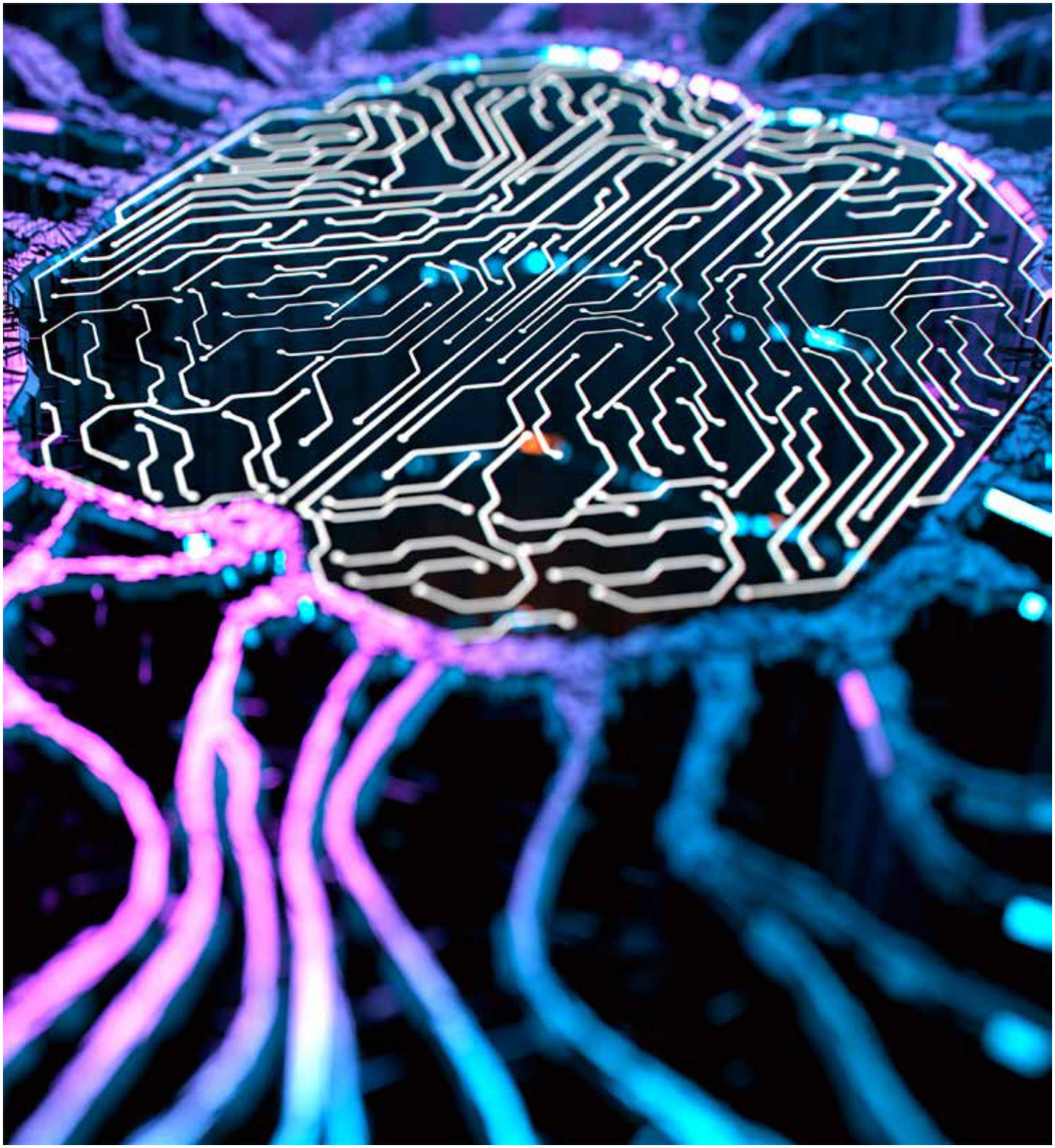
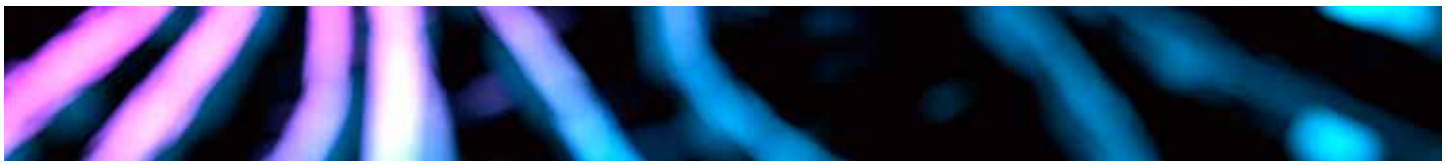
Linee guida per lo sviluppo di sistemi di intelligenza artificiale sicuri





