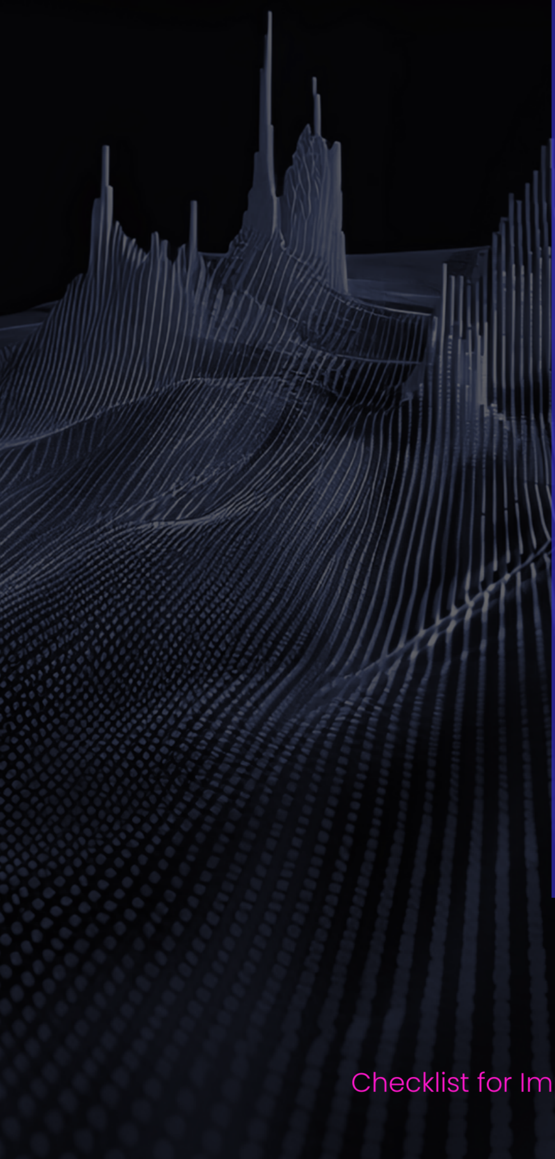
hackerone

Checklist for Implementing Safe and Secure AI

Whether your organization is looking to develop, secure, or deploy AI or LLM, or you're hoping to ensure the security and ethical adherence of your existing model, we've compiled a checklist for implementing safe and secure AI. While not exhaustive for every use case, this checklist can get you started. Ask the experts at HackerOne for more details on safeguarding your AI.



Joint AI safety & security measures:

- □ **Red teaming:** Incorporate both security and safety AI red teaming as a standard practice for AI models and applications.

- □ **Testing:** Establish continuous testing, evaluation, verification, and validation throughout the AI model life cycle. Provide regular executive metrics and updates on AI model functionality, security, reliability, and robustness. Regularly scan and update the underlying infrastructure and software for vulnerabilities.

- □ **Risk assessment:** Conduct comprehensive risk assessments to identify potential risks associated with the AI system, including unintended consequences, negative societal impacts, and misuse or abuse scenarios.

- □ **Regulations and governance:** Determine country, state, or government-specific AI compliance requirements. Some regulations exist around specific AI features, such as facial recognition and employment-related systems. Establish an AI governance framework outlining roles, responsibilities, and ethical considerations, including incident response planning and risk management.

- □ **Input and output security:** Evaluate input validation methods, as well as how outputs are filtered, sanitized, and approved.

- □ **Training:** Train all users on ethics, responsibility, legal issues, AI security risks, and best practices such as warranty, license, and copyright. Establish a culture of open and transparent communication on the organization's use of predictive or generative AI.

AI safety measures:

- □ **Ethical considerations:** Establish clear ethical principles and guidelines for the development and use of AI systems, addressing issues such as bias, transparency, accountability, and respect for human rights.
- □ **Human oversight:** Incorporate human oversight and control mechanisms into AI systems, allowing for human intervention and decision-making in critical situations.
- □ **Explainability and transparency:** Ensure that AI systems are explainable and transparent, enabling users and stakeholders to understand how decisions are made and the underlying reasoning.
- □ **Continuous monitoring:** Establish mechanisms for continuous monitoring of AI systems during operation, to detect and respond to any deviations from expected behavior or potential safety concerns.
- □ **Responsibility and accountability:** Clearly define roles, responsibilities, and accountability measures for the development, deployment, and use of AI systems, including processes for redress and remediation in case of harm or unintended consequences.
- □ **Stakeholder engagement:** Involve diverse stakeholders, including affected communities, experts, and regulators, in the development and deployment of AI systems to ensure a comprehensive understanding of potential impacts and concerns.

AI security measures:

- □ **Data security:** Verify how data is classified and protected based on sensitivity, including personal and proprietary business data. Determine how user permissions are managed and what safeguards are in place.
- □ **Access control:** Implement least-privilege access controls and defense-in-depth measures.
- □ **Training pipeline security:** Require rigorous control around training data governance, pipelines, models, and algorithms.
- □ **Monitoring and response:** Map workflows, monitoring, and responses to understand automation, logging, and auditing. Confirm audit records are secure.
- □ **Production release process:** Include application testing, source code review, vulnerability assessments, infrastructure security, and AI red teaming in the production release process.
- □ **Supply chain security:** Request third-party audits, penetration testing, and code reviews for third-party providers, both initially and on an ongoing basis.
- □ **Measurement:** Identify or expand metrics to benchmark generative cybersecurity AI against other approaches to measure expected productivity improvements. Stay updated with the latest advancements in AI security research and best practices.