ScaleSec

Putting it all Together: ScaleSec's Checklist for Securely Integrating Generative AI

It is paramount to backstop your security program and products to an industry-recognized framework, such as the NIST Cyber Security Framework (CSF). Below is a checklist of recommendations and goals to accomplish to ensure you can build and leverage generative AI systems securely and effectively, aligned with the August 2023 draft of NIST CSF 2.0. Along with our proposed implementations, we reference the NIST CSF guideline categories and subcategories so you can better understand how security integrates with your AI ecosystem.

Govern (GV)

- ☐ Clearly outline your goals and requirements for generative AI integration (GV.OC)
  - ☐ Consider compliance, regulatory, contractual requirements, and consumer privacy as part of your requirements gathering
- ☐ Identify priorities, risk tolerance, and assumptions to be used when designing your AI program (GV.RM)
- ☐ Establish processes to manage your AI platform and training data supply chain (GV.SC)
- ☐ Clearly define accountability for your AI program, including roles and responsibilities for all components (GV.RR)
- ☐ Establish and ingrain secure and ethical use of AI into your policies, standards, and procedures (GV.PO)
  - ☐ Ensure coverage of what types of data are acceptable as inputs, and guidelines on sharing content internally and externally
- ☐ Continually review and adjust your AI program, maintaining transparency and engagement to build public trust. (GV.OV)

Identify (ID)

- ☐ Maintain an inventory of your AI systems, and lineage of any training data utilized (ID.AM)
  - ☐ Know where your models and their training data is coming from
  - ☐ Verify the legitimacy and accuracy of training data sources
- ☐ Understand security risks relevant to your AI systems, maintain a threat matrix (ID.RA)
- ☐ Continually reassess your AI systems and address any findings (ID.IM)
  - ☐ Models and technologies continue to evolve rapidly, consider how you will measure the impact of changes
  - ☐ Consider required granularity of model tuning: general vs. specific for each use case
- ☐ Regularly review plugins and external services, expect them to evolve rapidly

Protect (PR)

- ❑ Control access by/to AI systems and data based on least-privilege access principles, utilize Role-based access control (RBAC) (PR.AA)
  - ❑ Build/train/deploy models and their respective services with trusted infrastructure patterns and verified IAM primitives to minimize administrative errors and reduce attack surface.
  - ❑ Enforce access controls across the chain: data, model, services, endpoints, etc.
  - ❑ Require Multi-Factor Authentication (MFA)
  - ❑ Establish human-in-the-loop for privileged operations, don't give AI systems agency beyond carefully controlled actions
  - ❑ Establish boundaries and treat LLMs as if they were untrusted users
  - ❑ Perform periodic IAM audits, including access reviews and cleanup
- ❑ Develop a training curriculum around AI usage and security, to ensure users treat inputs and outputs appropriately with healthy skepticism, and sensitive data is not leaked (PR.AT)
  - ❑ Educate users on the responsible and ethical use of AI
- ❑ Implement security controls to protect data-at-rest, data-in-transit, and data-in-use, including all training data, inputs, and outputs. (PR.DS)
  - ❑ Enforce utilization of encryption at all stages
  - ❑ Prohibit training models on sensitive data, consider sanitization and anonymization techniques where possible
  - ❑ Apply input and output validation, considering them untrusted at each step
  - ❑ Validate the accuracy of outputs to prevent downstream usage of returned misinformation
- ❑ Enforce a data lifecycle, including archival and/or destruction of unnecessary or obsolete data (PR.DS)
  - ❑ Document and enforce, preferably with automation, data/model/service lifecycle. Stale data sitting unused is both expensive and risky.

- ☐ Secure your AI platform itself and all underlying technologies, including your training and production systems and their endpoints (PR.PS)
- ☐ Ensure logs are generated and retained for all user and administrative use of the AI platform and data (PR.PS)
- ☐ Maintain resiliency and availability of your AI system, prevent Denial-of-Service scenarios such as from malicious network traffic, implement rate limits and resource caps (PR.IR)

Detect (DE)

- ☐ Monitor your AI platform, data stores, and users for indicators of compromise and malicious use (DE.CM)
- ☐ If utilizing an AI service provider, ensure they are monitoring and responding to anomalous activity (DE.CM)
- ☐ Ensure logs from your AI platform, data stores, and users are integrated into your SIEM for correlation and analysis (DE.AE)

Respond (RS)

- ☐ Establish an Incident Response plan for your AI systems (RS.MA)
  - ☐ Perform regular tabletop exercises to validate and improve your incident response capabilities
- ☐ During any incidents, record all findings and actions taken, maintain a documented chain of custody, and perform root cause analysis (RS.AN)
- ☐ Establish communication patterns to notify internal and external stakeholders of incident response progress (RS.CO)
- ☐ Have plans and train staff to rapidly contain and eradicate any incidents (RS.MI)
  - ☐ Ensure clear escalation procedures are in place to prevent unnecessary delays

Recover (RC)

- ☐ Ensure the recovery portion of your AI incident response plan includes verifying systems/data integrity before bringing the system back online, and documenting all steps and findings (RC.RP)
- ☐ Establish communication patterns to notify internal and external stakeholders of incident recovery including performing any necessary public disclosures (RC.CO)

Want to learn more about security, compliance and accelerating your business through automation?

Sign up for our newsletter and follow us on **LinkedIn**.

Have a question?

We'd love to hear from you.

Reach out today.

ScaleSec