National Cyber Security Centre
a part of GCHQ

CISA
CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Communications Security Establishment
Canadian Centre for Cyber Security
Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

CSIRT
Equipo de Respuesta ante Incidentes de Seguridad Informática

National Cyber and Information Security Agency

Republic of Estonia
Information System Authority

Department of Justice
Federal Bureau of Investigation

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Federal Office for Information Security

INCD
Israel National Cyber Directorate

AGENZIA PER LA CYBERSICUREZZA NAZIONALE
ACN
SAPIENTIA ADIUVAT

NISC
内閣サイバーセキュリティセンター
National center of Incident readiness and Strategy for Cybersecurity

National Cyber Security Centre

NITDA

NSM
NORWEGIAN NATIONAL CYBER SECURITY CENTRE

National Security Agency
United States of America

NASK

국가정보원
National Intelligence Service

Ministerstwo Cyfryzacji

CSA SINGAPORE
Cyber Security Agency of Singapore

Informazioni su questo documento

Questo documento è pubblicato dal Centro nazionale per la sicurezza informatica (National Cyber Security Centre, NCSC) del Regno Unito, dalla Agenzia per la sicurezza informatica e la sicurezza delle infrastrutture (Cybersecurity and Infrastructure Security Agency, CISA) degli Stati Uniti e dai seguenti partner internazionali:

- Agenzia per la sicurezza nazionale (National Security Agency, NSA)
- Ufficio federale di investigazione (Federal Bureau of Investigations, FBI)
- Centro australiano per la sicurezza informatica della direzione dei segnali australiani (Australian Signals Directorate's Australian Cyber Security Centre, ACSC)
- Centro canadese per la sicurezza informatica (Canadian Centre for Cyber Security, CCCS)
- Centro nazionale neozelandese per la sicurezza informatica (New Zealand National Cyber Security Centre, NCSC-NZ)
- Team di risposta agli incidenti di sicurezza informatica del governo cileno (Equipo de Respuesta ante Incidentes de Seguridad Informática, CSIRT)
- Agenzia nazionale ceca per la sicurezza informatica e delle informazioni (NUKIB)
- Autorità per i sistemi informativi dell'Estonia (RIA) e Centro nazionale per la sicurezza informatica dell'Estonia (NCSC-EE)
- Agenzia francese per la sicurezza informatica (ANSSI)
- Ufficio federale tedesco per la sicurezza delle informazioni (BSI)
- Direzione nazionale israeliana per la cibernetica (INCD)
- Agenzia per la cybersicurezza nazionale (ACN)
- Centro nazionale giapponese di preparazione agli incidenti e di strategia per la sicurezza informatica (NISC)
- Segretariato giapponese per la politica della scienza, della tecnologia e dell'innovazione, Ufficio del Gabinetto
- Agenzia nazionale nigeriana per lo sviluppo delle tecnologie dell'informazione (NITDA)
- Centro nazionale norvegese per la sicurezza informatica (NCSC-NO)
- Ministero degli Affari Digitali della Polonia
- Istituto nazionale di ricerca polacco NASK (NASK)
- Servizio nazionale di intelligence della Repubblica di Corea (NIS)
- Agenzia per la sicurezza informatica di Singapore (CSA)

Ringraziamenti

Le seguenti organizzazioni hanno contribuito allo sviluppo di queste linee guida:

