bias amplification, Insecure design
Identity/Authentication	Credential stuffing	Deepfakes
Malware/Exploits	Malware attacks	Data poisoning
Supply Chain/Third-party	Dependency exploits	Data supply chain attacks
Unauthorized access	Credential theft	Theft of services, Prompt injection

- **AI Attack Surfaces (OWASP AI Exchange)**

- **Development-time threats:** Occur during development (e.g., in the engineering environment or supply chain).
 - **Runtime security threats:** Attacks due to non-AI-specific security weaknesses (common IT vulnerabilities in software wrappers or infrastructure).
 - **Threats through use:** Occur during normal model usage, particularly around input and output.
- **Examples of AI Technical Threats**
 - **Training Data Leakage:** Data leaked/stolen from the training environment due to weak access controls or poor exfiltration controls (e.g., in Jupyter Notebooks).
 - **Data Poisoning:** Malicious or erroneous data inserted into datasets to change a model's output. Entry points include source, supplier, in-transit pipelines, preprocessing, or in-context training (RAG).
 - **Impact:** Creating a **backdoor/Trojan horse** or **sabotaging** the model's safety/performance.
 - **Model Poisoning:** Attacker-induced changes to model behavior via
 - data poisoning,
 - directly changing model parameters/architecture/libraries, or
 - poisoning a third-party, pre-trained model.
 - **Model Theft:** Stealing the proprietary model (source code, configurations, binaries) or inferring its parameters and features through crafted runtime inputs/outputs.
 - **Prompt Injections:** Adversarial attack using crafted text prompts to make a GenAI model ignore original instructions and follow the attacker's instructions (similar to SQL injection).
 - **Mitigation:** **Prompt templates** for structure and sanitization, **runtime output validation**.
 - **Model Evasion:** Attacker modifies data to circumvent the model's correct prediction, causing misclassification or errors.
 - **Model Inversion:** Attacker extracts sensitive or private information by analyzing the model's inputs and outputs to infer training data attributes (**attribute inference**) or whether a specific record was in the training set (**membership inference**).
 - **Nontechnical Threats:** - Nontechnical threats pose a unique challenge as they cannot be addressed with traditional technical patches (e.g., ethical concerns, biased decision-making).

Category	Threat	Description
Ethical concerns	Hallucinations	System generating misleading, harmful, or false content
Bias and fairness	Bias amplification	AI models reinforcing societal inequalities
Dependency	Overreliance	Users assuming decisions and information from AI are infallible

Workforce impact	Job displacement	Replacing humans in the decision-making process. Replacing human roles altogether
Societal and cultural impacts	Societal manipulation	Deepfakes, Automated misinformation, Content inauthenticity and fake media, Voice cloning
Financial risk	Compute costs	Cloud/compute expenses required for AI workloads
Staff competency	Skill gaps	Shortage of AI-trained professionals
Brand and reputation	AI scandals	Unethical, biased, or harmful decisions that lead to mistrust and financial losses
Regulatory and compliance infractions	AI-specific regulations	Strict AI transparency and bias laws (e.g., EU Artificial Intelligence Act)
Lack of transparency and explainability	Black-box problems	Inability to explain AI-driven decisions
Strategic misalignment	AI hype adoption	Adding AI to products without real business value
Consumer distrust	Deceptive concerns	Fear of manipulated content, AI-driven fraud
Legal violations	Legal risk	Bias in hiring AI, Facial recognition lawsuits, Contractual agency
Privacy intrusion	AI-specific privacy concerns	Facial recognition, AI-powered surveillance
Market manipulation and abuse	Market manipulation	Algorithmic trading exploits, Dynamic pricing abuse
Rogue AI behavior	AI unpredictability	Self-learning models making harmful decisions
Cybercrime	Autonomous attacks	AI-generated phishing, Deepfake scams
Environmental and sustainability concerns	Excessive power and water consumption	High energy demand for training AI models decreasing availability for consumers, High demand of water supply to cool servers needed for generative AI queries
Global geopolitical and trade restrictions	Export controls	AI chip and algorithm restrictions in global trade
Human-machine collaboration challenges	Trust issues	User misunderstandings or misinterpretations of AI outputs
Content inauthenticity and fake media	Deepfakes	Fake media, Voice cloning, Automated disinformation

- **AI-Enabled Threats:** - AI is leveraged by threat actors to create or optimize attacks:
 - **Adversarial Models:** Models created specifically to exploit LLM capabilities (e.g., DeepSeek, Owen).

- **Convincing Fabricated Content:** Deepfake audio/video circumventing traditional phishing training.
- **Optimizing Cyber Kill Chain:** Using AI (like ChatGPT) to optimize code for sophisticated malware, ransomware, and phishing.
- **Mitigation Strategies:** - Mitigation should be chosen based on the threats and include:

Threat Category	Potential Mitigations
Advanced threats (Model backdoor)	Continuous monitoring, AI fairness/bias testing.
Availability/Reliability (Model drift)	Automated monitoring, AI model retraining pipelines.
Bias and Fairness (Bias amplification)	Bias detection/fairness audits, diverse training datasets.
Content Authenticity (Deepfakes)	Watermarking , Digital signatures, Media forensics, Public awareness.
Data Security (Model inversion)	Strong cryptography, IAM, Query parameterization.
Denial of Service (Adversarial attacks)	Rate limiting , Baseline modeling, Anomaly detection.
Dependency (AI overreliance)	Human in the Loop (HITL) requirement, Redundant decision pathways.
Emerging AI-specific risk (Hallucinations)	AI validation/adversarial testing, Structured prompt engineering .
Ethical concerns (Ethical boundaries)	AI ethics frameworks/guidelines, AI governance programs.
Unauthorized Access (Prompt injection)	Input sanitization , Output validation , MFA, Secure session management.
Transparency/Explainability (Black-box problem)	Explainability policies, detailed AI model documentation.

3. AI Vendor and Supply Chain Management

Enterprises rely on third parties, and AI adoption introduces new vendor relationships and impacts existing ones.

- **Vendor, Provider, and Supply Chain Risk:** - Third-party relationships introduce various risks:
 - **Operational:** Production delays, manufacturing defects, labor issues.
 - **Financial:** Supplier bankruptcy, credit risk, price increases.
 - **Geopolitical:** Political instability, sanctions, regional conflict.
 - **Environmental:** Floods, wildfires.
 - **Compliance:** Regulatory, contractual, ethics, safety.
 - **Cybersecurity:** Data breaches, ransomware, IT disruptions.
 - **Reputational:** Negative publicity from supplier misconduct.

