SECURITYLAB · GUIDA OPERATIVA

Sicurezza e Conformità GDPR nell'Adozione dei Modelli AI Open-Weight

Edizione: Novembre 2026
Uso: Interno / Clienti

> **Obiettivo del Documento:** Questa guida operativa fornisce a Responsabili IT, CISO, DPO e sviluppatori le procedure tecniche e organizzative necessarie per validare, isolare e monitorare i modelli linguistici open-weight eseguiti su infrastrutture locali o cloud private. L'obiettivo primario è prevenire backdoor, avvelenamento dei pesi (data poisoning), esfiltrazione di dati aziendali e contestazioni di non conformità ai sensi degli articoli 25, 32 e 35 del GDPR.

Fase 1: Governance, Policy e Valutazione d'Impatto (Pre-Deployment)

Prima di scaricare o integrare qualsiasi modello di intelligenza artificiale, l'organizzazione deve stabilire un perimetro di controllo formale.

1.1 Blocco dello "Shadow AI" e Autorizzazione Centralizzata

- **Regolamento interno:** Aggiornare la policy aziendale sull'uso degli strumenti informatici, vietando espressamente il download, l'esecuzione locale e l'integrazione di pesi o modelli AI non preventivamente validati dal team IT e approvati dal DPO.
- **Canale di richiesta:** Istituire una procedura interna tramite cui i team di sviluppo e di business devono documentare la finalità del trattamento, le categorie di dati trattati (es. segreti commerciali, dati ex art. 9 GDPR) e il repository d'origine.

1.2 Redazione della DPIA Specifica per l'AI (Art. 35 GDPR)

Prima del collaudo in produzione, è obbligatorio documentare la Valutazione d'Impatto sulla Protezione dei Dati:

- **Mappatura del flusso dati:** Tracciare dove risiedono i prompt inviati dagli utenti e con quale frequenza vengono distrutti.
- **Analisi dei rischi:** Formalizzare il rischio di violazione della riservatezza e di integrità (alterazione degli output causata da backdoor).
- **Parere del DPO:** Registrare le misure compensative individuate per attenuare i rischi.

Fase 2: Validazione della Supply Chain del Software e del Modello

I pesi e i file dei modelli AI sono vettori di esecuzione arbitraria di codice. Il download da repository aperte (es. Hugging Face, GitHub) richiede controlli rigorosi.

2.1 Verifica di Provenienza e Integrità Crittografica

- **Autorialità accreditata:** Scaricare modelli solo da organizzazioni con account verificati ed evitare fork, mirror non ufficiali o ri-upload da utenti terzi.
- **Controllo Hash SHA-256:** Calcolare sempre l'impronta crittografica dei file scaricati e confrontarla con l'hash dichiarato dall'autore ufficiale.

  ```
  sha256sum nome_modello.safetensors
  ```

- **Firma crittografica:** Verificare la firma digitale del commit per scongiurare sostituzioni silenziose degli artefatti.

2.2 Standardizzazione sui Formati Sicuri

- **Divieto di formati a rischio:** Bloccare l'uso di `.bin`, `.pkl`, `.pt` (basati su PyTorch standard). Questi utilizzano `pickle`, che consente l'esecuzione immediata di codice arbitrario all'atto del caricamento.
- **Adozione esclusiva di safetensors:** Consentire l'importazione solo di modelli distribuiti nel formato aperto `safetensors`, progettato per archiviare tensori senza consentire payload eseguibili.

Fase 3: Architettura di Esecuzione e Hardening di Rete

Un modello compromesso non deve avere la possibilità fisica di comunicare all'esterno né di infettare il nodo host aziendale.

3.1 Isolamento di Rete (Zero-Egress Filtering)

- **Firewall perimetrale:** Configurare il server di inferenza in una DMZ o subnet isolata con blocco totale del traffico in uscita verso Internet (egress deny-all).
- **Inibizione DNS:** Rimuovere server DNS pubblici dalle configurazioni di rete della macchina host.
- **Proxy autorizzativo:** Se il modello deve interrogare database vettoriali aziendali interni (es. RAG), abilitare unicamente connessioni verso quegli specifici indirizzi IP interni.

3.2 Sandboxing e Privilegi Minimi

- **Esecuzione non root:** Eseguire i processi (es. vLLM, Ollama) in container configurati con utente non privilegiato (`USER nonroot`).
- **Filesystem in sola lettura:** Montare la cartella contenente i pesi del modello in modalità sola lettura (`read-only`).

Fase 4: Protezione Operativa dei Dati (Runtime Security)

4.1 Filtro Preventivo dei Prompt (Data Masking)

Interporre un gateway API tra l'utente finale e il modello AI configurato per sostituire automaticamente dati personali (codici fiscali, IBAN, indirizzi email) con identificatori fittizi prima che il testo raggiunga il modello. E' inoltre fondamentale bloccare stringhe note di *Prompt Injection*.

4.2 Ispezione e Monitoraggio

Inviare i log di accesso, le metriche di carico e gli errori di runtime al SIEM aziendale, avendo cura di escludere dal payload il testo in chiaro dei prompt per evitare log-poisoning o fughe di dati involontarie.

Checklist Operativa di Rilascio in Produzione

Prima di rendere accessibile un modello AI all'interno dell'ambiente aziendale, completare obbligatoriamente la seguente tabella:

Ambito	Controllo Operativo	Responsabile	Esito Richiesto
Governance	DPIA redatta e approvata	DPO / Legal	**Formalizzata**
Governance	Registro dei trattamenti aggiornato	DPO	**Registrato**
Supply Chain	Provenienza e Account Autore verificati	IT Security	**Accreditata**
Supply Chain	Formato file limitato a safetensors	Dev / MLOps	**100% Conforme**
Supply Chain	Corrispondenza Hash SHA-256 validata	IT Security	**Validata**
Infrastruttura	Egress filtering attivo (No Internet)	Network Admin	**Bloccato**
Infrastruttura	Esecuzione processo in container non-root	SysAdmin	**Verificato**
Sicurezza	Mascheramento dati nei prompt attivo	IT / Dev	**Operativo**

Piano di Risposta agli Incidenti (Incident Response)

Qualora vengano rilevate anomalie, risposte sospette o tentativi di connessione esterna da parte dell'ambiente di inferenza:

1. **Isolamento Immediato:** Disconnettere la scheda di rete del server senza spegnere la macchina (per preservare la memoria per l'analisi forense).
2. **De-provisioning:** Bloccare l'accesso al modello da parte di qualsiasi applicativo o utente aziendale.
3. **Analisi Rischio Data Breach:** Determinare quali prompt sono stati elaborati e se c'è stata esfiltrazione. In caso di rischio per gli interessati, effettuare la notifica al Garante Privacy entro 72 ore (Art. 33 GDPR).