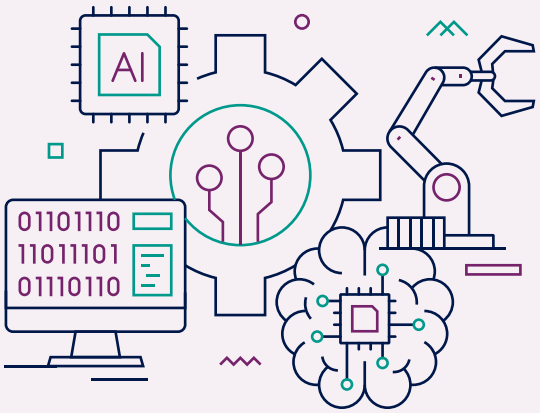
- Alan Turing Institute
- Anthropic
- Databricks
- Centro per la sicurezza e le tecnologie emergenti (Center for Security and Emerging Technology) dell'Università Georgetown
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Istituto per l'ingegneria del software (Software Engineering Institute) presso la Carnegie Mellon University
- Centro Stanford per la sicurezza dell'intelligenza artificiale
- Programma Stanford su geopolitica, tecnologia e governance

Dichiarazione di non responsabilità

Le informazioni contenute in questo documento sono fornite "come tali" dal NCSC e dalle organizzazioni che ne sono autrici, le quali non sono responsabili per eventuali perdite, lesioni o danni di qualsiasi tipo causati dal loro utilizzo, salvo quanto previsto dalla legge. Le informazioni contenute in questo documento non costituiscono o implicano l'approvazione o la raccomandazione di alcuna organizzazione, prodotto o servizio fornito da terzi da parte dell'NCSC e delle agenzie autrici. I link e i riferimenti a siti web e a materiali di terzi sono forniti a titolo puramente informativo e non rappresentano un'approvazione o una raccomandazione di tali risorse rispetto ad altre.

Questo documento è disponibile su base TLP:CLEAR (https://www.first.org/tlp/).

Indice dei contenuti



Sintesi dei contenuti

Questo documento raccomanda delle linee guida per i fornitori di qualsiasi sistema che utilizzi l'intelligenza artificiale (IA), sia che tali sistemi siano stati creati da zero sia che siano stati costruiti sulla base di strumenti e servizi forniti da altri. L'implementazione di queste linee guida aiuterà i fornitori a costruire sistemi di IA che funzionino come previsto, siano disponibili quando necessario e lavorino senza rivelare dati sensibili a parti non autorizzate.

Questo documento si rivolge principalmente ai fornitori di sistemi di IA che utilizzano modelli ospitati da un'organizzazione o che utilizzano interfacce di programmazione delle applicazioni (API) esterne. Invitiamo **tutti** gli stakeholder (compresi i data scientist, gli sviluppatori, i manager, i decisori e i gestori del rischio) a leggere queste linee guida affinché possano prendere decisioni informate su **progettazione**, **sviluppo**, **dispiegamento** ed **esercizio** dei loro sistemi di intelligenza artificiale.

Informazioni sulle linee guida

I sistemi di IA hanno il potenziale per apportare molti benefici alla società. Tuttavia, affinché le opportunità offerte dall'IA possano essere realizzate appieno, è necessario che essa venga sviluppata, dispiegata e gestita in modo sicuro e responsabile.

I sistemi di intelligenza artificiale sono soggetti a nuove vulnerabilità di sicurezza che devono essere considerate unitamente alle minacce standard relative alla sicurezza informatica. Quando il ritmo di sviluppo è elevato, come nel caso dell'IA, la sicurezza può spesso diventare una considerazione secondaria. La sicurezza deve essere un requisito fondamentale, non solo nella fase di sviluppo, ma durante tutto il ciclo di vita del sistema.

Per questo motivo, le linee guida sono suddivise in quattro aree chiave all'interno del ciclo di vita dello sviluppo del sistema di IA: **progettazione sicura**, **sviluppo sicuro**, **dispiegamento sicuro** ed **esercizio e manutenzione sicuri**. In ciascuna sezione vengono suggerite considerazioni e mitigazioni che contribuiscono a ridurre il rischio complessivo del processo di sviluppo di un sistema di IA di un'organizzazione.

1. **Progettazione sicura**
   Questa sezione contiene linee guida che si applicano alla fase di progettazione del ciclo di vita dello sviluppo del sistema di intelligenza artificiale. Si occupa della comprensione dei rischi e della modellazione delle minacce, nonché di argomenti specifici e compromessi da considerare nella progettazione di sistemi e modelli.