- **Roles in the AI Supply Chain:** - Understanding the enterprise's role is critical as it impacts responsibilities and risk.
 - **AI Provider (Vendor):** Entity supplying the AI solution.
 - **AI Deployer (User/Subscriber):** Entity using/integrating the AI solution.
- **Key Differences (Based on EU AI Act)**

Providers (Vendors)	Deployers (Users/Subscribers)
Compliance: Ensuring AI systems meet legal requirements.	Use as Directed: Using systems per provider instructions.
Conformity assessment: Verifying high-risk AI meets standards.	Oversight and HITL: Assigning qualified individuals to oversee AI use.
Documentation: Creating and maintaining detailed AI system documentation.	Data management: Controlling data provided to AI systems.
Monitoring: Tracking AI performance and addressing issues.	Monitoring: Monitoring AI responses and reporting issues.
Incident reporting: Reporting serious incidents to authorities.	Transparency: Informing people when they interact with AI systems.

Contextualization: An internal AI system gives the enterprise the **most control** over risk. A third-party AI system gives the enterprise the **least control**. An enterprise can be **both** a provider and a deployer.

- **AI Vendor Management:** - Existing vendor management programs must be updated to include AI-specific risks:
 - **Business Risk:** Disruptions from new AI features/functions.
 - **Privacy Risk:** Violations from how vendor models are trained and data is handled.
 - **Security Risk:** New exposures/vulnerabilities from new AI features.
 - **Compliance Risk:** Vendor's ability to adapt to emerging AI regulations.
- **An Approach to AI Vendor Management:** - This approach prioritizes due diligence **before** vendor lock-in:
 - Define internal diligence and the need for AI.
 - AI Vendor Inventory and Scope Definition.
 - **Adapt Existing Due Diligence to Include AI-Specific Risk:**
 - Evaluate Vendor Reputation and Financial Stability.
 - Assess Vendor Practices, Processes, and Alignment.
 - **Define Risk Appetite:** AI System Instrumentation, Use, Abuse, and Misuse.
 - Establish Compliance and Incident Notification Criteria.
 - Develop Practices for Routine Monitoring and Reassessment.
 - Establish Triggers for Renewed Risk Assessment.

- **High-Risk AI Vendor Considerations:** - Primary concerns when evaluating a third-party AI solution:
 - Lack of **transparency and explainability**.
 - Unacceptable **ethical risk** and high potential for **data privacy violations**.
 - **Vendor lock-in**.
 - **Immature** or unproven AI technology (especially if impacting health/safety).
 - Lack of a risk management framework that incorporates AI.
 - Inability to meet AI security requirements (including **content provenance** for GenAI).
 - Resistance to due diligence/auditing.
- **AI Deployer Considerations and Shared Responsibility**
 - **Deployer Responsibility:** Even with a third-party AI system, the deployer is responsible for managing the **risk and impacts** of the system in its deployed environment.
 - **Example Liability:** If an airline's chatbot (provider's AI) grants a non-policy refund, the airline (deployer) is legally bound, as the chatbot is its authorized agent.
- **AI Shared Responsibility Model:** - Similar to cloud computing, the responsibilities are divided between the **Provider** and the **Deployer**. Clear delineation of responsibilities should be **contractually addressed** (SLAs, incident response, notification).

AI Deployer Responsibilities (Consumers)	AI Provider Responsibilities (Vendors)
Audits and testing (e.g., adversarial testing).	Compliance and audit certifications (adherence to standards).
Strong data governance (quality, integrity, provenance).	Data security (protecting training data, model weights, artifacts).
Identity and access management.	Incident response (robust plan, coordination with subscribers).
Incident response planning (AI-centric plan development, training, testing).	Model development (ethics framework, secure coding, biased-reduced data).
Model validation (including adversarial testing to remove bias/vulnerabilities).	Transparency and explainability (notifications, tools, documentation).
Supply chain risk management (vetting security, reputation, including contractual requirements).	Vulnerability management (proactive identification and patching).
Transparency and explainability (ensuring user awareness, monitoring performance/outputs).	System Documentation, Sources coding and ethics framework

- **Integration Risk**
 - **Legacy Systems:** - AI integration with legacy systems poses challenges:
 - They may be **difficult to integrate** and often **not well-suited for AI**.
 - **Considerations:** Determine if AI makes sense, check integration capabilities, assess need for complete overhaul, evaluate data

quality/availability, define flexible methodology, and secure required support.

- **Intellectual Property (IP):** - IP risk is a major concern with AI-generated content.
 - **Training Data Sourcing:** AI models are trained on vast data, often from the public web, which may include **copyrighted material**.
 - **Litigation Risk:** Organizations can unknowingly introduce pilfered content, leading to lawsuits and penalties.
 - **Ownership Clarity:** The true IP holder of AI-generated content or code is often uncertain, especially if the model pulls from the provider's dataset. **Ownership provisions must be clearly defined in contracts.**
- **AI Software Supply Chain Risk:** This risk extends beyond individual components to **entire systems**, requiring management to protect brand and reputation. The supply chain involves multiple dimensions: **people, processes, technology, data, and models**.
 - **Supply Chain Parties:** - The ecosystem extends far beyond the direct provider:
 - **First Party:** The enterprise (consumer).
 - **Third Party:** Entities directly contracted by the provider (e.g., cloud, data providers, model developers).
 - **Fourth Party:** Vendors relied upon by the third party (vendors of vendors).
 - **Nth Parties:** Continuing the chain of outsourcing.
 - **Emerging Best Practices for Monitoring:** - Continuous monitoring and robust security are essential:
 - Choose the right **LLM monitoring metrics** (intrinsic and extrinsic).
 - Set up effective **logging, alerting, and monitoring** (for prompt quality, output relevance, bias, and toxicity).
 - Run **adversarial tests** (token manipulation, jailbreak attacks).
 - Maintain strict **data integrity and model input standards** (check data quality, monitor for model drift).
 - **Never trust, always verify** (implement tests for accuracy, push systems to cite sources).
 - **Perform a sanity check** (use multiple AI systems and compare results).

Domain 3: AI Security Architecture and Life Cycle

Overview and Context

- **Evolving Landscape:** The information security professional's skillset must continuously evolve due to rapid technological changes, especially in AI.
- **AI Development vs. Regulation:** AI technology is advancing faster than standardization and regulation. Security professionals must establish **guardrails** to safeguard AI solutions while supporting **business strategy**.
- **AI Complexity:** Unlike traditional IT adoption (e.g., cloud), AI is complex. Inputs and outputs in AI models pose significant risks regarding **data privacy, ethics, and misleading results**.
- **Increased Attack Surface:** The evolving use of AI increases the **attack surface** and introduces new types of exploits.
- **Comprehensive Strategy:** An essential AI strategy must cover **design and architecture, life cycle management, and data classification**.

