AGID | Agenzia per l'Italia Digitale

LINEE GUIDA IA: L'ADOZIONE NELLA PUBBLICA AMMINISTRAZIONE

SICUREZZA CIBERNETICA

Sicurezza dei sistemi di IA

Insieme di strumenti, strategie e processi implementati per identificare e prevenire le minacce che potrebbero compromettere la riservatezza, l'integrità o la disponibilità di un modello di IA o di un sistema abilitato all'IA.

Gestione rischio cibernetico

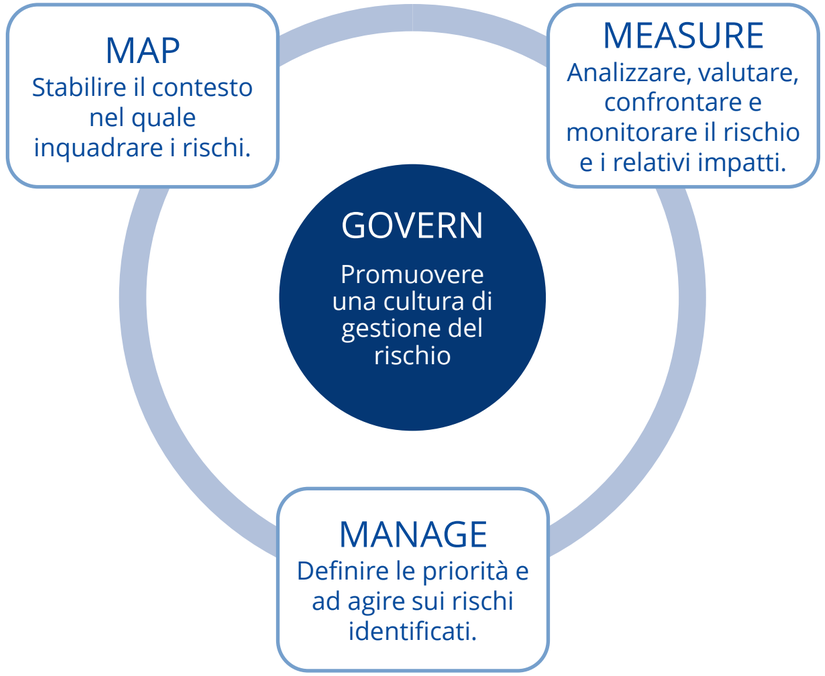
I sistemi di IA sono caratterizzati da specifici rischi.

L'AI Act , stabilisce regole armonizzate sulla base di del rischio e del livello d'impatto.

Devono essere considerate le caratteristiche distintive dei sistemi.

E' possibile fare riferimento a framework standard come l' AI RMF sviluppato dal NIST.

Risk Management Framework



Tassonomie di attacco

Evasion attacks (P/G)

- Errori nella classificazione del modello (Adversarial examples)

Poisoning attacks (P/G)

- Degradare le prestazioni del modello o far generare uno specifico risultato alterando i dati di addestramento del modello.

Privacy attacks (P/G)

- Compromettere le informazioni degli utenti ricostruendole a partire dai dati di addestramento.

Abuse attacks (G)

- Alterare il comportamento di un sistema di IA generativa per adattarlo ai propri scopi.

Obiettivi di sicurezza

Adottare l’IA in modo responsabile	Identificare, tracciare, mantenere e proteggere gli asset	Proteggere la catena di approvvigionamento
Proteggere il modello e i dati	Monitorare il comportamento del sistema e degli input	Sviluppare un piano di risposta agli incidenti
Formare e sensibilizzare il personale sulle minacce e sui rischi	Proteggere l'infrastruttura ICT	Proteggere le identità e gli accessi

Gestione integrata della sicurezza dei sistemi di IA

AGID | Agenzia per l'Italia Digitale

Grazie