2. **Sviluppo sicuro**
   Questa sezione contiene linee guida che si applicano alla fase di sviluppo del ciclo di vita del sistema di IA, compresa la sicurezza della catena di approvvigionamento, la documentazione e la gestione degli asset e del debito tecnico.

3. **Dispiegamento sicuro**
   Questa sezione contiene linee guida che si applicano alla fase di dispiegamento del ciclo di vita del sistema di IA, compresa la protezione dell'infrastruttura e dei modelli da compromissioni, minacce o perdite, lo sviluppo di processi di gestione degli incidenti e il rilascio responsabile.

4. **Esercizio e manutenzione sicuri**
   Questa sezione contiene linee guida che si applicano alla fase di esercizio e manutenzione del ciclo di vita dello sviluppo del sistema di IA. Questa sezione fornisce linee guida riguardanti azioni particolarmente rilevanti una volta che un sistema è stato dispiegato, tra cui la registrazione e il monitoraggio, la gestione degli aggiornamenti e la condivisione delle informazioni.

Le linee guida seguono un approccio "sicuro per impostazione predefinita" e sono strettamente allineate alle pratiche definite nella Secure development and deployment guidance (Guida allo sviluppo e alla distribuzione sicura) dell'NCSC, nel Secure Software Development Framework (Quadro di sviluppo di software sicuro) del NIST e nei "Secure by design principles (Principi di sicurezza sin dalla progettazione)) pubblicati dal CISA, dall'NCSC e dalle agenzie cibernetiche internazionali. Le linee guida danno priorità:

- all'assunzione di responsabilità dei risultati di sicurezza per i propri clienti;
- all'adozione di pratiche di trasparenza e di una responsabilità radicali;
- alla costruzione di una struttura organizzativa e di una leadership tali da rendere la sicurezza una priorità aziendale assoluta.

Introduzione

I sistemi di intelligenza artificiale (IA) hanno il potenziale per apportare molti benefici alla società. Tuttavia, per sfruttare appieno le opportunità offerte dall'IA, è necessario che essa venga sviluppata, dispiegata e gestita in modo sicuro e responsabile. La sicurezza informatica è un prerequisito necessario per la sicurezza, la resilienza, la privacy, l'equità, l'efficacia e l'affidabilità dei sistemi di IA.

I sistemi di intelligenza artificiale sono soggetti a nuove vulnerabilità di sicurezza che devono essere considerate unitamente alle minacce standard alla sicurezza informatica. Quando il ritmo di sviluppo è elevato, come nel caso dell'IA, la sicurezza può spesso diventare una considerazione secondaria. La sicurezza deve essere un requisito fondamentale, non solo nella fase di sviluppo, ma durante tutto il ciclo di vita del sistema.

Questo documento raccomanda delle linee guida per i fornitori[1] di qualsiasi sistema che utilizzi l'intelligenza artificiale (IA), sia che tali sistemi siano stati creati da zero sia che siano stati costruiti sulla base di strumenti e servizi forniti da altri. L'implementazione di queste linee guida aiuterà i fornitori a costruire sistemi di IA che funzionino come previsto, siano disponibili quando necessario e lavorino senza rivelare dati sensibili a parti non autorizzate.

Queste linee guida devono essere prese in considerazione insieme alle migliori pratiche consolidate in materia di sicurezza informatica, gestione del rischio e risposta agli incidenti. In particolare, invitiamo i fornitori a seguire i "Secure by design principles" (principi di sicurezza fin dalla progettazione)[2] sviluppati dalla Cybersecurity and Infrastructure Security Agency (CISA) statunitense, dal National Cyber Security Centre (NCSC) britannico e da tutti i nostri partner internazionali. I principi danno priorità:

- all'assunzione di responsabilità dei risultati di sicurezza per i propri clienti;
- all'adozione di pratiche di trasparenza e di una responsabilità radicali;
- alla costruzione di una struttura organizzativa e di una leadership tali da rendere la sicurezza una priorità aziendale assoluta.