1. AI Security Architecture and Design

- **AI Architecture Fundamentals:** - The architecture of AI includes a nested hierarchy of concepts:

Concept	Description
Artificial Intelligence (AI)	Techniques enabling computers to emulate human behavior (learn, make decisions, recognize patterns, solve complex problems). Broad term, from simple rule engines to self-driving cars.
Machine Learning (ML)	A subset of AI using advanced algorithms to detect patterns in large datasets, allowing machines to learn and adapt (supervised or unsupervised learning).
Deep Learning (DL)	A subset of ML that uses neural networks (NNs) with multiple layers for in-depth data processing and analytical tasks, simulating how human brains perceive data.
Generative AI (GenAI)	A subset of DL models that generates new content (text, images, code) based on provided input.
Data Science	A multidisciplinary field combining computer science, math, statistics, AI, ML, DL, and visualization to extract insights from large amounts of data.

- **Types of AI:** - AI can be categorized by Functionality or Capability.
 - **Types of AI Based on Functionality**

Type	Overview	Examples
Reactive Machines	Most basic, generates an automatic response to a specific, limited set of inputs . No memory of past experiences.	Gaming (Deep Blue), Industrial Robotics.
Limited Memory	Possesses a brief memory recollection ; uses recent observational data for a specific task but cannot store intricate repositories of past experiences.	Autonomous vehicles, Virtual assistants (Siri, Alexa), Recommendation systems, Chatbots.

Theory of Mind (ToM)	Hypothetical. Machines capable of understanding and mimicking human thought patterns, ideas, and social-emotional behavior . Requires understanding that thoughts and emotions impact human behavior.	Social robots (Pepper, Sophia) - <i>Early examples, not full ToM.</i>
Self-Aware AI (ASI)	Hypothetical. Machines aware of others' emotions/mental states and their own . Has not been achieved.	None currently.

• Types of AI Based on Capabilities

Type	Overview	Connection to Functionality
Artificial Narrow Intelligence (ANI)	"Weak AI." Limited to a specific domain or area of knowledge. High accuracy/efficiency within its scope, but lacks self-awareness, consciousness, or general human intelligence.	Classified as Reactive AI and Limited Memory AI .
Artificial General Intelligence (AGI)	"Strong AI." Capable of undertaking any intellectual task a human can perform . Must possess sensory perception, fine motor skills, problem-solving, navigation, creativity, etc. (still uncertain timeline).	Connects to the eventual possibility of Theory of Mind .
Artificial Superintelligence (ASI)	Hypothetical stage that transcends human intelligence . Surpasses human capabilities in all intellectual tasks.	Connects to the goal of Self-Aware AI .

• Generative Models

- **Generative AI (GenAI):** Built on patterns and examples from training data to generate **new and original content**.
- **Generative Adversarial Networks (GANs):** A revolutionary class of DL models using a **dual neural network system** (Generator vs. Discriminator) that compete.
 - **Generator Network:** Creates new, realistic data samples.
 - **Discriminator Network:** Distinguishes between real and generated samples.
 - **Result:** The generator continuously improves until it produces data indistinguishable from real data.
 - **Risk:** Legal and ethical risk associated with harmful content distribution, **copyright ownership, infringement, and intellectual property (IP)**.

• Agentic and Predictive Models

- **Agentic AI:** Next evolution of AI assistants. AI agents make **autonomous decisions and act to achieve specific objectives** (e.g., maximizing supply chain efficiencies).

- **Benefits:** Greater workplace specialization, ease of innovation/experimentation, increased trustworthiness (less likely to "hallucinate" than GenAI).
- **Requirement:** Requires **human oversight** to optimize effectiveness and ensure goals are clearly articulated.
- **Predictive Models:** Use historical data and ML algorithms to **forecast events and trends**.
 - **Risk:** Need to ensure full understanding of how models derive predictions to ensure decisions are based on **verifiable facts**.
- **Machine Learning Models and Paradigms:** - ML models learn from data without explicit programming, based on mathematics (statistics, probability, etc.).
 - **Basic ML Paradigms**

Paradigm	Data Type	Key Feature	Risk/Considerations
Supervised Learning	Labeled data (input right arrow output pairs).	Maps input data to an output label to make predictions on new data.	Data Quality and Bias are the greatest risks, as it relies heavily on high-quality, non-biased labeled data.
Unsupervised Learning	Unlabeled (raw) data.	Learns by detecting underlying patterns, relationships, and structures without explicit guidance (e.g., clustering, dimensionality reduction).	Lack of explainability of results. Requires Human-in-the-loop (HITL) for accuracy and reliability.
Reinforcement Learning (RL)	Learns through consequences of its actions (rewards or penalties) in a specific environment.	Trains intelligent agents for autonomous decision-making and control tasks to maximize expected cumulative reward.	Autonomous nature of the model; safety and ethics during the exploration phase are imperative. HITL is crucial.

- **Supervised Learning Types:**
 - **Regression:** Predicts the most likely outcome from a **continuous set** (e.g., forecasting sale prices). (Quantitative results).
 - **Classification:** Identifies the most probable **category** (e.g., classifying email as spam or legitimate). (Qualitative results).
- **Unsupervised Learning Types:**
 - **Clustering Methods:** Grouping data points that share similar characteristics (e.g., K-means clustering).
 - **Association Rules:** Uncovering how items within large datasets are associated.

- **Dimensionality Reduction:** Transforming high-dimensional data into a lower-dimensional space while retaining essential features (e.g., Principal Component Analysis - PCA).
- **Neural Networks (NNs) and Deep Learning (DL)**
 - **Neural Networks (NNs):** Complex algorithms designed to mimic the human brain, identifying patterns and generalizing across tasks.
 - **Layers of an NN:**
 - **Input Layer:** Receives the data.
 - **Hidden Layer(s):** Where computation and learning occur (neurons process signals).
 - **Output Layer:** Produces the final prediction or result.
 - **Deep Learning (DL):** A neural network with **more than three layers** (including inputs and outputs) is considered a DL algorithm. DL's key advantage is its ability to perform **feature extraction and transformation autonomously**.

NN Type	Description	Examples
Feedforward NN (FNN)	Information flows in one direction (input to output), no loops.	Email filtering, Image recognition.
Recurrent NN (RNN)	Designed for sequential data (time-series), with loops allowing information to pass from one step to the next.	Language translation, Speech recognition, Text generation.
Convolutional NN (CNN)	Specialized for processing grid-like data (images); uses convolutional layers to detect patterns like edges and textures.	Photo tagging, Object detection, Medical imaging.

