Responsible AI

AI Safety & Security

Components of AI Safety & Security

- Robustness
- Reliability
- Security controls
- Data protection
- Threat assessment
- Incident response
- Regular security audits

Discuss/ask these key questions with your team during the implementation of AI solutions:

Data Integrity and Protection

- Are all datasets for our AI systems collected from verified, trusted sources?
- Do we have clear data policies addressing privacy, protection, and classification of sensitive data?
- How do we ensure secure storage and processing of datasets, including the use of encryption and access controls?
- Are datasets tracked and verified (for example, via cryptographic hashes) before use?

System Robustness and Reliability

- How do we test the AI system for resilience against failures, attacks, or adversarial inputs before deployment?
- Are model training and deployment conducted and validated under real-world conditions?
- How is continuous monitoring implemented to quickly detect unexpected behavior or errors?

Security Controls and Access

- What technical (e.g., encryption, firewalls), administrative (e.g., access policies), and physical controls are in place to protect our AI systems?

- Are access requests reviewed regularly, and do we apply the principle of least privilege for system access?

- Do we have effective separation between development, test, and production environments

Threat Assessment

- What processes are followed for identifying and prioritizing potential threats and vulnerabilities in our AI systems?
- How often do we conduct security reviews, penetration testing, or vulnerability assessments?

Incident Response and Recovery

- Is there a documented incident response plan specifically for AI-related security breaches?
- Do we have routine drills or testing for incident detection, response, and recovery to ensure rapid containment?
- Are business continuity and recovery plans in place for critical AI assets?

Transparency and Human Oversight

- How transparent are our AI models and decision-making processes to stakeholders? Are outputs explainable and auditable?
- What mechanisms exist for human oversight and intervention when the AI behaves unexpectedly or incorrectly?

Compliance and Governance

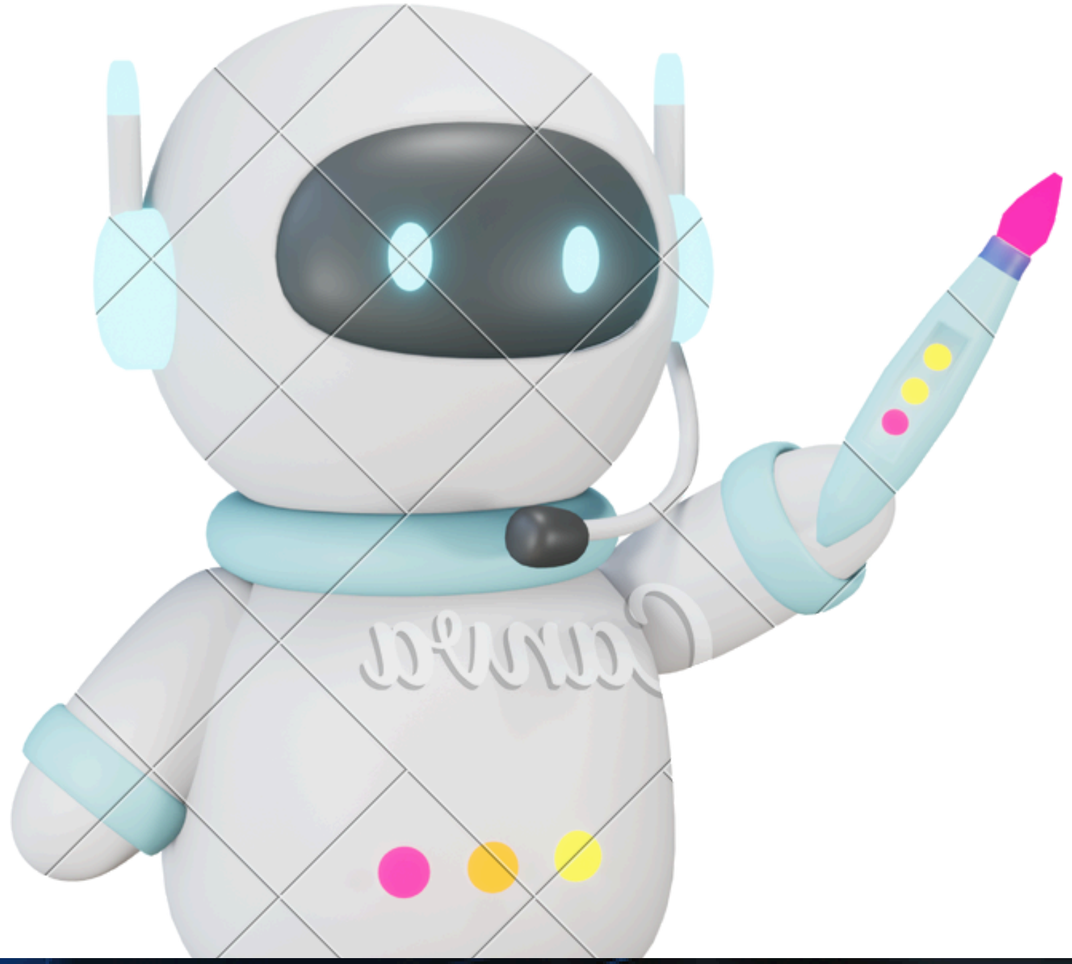
- How do we ensure ongoing compliance with relevant regulations (GDPR, ISO/IEC 27001, etc.)?
- Do we have clearly defined roles and responsibilities for AI safety and security within the team?

References

- https://abnormal.ai/blog/questions-for-security-vendors-ai
- https://mitratech.com/resource-hub/blog/ai-security-policy-questions-to-ask-third-party-vendors/

Further Readings

- https://www.federalregister.gov/documents/2025/01/06/2024-30983/hipaa-security-rule-to-strengthen-the-cybersecurity-of-electronic-protected-health-information
- https://www.isaca.org/resources/glossary
- https://www.secondsight-ts.com/threat-assessment-blog/threat-and-risk-assessment-approaches-for-security
- https://www.cycognito.com/learn/exposure-management/security-controls.php



Thank You

Build Powerful AI.
Build with Ethics.
Build with Governance.
Build with Safety & Security
Build Together.