Seguire i principi di "sicurezza fin dalla progettazione" richiede risorse significative durante tutto il ciclo di vita di un sistema. Significa che gli sviluppatori devono investire per dare priorità a **caratteristiche**, **meccanismi** e all'**implementazione** di strumenti che proteggano i clienti a ogni livello della progettazione del sistema e in tutte le fasi del ciclo di vita dello sviluppo. In questo modo si potranno evitare costose riprogettazioni successive e si potranno salvaguardare i clienti e i loro dati nel breve termine.

Perché la sicurezza dell'IA è differente?

In questo documento usiamo l'acronimo "IA" per riferirci specificamente alle applicazioni di apprendimento automatico (Machine Learning - ML)[3]. Nel suo campo di applicazione sono compresi tutti i tipi di apprendimento automatico. Definiamo le applicazioni di apprendimento automatico come applicazioni che:

- prevedono componenti software (modelli) che consentono ai computer di riconoscere e contestualizzare gli schemi nei dati senza che le regole debbano essere programmate esplicitamente da un essere umano;
- generano previsioni, raccomandazioni o decisioni basate su ragionamenti statistici.

Oltre alle minacce alla sicurezza informatica esistenti, i sistemi di intelligenza artificiale sono soggetti a nuovi tipi di vulnerabilità. Il termine "apprendimento automatico antagonistico" (adversarial machine learning - AML) è utilizzato per descrivere lo sfruttamento di vulnerabilità fondamentali nelle componenti di apprendimento automatico, compresi hardware, software, flussi di lavoro e catene di approvvigionamento. L'apprendimento automatico antagonistico consente agli aggressori di provocare comportamenti indesiderati nei sistemi di apprendimento automatico che possono includere:

- effetti sulle prestazione di classificazione o regressione del modello;
- la possibilità per gli utenti di eseguire azioni non autorizzate;
- l'estrazione di informazioni sensibili sul modello.

Ci sono molti modi per ottenere questi effetti, come ad esempio gli attacchi di tipo prompt injection nel dominio dei modelli linguistici di grandi dimensioni (large language model - LLM), o la corruzione deliberata dei dati di addestramento o dei feedback degli utenti (nota come "avvelenamento dei dati").

Chi dovrebbe leggere questo documento?

Questo documento si rivolge principalmente ai fornitori di sistemi di IA, sia che si basino su modelli ospitati da un'organizzazione sia che utilizzino interfacce di programmazione delle applicazioni (API) esterne . Tuttavia, invitiamo **tutti** gli stakeholder (compresi i data scientist, gli sviluppatori, i manager, i decision maker e i proprietari del rischio) a leggere queste linee guida per aiutarli a prendere decisioni informate su **progettazione**, **dispiegamento** ed **esercizio** dei loro sistemi di IA ad apprendimento automatico.

Detto questo, non tutte le linee guida saranno direttamente applicabili a tutte le organizzazioni. Il livello di sofisticazione e i metodi di attacco variano a seconda dell'avversario che prende di mira il sistema di intelligenza artificiale, pertanto le linee guida devono essere considerate insieme ai casi d'uso e al profilo di minaccia della propria organizzazione.

Chi è responsabile dello sviluppo di un'IA sicura?

Nelle moderne catene di approvvigionamento dell'IA sono spesso presenti molti attori. Un approccio semplice presuppone due entità:

- il "fornitore" che è responsabile della cura dei dati, dello sviluppo di algoritmi, della progettazione, del dispiegamento e della manutenzione;
- l'"utente", che fornisce gli input e riceve gli output.

Sebbene questo approccio fornitore-utente sia utilizzato in molte applicazioni, sta diventando sempre più raro[4], poiché i fornitori possono cercare di incorporare nei propri sistemi software, dati, modelli e/o servizi remoti forniti da terze parti. Queste complesse catene di approvvigionamento rendono più difficile per gli utenti finali capire dove risiede la responsabilità per la sicurezza dell'IA.

Gli utenti (siano essi "utenti finali" o fornitori che incorporano una componente esterna di IA[5]) non hanno in genere visibilità e/o competenze sufficienti per comprendere, valutare o affrontare appieno i rischi associati ai sistemi che utilizzano. Pertanto, in linea con i principi di "sicurezza fin dalla progettazione", **i fornitori di componenti di IA dovrebbero assumersi la responsabilità dei risultati in termini di sicurezza degli utenti lungo la catena di approvvigionamento**.