- **Algorithms and Hyperparameters**
 - **Algorithm:** A set of step-by-step instructions or rules to solve a specific problem or perform a task. They are the **building blocks** enabling models to analyze data, identify patterns, and make predictions.
 - **Hyperparameters:** Settings that can be **adjusted prior to training** the model to optimize its performance (e.g., learning rate, depth of decision trees). Selecting the right hyperparameters is as important as choosing the right algorithm.
- **AI Security Architecture and Design Considerations:** - The AI security architecture depends on the organization's role: AI provider, AI producer, or AI customer.
- **AI Secure by Design (SBD):** - This process ensures security concepts are **enmeshed in development from inception through end of life**. The framework does not change due to AI, but it mandates that AI security is considered throughout the life cycle.
 - **Three Principles of Secure by Design:**

- Take ownership of customer security outcomes (e.g., proper consent, securing data pipelines).
 - Embrace radical **transparency and accountability** (e.g., explainability for AI systems).
 - Build organizational structures and leadership to achieve security goals (e.g., proper AI governance).
- **AI and DevSecOps:** - AI can be infused across the DevSecOps continuum (Plan, Code, Build, Test, Release, Deploy, Monitor, Operate) to enhance security and development activities.
- **Data Dependency and Storage**
 - **Sensitivity:** AI systems are highly sensitive to changes in training data, data sources, and preprocessing.
 - **Documentation:** The organization must document data input requirements and align them with production data.
 - **Storage:** AI models create an expansive amount of data. The IT security professional must identify **how and where** AI input/output data will be stored to mitigate the risk of **data leakage, shadow IT, drift, and access control issues**.
- **Configuration Management and Model Selection**
 - **Consistency:** Maintaining consistency between **data preprocessing** during training and **data preprocessing** during postproduction (inference) is essential for reliable performance. The same **tokenization framework** should be used.
 - **Model Behavior:** AI model behavior can be modified or tuned based on **post-training parameters** (e.g., *temperature* in GPT-4 can dramatically change output variance).
 - **Model Opacity:** AI models are often opaque, varying significantly based on data sources, training methods, and ethical practices of developers.
- **AI Change Management:** - Changes to AI security architecture require a specific change management plan.
 - **Key Elements:** Research emerging technologies, technical requirements for integration, **shared responsibility** (especially with third parties), **provenance of the model**, data and model governance, business continuity planning, access control, regulatory/societal impact monitoring, scalability, and workforce skills/competencies.
 - **Emergency Changes:** Urgent fixes are difficult, as re-training an AI model can take hours to weeks. Alternatives include:
 - **Rollback to an earlier AI model.**
 - **Implement input/output validation** (data preprocessing and postprocessing) *outside* of the model itself.

2. AI Life Cycle (OECD Model): - The AI life cycle is similar to the System Development Life Cycle (SDLC), requiring continuous risk consideration.

- **Plan and Design**
 - **Objective:** Define the system's concept, objectives, assumptions, context, and requirements. Define clear goals and **success metrics**.
 - **Documentation:** Problem and solution documented in a **business use case**.
 - **Risk:** Conduct an initial risk assessment, identifying potential **ethical considerations and societal impacts** (including bias).
 - **Process:** Establish a robust **data governance** framework and use **prototyping** to validate concepts and assumptions early.
- **Collect and Process Data**
 - **Objective:** Ensure high-quality data, as it directly impacts performance and reliability.
 - **Process:** Gather data, ensure proper **consent/licensing**, and **clean** the data (removing inconsistencies, duplicates, errors). Perform checks for **completeness, quality, relevance, and timeliness**.
 - **Documentation:** Critical step is documenting **data provenance** (creation, sources, preprocessing, intended uses, maintenance).
- **Build and/or Adapt Model(s)**
 - **Objective:** Transform collected data into actionable insights by creating or selecting models/algorithms and training/calibrating them.
 - **Risk:** Focus on mitigating the risk of lacking **explainability** and **robustness**.
 - **Process:** Select the appropriate model, **calibrate** (tune parameters), and ensure **explainability** to build stakeholder trust. Integrate **Human-in-the-loop (HITL)** oversight for optimal accuracy and security.
- **Test, Evaluate, Verify, and Validate (TEVV)**
 - **Objective:** Ensure the AI system is accurate, effective, **trustworthy, and secure**.
 - **Verification:** Model is built **correctly** (according to design specifications).
 - **Validation:** Model meets the **intended purpose** and provides accurate output.
 - **Testing Types:** Model testing (accuracy, precision), **Stress testing**, Comparative analysis, **Bias and fairness checks**, Scenario analysis.
 - **AI Red Teaming:** Essential practice of emulating real-world adversaries to identify risk, uncover blind spots, and validate assumptions.
 - **Base Model Red Teaming:** Identifies misuse, capabilities, and limitations of the base model.
 - **Application-Level Red Teaming:** Includes base model testing but reviews application-specific safety and failures beyond base model safety features.
 - **Assurance:** TEVV gives reasonable assurance of the validity of outcomes and adherence to the organization's **risk appetite**.
- **Make Available for Use/Deploy**

- **Objective:** Implement the AI solution per the established plan.
 - **Process:** Conduct **piloting** in a controlled environment, check **compatibility** and **interoperability** with legacy/existing systems, manage organizational change (training, new roles), and evaluate **user experience**.
 - **Third Parties:** Provide clear instructions and ensure compatibility/integration with downstream deployers.
- **Operate and Monitor**
 - **Objective:** Continuous oversight to ensure optimal performance and ethical integrity.
 - **Process:** Day-to-day operations, systematic evaluation of performance, documenting quality metrics, and assessing both **intended and unintended consequences**. Establish a **quality management system** to ensure compliance with regulatory and industry standards.
- **Retire/Decommission**
 - **Objective:** Conclude the life cycle when the solution is obsolete, new alternatives are available, or overall risk exceeds management's **risk appetite**.
 - **Process:** Follow established **change management procedures** for secure data and software component removal, archiving, and system dismantling.
- ### 3. Data Management Controls
- Data is central to AI, requiring information security teams to understand and secure data throughout the entire **data life cycle** (acquisition, preprocessing, utilization, storage, retention, destruction). Risk and security must be evaluated at each stage.
- **Data Governance**
 - **Alignment:** AI data management must align with the enterprise's overall **data governance program** and baseline governance principles.
 - **Monitoring:** Governance monitors how the phases of the data life cycle pertain to AI.
 - **Security:** AI security enhancements should build upon existing organizational security practices.
 - **Data Acquisition:** - Data acquisition involves collection or receipt (e.g., through agreements or purchase). All data should have **data processing agreements** and **standard operating procedures** detailing its **origin** and **legal basis** for use.
 - **AI Data Acquisition Questions (Security Team):**
 - What is the **origin** of the data? Is the initial source detailed in agreements?
 - If non-personal, are the responsibilities, parameters, and uses outlined (proprietary/confidential definitions)?
 - Is the data **personal data**?

- Are the **necessary consents** in place and do they align with the intended processing activity? (e.g., routine process consent, uncommunicated process consent).
- Are consents **up to date** (consistent with policies and jurisdiction laws)?
- **Data Organization and Preparation:** - The AI development team and data governance team must work closely to keep data organized and clear. AI processes must meet the same **security review and mapping** standards as any other data processing.
 - **Security Controls for AI Data Management (Same as any dataset):**
 - **Dataflow mapping:** Clear map of all data elements, source, and hardware location (on-premises, cloud, hybrid) throughout the AI process. Must align with incident response evidence collection.
 - **Access control review:** Review rights at different acquisition stages. Implement **principle of least privilege** for training/testing datasets and the model.
 - **Data integrity monitoring:** Monitor processes like **data minimization, cleansing, and standardization**. Data steward provides manipulation details.
 - **Additional Controls for AI Datasets:**
 - Review **data sharing agreements** and mapping specific to the AI model.
 - **Location of temporarily housed data** must be known and securely deleted *immediately* after processing.
 - Follow **risk management processes** when importing external data.
 - Implement **security monitoring specific to AI processes** (added layer).
 - Decisions should involve an **informed and diverse team**.
 - **Possession:** The organization is responsible for meeting privacy and security requirements, regardless of the length of possession (nanosecond or decades). CIA triad is always relevant. Security risk assessments must be conducted **prior to, during, and after** AI processing.
- **Data Utilization and Storage:** - Utilization becomes specialized (e.g., LLM development/deployment).
 - **Security Considerations (Examples):**
 - **Privacy-by-design:** Monitor adherence with greater frequency when using automated decision-making (generating new data, subject to **hallucinations** and security risk).
 - **New Data Generation:** Follow data acquisition processes (3.4.1).
 - **Dynamic information flow controls:** Policies guiding permissible information flows based on threat environment or risk tolerance.
 - **Data Storage Security:** Use enterprise-wide controls; AI-generated data must align with its **classification**.
 - **General:** Security should align with all other organizational data processes. Pay particular attention to **encryption in all states** and consistency of use.
- **Data Retention and Destruction**

- **Retention Schedules:** Must cover both **input data** and **data generated** by the AI process.
 - Stored data must have clear reasons for storage described in the schedule.
 - Audit capabilities should review access frequency for non-operational data.
 - Access to AI-related storage should **not be automatically granted** to users with access to similar data for other functions.
- **Destruction:** Must be consistent with agreed-upon responsibilities (contracts, laws). **Security-by-design** principles outline destruction parameters. The retention schedule for AI-generated data may benefit from **automated destruction**.
- **Special Consideration (DL/NNs):** Creation and deletion of data may not be transparent, affecting integrity and creating vulnerabilities. Requires **continual monitoring** at higher levels than typical.

- **Summary of Data Governance Controls**

Control	Description
AI data retention and encryption protocols	Deploy encryption and implement clear retention timelines.
Centralized AI database management	Establish a centralized database for high-risk AI system details, conduct compliance checks, and validate accuracy.
Data privacy impact assessment	Use tools (encryption, anonymization, access controls), train personnel, and use penetration testing to validate compliance.
Data storage and archiving	Implement automated data backup solutions; regularly test recovery. Securely store training data using encryption , secure transfer protocols, and firewalls.

- **Data Security:** - Implementing AI requires additional data security considerations throughout the AI life cycle, often balancing the need for data access with security restrictions (Principle of Least Privilege).
- **Data Encoding**
 - **Tokenization:** Raw data is often **tokenized** (e.g., BPE, Word Piece) and stored as binary files (e.g., TF Records) or vectors (embeddings).
 - **Vector Databases:** Store vector embeddings, making comparison/searching easier. Must be secured with access controls, encryption, and specific techniques (e.g., vector index management).
 - **Intermediate States:** Model weights and intermediate states are stored (in-memory, GPU, or checkpoint files like .ckpt, .pth) and require access/encryption controls.
- **Data Access:** - Aggregating data from various sources into a centralized location (e.g., **data lake**) introduces risk:
 - **Concentration:** Centralized data becomes a prime target for attackers (easier to exfiltrate).

- **Control Translation:** Access controls might not be properly translated during data transfer to the target source.
 - **Least Privilege Expansion:** Data exploration by data scientists may broaden the definition of least privilege necessary for the original collection purpose.
 - **Replication:** Data is replicated, increasing the number of secured systems and the scope of compliance programs.
 - **Action:** Review AI system dataflow; apply additional access controls for congruency; review **networking access policies** for containment.
- **Data Confidentiality/Secrecy**
 - **Challenge:** AI model training often requires data to be **machine readable**, meaning sensitive data might be **decrypted** during the AI development life cycle, potentially diluting confidentiality controls (like obfuscation/encryption).
 - **Future Solution:** New techniques like homomorphic **encryption** may allow training on encrypted data, but currently, high computational costs limit use.
 - **Current Controls:** Use **compensating controls** like:
 - Limiting access to unencrypted production data.
 - Monitoring the use of sensitive data.
 - Applying **disk-level encryption** (defense in depth).
 - **Data Backup:** - While source data may be redundant, new AI artifacts must be backed up:
 - **Artifacts to Backup:** Postprocessed data, **tokenized training data**, **model weights**, and model architecture parameters/definitions (JSON/YAML).
 - **Explainability:** GenAI models are **nondeterministic**. To provide better explainability, organizations should preserve copies of **training and testing datasets**, as well as model performance and bias testing results.
 - **Data Integrity:** - Preserving the integrity of datasets, model weights, and architecture is critical.
 - **Compromise Scenarios:**
 - **Data poisoning:** Attacker modifies training data to influence model behavior.
 - **Model tampering:** Modifying weights, architecture, or parameters for incorrect/biased/malicious results.
 - **Embedding tampering:** Tampering with embedding matrices to skew model results.
 - **ETL:** Errors can be introduced during preprocessing (Extract, Transform, Load). Requires defining, documenting, testing, and properly implementing ETL logic.
 - **Additional Controls:**
 - **Anomaly detection** in training data.
 - Separating **training data from production data**.
 - Using **data validation** and **model ensembles**.
 - Applying **model hardening**.

4. Privacy, Ethical, Trust, and Safety Controls

These four groups are interwoven, with a common focus on **harm**. Assessing potential harm is complex.

- **Privacy:** - Privacy protection is crucial in AI processes because it's easy to trace data points to a single individual.
- **Legal Basis and Consent:** A legal basis for processing is required, typically **consent**. Consent must be specific (e.g., acceptance of terms for one service like consent for marketing).
- **Consumer Rights (PII):** AI processes must maintain these rights:
 - **Right to be informed:** How data is collected, used, shared, and retained.
 - **Right to access:** Gain access to data held about them (e.g., a copy).
 - **Right to rectification:** Request correction of incorrect information.
 - **Right to be forgotten:** Revoke consent; data must be destroyed.
 - **Right to restrict processing:** Request data not be processed in a certain way, while still allowing other uses.
- **Compliance:** If an AI use case (e.g., travel recommendations) is not covered by initial consent, **additional consent is required**.
- **Third-Party Models:** Using a third-party AI solution (e.g., for custom messages) requires a **Data Processing Agreement (DPA)** on how personal data will be managed, in addition to a Data Protection Impact Assessment (DPIA).
- **Development Environment:** Need protections (change management, security controls) to prevent **data leakage** when AI systems are not closed-loop.
- **Controls for AI-related Data Privacy**

Control	Description
Data privacy and handling protocols	Implement/document protocols for encryption, anonymization, and access controls.
AI data retention and encryption protocols	Deploy encryption and implement clear retention timelines.
Privacy-first data handling	Models handle data with privacy as a priority.
Differential privacy	Ensure the model does not leak individual training data information.
Privacy-enhancing technologies	Integrate technologies to meet or exceed privacy standards.

- **Ethics:** - AI introduces ethical concerns due to **automatic decision making** and **self-learning**, which can create undue harm or risk without human involvement.
 - **EU AI Act (Examples of Unacceptable Uses):** Social credit scoring, emotion-recognition in work/education, exploitation of vulnerabilities, behavioral manipulation, untargeted facial image scraping, certain biometric categorization, and specific predictive policing.
 - **Strategy for Ethical Considerations:**
 - Consider intended purpose, enterprise context, applicable laws, and societal norms.

- Ensure **collaboration and diversity** in development (demographic, departmental, and skill diversity).
 - Gather **benchmarking data**.
 - Develop **feedback mechanisms** among users to address negative impacts.
- **Common Ethics Controls:**
 - Ensure **Human Oversight (HITL)**.
 - Establish an **AI ethics and compliance committee**.
 - Establish and evaluate an **AI code of conduct**.
 - Document adherence to standards (health, safety, laws, fundamental rights).
 - Ensure **transparency, fairness, and lack of bias**.
 - Evaluate **environmental impact**.
- **Trust:** - A key concern is the **lack of clarity** into how AI makes decisions (especially DL and NNs).
 - **Transparency:** Document the data used, data created, and the AI process design.
 - **Inscrutability/Black Box Problem:** When an AI process is opaque and results cannot be explained.
 - The **governance process** must incorporate risk management to identify potential black box use cases and prepare for remediation.
 - **Documentation** of what is **not known** about how the model fails is as important as what is known.
 - **HITL Standard:** Derived from the nonexperimental human-controlled traditional software development life cycle. AI is experimental, and its ability to adapt can lead to **unexpected, unexplainable outcomes**, which changes the **risk acceptance paradigm**.
- **Safety:** - Concerns range from **physical safety** (e.g., autonomous robots) to the **agency** of AI to act independently.
 - **Agency and Accountability:** Practical issues arise when AI can act autonomously:
 - How can AI or its human creators be held **accountable** for the consequences of its decisions?
 - What self-regulation does AI have for ethical/safe behavior?
 - What happens if an AI decision conflicts with user values?
 - **Nefarious Use:** Exploitation of collected data by fraudsters (e.g., creating false identities), and **Deepfakes** (computer-generated counterfeit video/audio, often created using **GANs**).
 - **Control:** **Robust governance** and **increased supervision** over AI outcomes are key.
- **Controls Related to AI Supervision**

Control	Description
---------	-------------

Logging and monitoring	Detailed logging of decision-making paths ("chain of thoughts") for complex DL models. Creates an audit log for debugging.
AI observability	Practices/tools to ensure availability, reliability, performance, and trustworthiness . Categories include:
	Data pipelines: Monitoring for prompt injection, data quality issues.
	Monitoring infrastructure and the system: System health, consumption, and processing telemetry.
	Interpretability: Strategies for complex models (e.g., counterfactual explanations, feature virtualization).
Human in the loop (HITL)	A person oversees and monitors AI, making final decisions for critical actions. Requires the workflow to be designed to halt for human approval.

- **Human-in-the-loop (HITL)**
 - **Function:** For critical decisions, a human reviews and approves the AI model's decision (e.g., a doctor reviewing an AI diagnosis).
 - **Implementation:** The AI workflow pipeline must be designed to require HITL for specified conditions and explicit approval. This will slow the AI system's speed.
 - **AI in the loop (AITL):** A paradigm shift where AI actively assists and augments human decision-making by providing insights/recommendations, but **humans maintain primary control** over critical decisions.

5. AI Security Controls and Monitoring

- **AI Security Controls (General)**
 - **Complexity:** AI security controls are more complex than traditional IT security (firewalls, antivirus) due to how AI is applied (internal business solutions, product enhancement, models for sale).
 - **Guardrails:** Needed to protect the organization from **inside and outside threats**, based on risk assessments or change management outcomes.
 - **Implementation:** Organizations must consider controls specific to using or developing AI, which may involve **augmenting existing controls** or implementing **new ones**.
 - **Foundation:** Understanding the **type of AI** (as detailed in 3.1) is essential for applying the correct controls.
 - **Change Management:** The **AI change management process** is critical to ensure changes and impacts to the AI security architecture are considered.

• AI Security Controls by Category

Control Category	Controls	Explanation
	Data Quality	Document definitions for: thresholds and rules , assessing/resolving quality issues, monitoring , and

Technical, Security, and Privacy		communication/transport protocols.
	Privacy	Encryption, hashing, and de-identification ; data anonymization; key management ; monitoring/logging; Identity and Access Management (IAM) ; regular Privacy Impact Assessments (PIAs) ; reporting.
	Data Management	Access controls (privileged/administrator); encryption ; data quality management (aggregation/integration); documentation and modeling; data integrity controls .
	AI Incident Management	Focuses on data breach and operational interference . Includes attacks on the model (e.g., model inversion , model poisoning) or AI misuse . Must be integrated into the organization's existing incident response plan.
Operation and Life Cycle Management	Dataflow Diagrams (DFDs) & Process Flow Documentation	Aids stakeholder understanding of data handling (logical, physical, context flows) and complex processes (collection, processing, enrichment).
	AI Decision-Making Documentation	Explainability documentation (how the process works); decision-making process (steps taken, includes DFDs/process flows); end-user documentation ; model documentation (architecture, algorithms, training data, performance); compliance and ethical documentation (bias/fairness, regulatory); data management and traceability (sourcing, consent, transformations, logging).
	Applicable Standards	Change management, data quality, and validation.
	Version Control and Management	Ensures reproducibility ; includes testing, training,

		environment management, virtualization/containerization, automated tools, and deployment processes.
	Change Management	Documented process, formal record keeping, retention of approvals, risk and potential impact identification , system of record, communication, continuous auditing.
	Continuous Monitoring and Logging	Monitoring of critical systems for AI product delivery and processing; includes data quality management controls (input validation, consistency, model drift).
System Development Life Cycle (SDLC)	Project Management	SDLC controls depend on approach (e.g., DevOps/DevSecOps vs. Waterfall). AI projects often fail due to unclear requirements/poor leadership.
	Ideation & Design Phases	Technical design documentation. Data collection (acquired/purchased); Data quality; Data optimization; Data fitness (completeness, bias mitigation); Requirements defined (including model types).
	Access Controls	Evaluation of existing controls against new regulations. Access adjusted based on data classification and AI system classification . Use Multifactor Authentication (MFA) . Regular monitoring, certification, and recertification of user access.
	Deployment Management	Automation may be used; must be monitored by operations and audited.
	Ethical Impact Assessment	Conduct for AI systems; align with available industry/corporate ethical

Ethical and Human Values		frameworks (e.g., NIST, Microsoft, IBM, Google).
	Consent Management	Ensure stakeholders have the right to opt out of AI processing/decision-making; validate processes for documenting/processing consent for training data; review data retention and "right to be forgotten" procedures.
	Bias Management	Validate processes to evaluate training data for bias and ensure data management controls are operating effectively.
	Transparency and Explainability	Disclosure of AI decision-making ; validate documentation of capabilities/limitations and training data collection; confirm use of explainability tools like: LIME (Local Interpretable Model-Agnostic Explanations for individual predictions) and SHAP (SHapley Additive exPlanations for feature importance).
	Human Oversight	Integrate an AI steering committee ; ensure processes for human oversight of AI decisions/output; validate an appeal process for AI decisions.
	Configuration Management	Monitoring provides risk detection and mitigation in AI development (e.g., correcting vulnerabilities before exploitation).
	Testing Frameworks and Protocols	End-to-end testing, scenario analysis, bias and fairness testing.

- **AI for Security and Monitoring Tools**
 - **Advancement:** AI models can be used as tools to **support security**, helping defenses evolve quickly to proactively detect and address threats.

- **Automation:** AI automates repetitive tasks (monitoring, scanning) and assists in analyzing data-heavy logs, reducing human error.
- **Leveraged Tools:** Existing tools now leverage AI capabilities:
 - **SIEM** (Security Information and Event Management)
 - **SOAR** (Security Orchestration, Automation, and Response)
 - **XDR** (Extended Detection and Response)
 - **UEBA** (User and Entity Behavior Analytics) right arrow Employs AI to detect **anomalous user behavior**.
- **AI Security Control Frameworks:** - The adoption of AI requires reexamining existing security controls and implementing new ones. Organizations can leverage guidance from evolving frameworks:
 - **NIST AI Risk Management Framework**
 - **Cloud Security Alliance AI Controls Matrix**
 - **MITRE ATLAS Matrix**
 - **OWASP GenAI Security Project Top 10 for LLMs and GenAI**
 - **ISACA Artificial Intelligence Audit Toolkit**
 - **IBM GenAI Controls Framework**
 - **Google's Secure AI Framework**
- **Specific AI Security Controls**
 - Access Controls
 - Protecting access to **data used for and created by AI** requires additional consideration.
 - An AI access control policy should cover authentication, authorization, and type of data.
 - **Key Considerations:**
 - ◆ **Centralized** AI access control policy.
 - ◆ **Role-Based Access Controls (RBACs).**
 - ◆ Validation to test access to **mixed data models** (where restricted data may be available within the model).
 - ◆ Review access controls after each development release.
 - **Zero Trust in AI:** - The concept of "never assume security unless verified" applies strongly to the AI model and its data.
 - **Traditional Zero Trust Applied to AI:**
 - ◆ **IAM:** Apply **MFA** and **principle of least privilege**.
 - ◆ **Network Segmentation:** Secure connections between data sources and the AI model.
 - ◆ **Data Encryption:** Safeguard data throughout its life cycle.
 - ◆ **Data Loss Prevention (DLP):** Monitor and prevent unauthorized processing/transfer; ensures sensitive data isn't sent to **public AI models**.
 - ◆ **UEBA:** Detects unusual behavior pointing to potential attacks.

- ♦ Continuous monitoring and auditing, incident response, security analytics, and threat intelligence.
- **Zero Trust for AI Decisions:** Trusting AI decisions requires considering three key areas:
 - ♦ **Ability:** Can the AI perform the task **robustly, safely, and reliably**?
 - ♦ **Integrity:** Will the AI process data properly, or could information be manipulated?
 - ♦ **Benevolence:** Does the AI adhere to "**do no harm**" principles?
- **Controls for Internet Access (Especially LLMs):**
 - ♦ **Deny access** where possible.
 - ♦ Implement **access controls** following organizational policies.
 - ♦ **Block restricted websites** (URL/spam filtering).
 - ♦ **Control access to available data** based on approved use cases.
 - ♦ Ensure **data validation** (accuracy, non-harmful).
- **AI Acceptable Use Policy (AUP)**
 - A **required security control** that states the framework for ethical and responsible deployment/use of AI.
 - Provides clear guidance on usage, balancing benefits against risk.
 - Communicates required and prohibited activities and behaviors to staff.
- **AI Audits and Traceability**
 - **Audits:** Critical to ensure requirements and policies are followed; can be augmented by AI tools.
 - **Traceability:** The ability to track how **data moves through the system** and **how decisions are made**.
 - Verifies origin of data, processes, and other factors key to model development.
 - Helps avoid the "**black box**" nature of many models.
 - **Metadata Logging:** Provides information about AI models, inputs, and outputs to ensure trustworthiness.
 - **Model Cards** are one method used to record this.
- **Security Controls for the Supply Chain**
 - Reliance on vendors requires expanding typical due diligence.
 - **AI-Specific Inquiries:** Ask vendors about AI technology integration, data privacy, data retention, **AI model validation**, and maintenance.
 - Inquire about the vendor's reliance on **sub-vendors** for data or models.
- **Shadow AI**
 - Use of **unapproved AI applications** (like unapproved Shadow IT).
 - Introduces significant risks: accidental data breaches, compliance violations, reputational damage.
 - Guardrails are needed to identify and protect against proprietary data leaking into **public domain models**.
 - **Controls to Identify/Protect:** Shadow IT policy, IT as a service-delivery org, IT budgeting/procurement, IT system consolidation, **User Education**, **CASBs** (Cloud Access Security Brokers), **DLP**, Network analysis, **UEBA**, Dataflow monitoring.

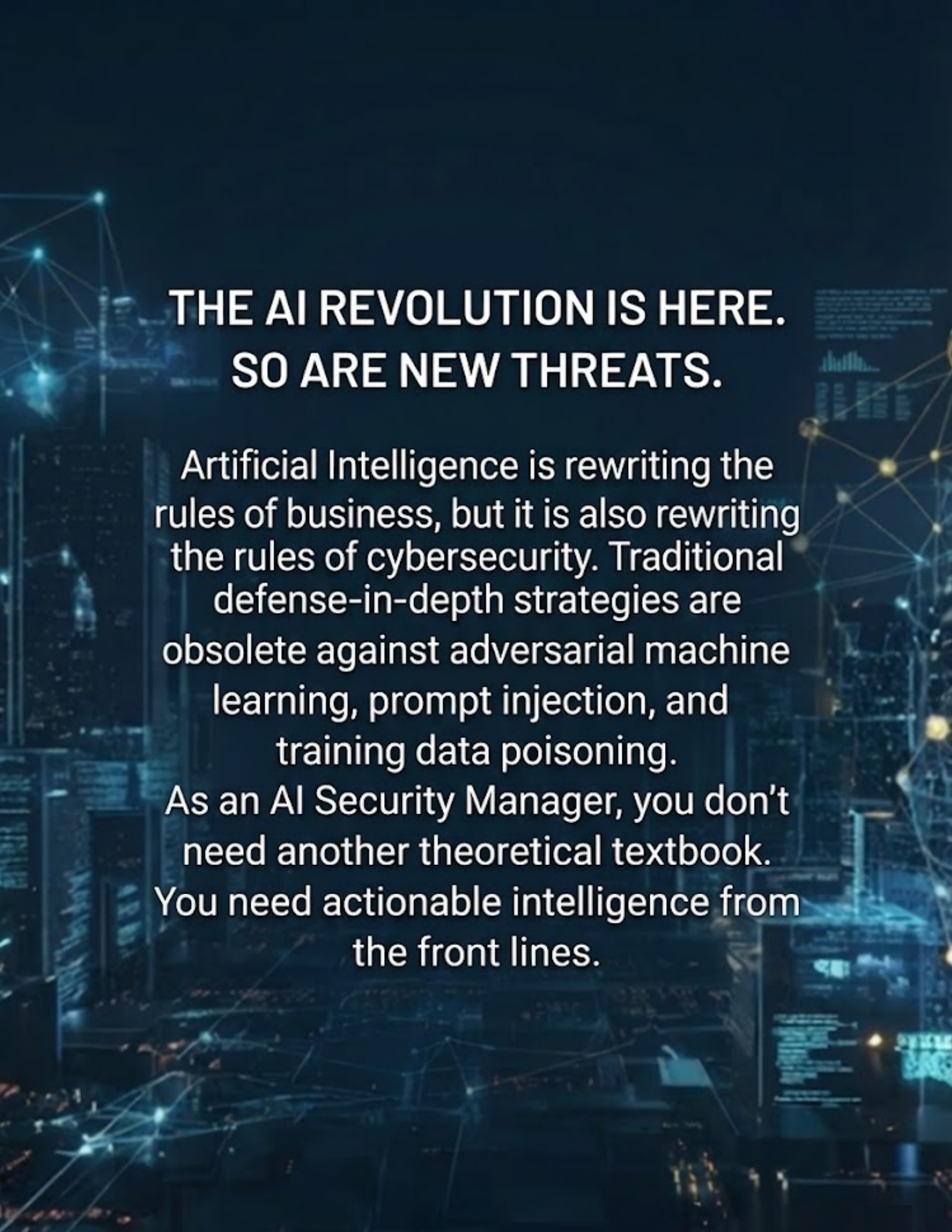
- **AI Incident Management**
 - Existing incident management programs must be updated to include the definition and process to identify and remediate incidents on **AI systems**.
 - Additions help address risks like **data breaches, algorithmic bias, safety issues, operational failures, and ethical violations**.
- **AI Security Awareness Training**
 - Prioritizing the safe use of AI requires **full engagement of employees**.
 - Should be more than a once-a-year presentation; leadership must be clear about intent and policy alignment.
 - **Components of an Effective Program:**
 - Security, privacy, and ethics.
 - Intended benefits to organization and users.
 - Understanding and training to interpret **AI model outputs**.
 - **Tabletop exercises.**
 - Ability to detect and manage **bias** in data and outputs.
 - Copyright and **IP concerns**.
 - Understanding **AI explainability** and **black-box risk**.
 - Identification of **model drift** or bias in outputs.
 - Awareness of **adversarial AI** and model manipulation techniques.
- **Addressing AI Skills Gaps**
 - Critical to identify necessary skills for design, development, deployment, assessment, and monitoring.
 - The organization must recruit, develop, and retain a workforce with the necessary backgrounds.
 - Investment in **training current staff** is essential for ethical and responsible actions.
 - Skill gap requirements should be included in the **AI governance program**.
- **Continuous Monitoring**
 - Modern security monitoring tools provide assurance by tracking trends and stopping incoming threats.
 - **AI-Specific Monitoring Considerations:**
 - AI models may produce outputs that are **factually incorrect** (e.g., hallucinations).
 - Need to detect threats like **data poisoning** (manipulated training data).
 - Anomaly detection tools (e.g., pattern recognition algorithms) identify discrepancies in inputs/outputs.
 - **Monitoring Examples for Model Integrity:**
 - **Drift**
 - Model behavior monitoring
 - Threat intelligence (including third-party integrity)
- **Drift**
 - **Core component of model life cycle management.**
 - **Definition:** Degradation of ML model performance due to changes in data or in the relationships between input and output variables.

- **Impact:** Faulty decision-making and bad predictions.
- **Detection/Response:**
 - Regular monitoring of performance and trustworthiness.
 - Utilizing **Human-in-the-Loop (HITL)** as part of monitoring outputs.
 - AI drift detection tools automatically detect when accuracy drops below a **preset threshold**.
 - Set **quantitative thresholds** (e.g., accuracy below 85%) that trigger **retraining or review protocols**.
- **Threat Intelligence**
 - Expands monitoring to include AI third-party tools, which increase the attack surface.
 - Provides assurance by monitoring for biased or malicious use of AI algorithms.
- **Controls for AI-Enabled Security Threats**

Threat	Control(s)
Social Engineering/Deepfakes	Regular employee training on emerging deepfake threats; AI-powered cybersecurity platforms for anomaly/NLP screening; phishing simulations.
Adversarial Models	Leverage adversarial training , randomized smoothing, and formal verification; monitor model performance metrics (precision, recall, F1); sanitize training data .
Credential Stuffing	Strong passwords and MFA ; network monitoring; password-less authentication; AI-driven adaptive/risk-based authentication.

- **Tracking Security Metrics:** -Tracking metrics is essential to measure the effectiveness of AI utilization and safeguard AI models.

Metric	Description
Level of Preparedness	Device/software updates, high-risk vulnerability identification.
Shadow IT	Device count.
Intrusion Attempts	Breach counts, frequency analysis, source identification.
DLP Effectiveness	Incident prevention rate, response time, false positives/negatives.
Awareness Training	Effectiveness/engagement, knowledge improvement, behavior change, regular updates.
AI Solution Performance	False positive rate, model drift frequency , latency in AI response time.



THE AI REVOLUTION IS HERE. SO ARE NEW THREATS.

Artificial Intelligence is rewriting the rules of business, but it is also rewriting the rules of cybersecurity. Traditional defense-in-depth strategies are obsolete against adversarial machine learning, prompt injection, and training data poisoning.

As an AI Security Manager, you don't need another theoretical textbook. You need actionable intelligence from the front lines.