I fornitori devono implementare i controlli e le mitigazioni di sicurezza, ove possibile, all'interno dei loro modelli, pipeline e/o sistemi e, laddove si utilizzano le impostazioni, implementare come predefinita l'opzione più sicura. Laddove i rischi non possano essere mitigati, il fornitore dovrebbe avere la responsabilità di:

- informare gli utenti che si trovano più in basso nella catena di approvvigionamento in merito ai rischi che essi e (se del caso) i loro stessi utenti stanno accettando;
- consigliare loro come utilizzare il componente in modo sicuro.

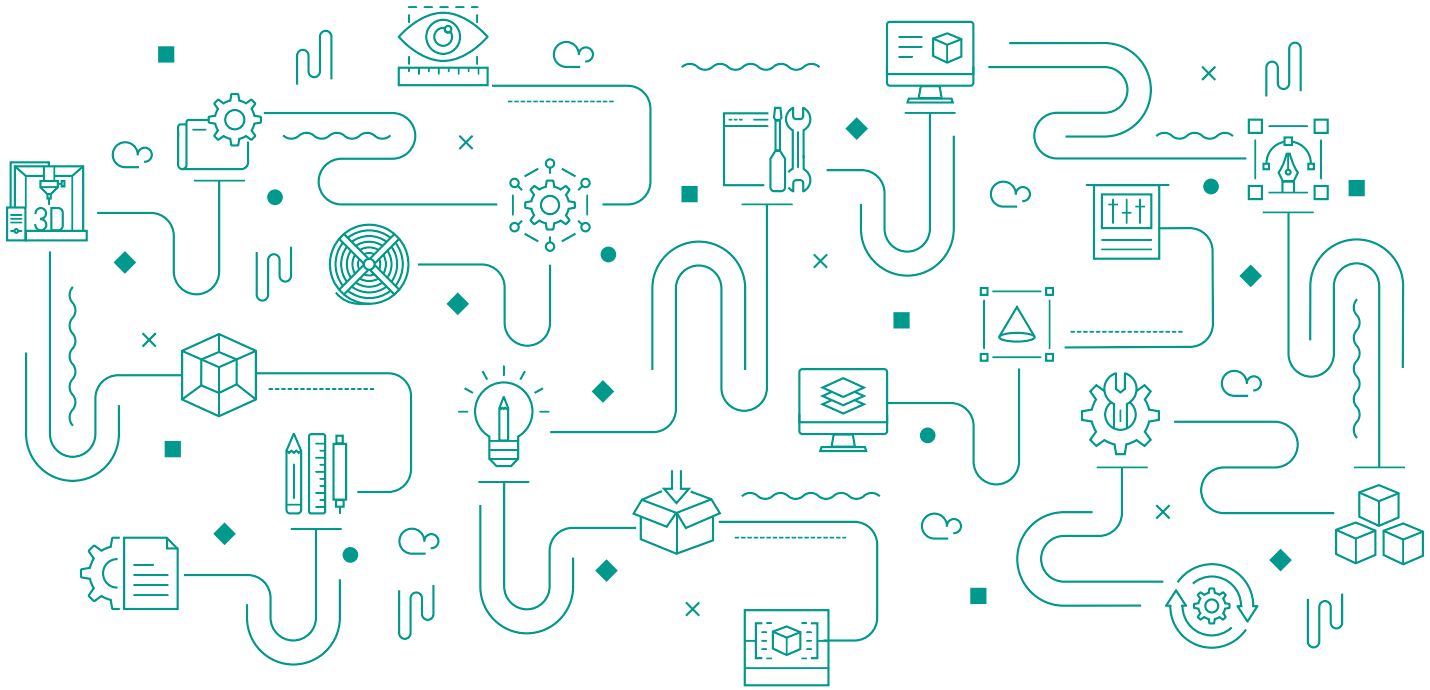
Quando la compromissione del sistema potrebbe portare a danni fisici o reputazionali tangibili o diffusi, a una perdita significativa delle attività aziendali, alla fuga di informazioni sensibili o riservate e/o a implicazioni legali, i rischi per la sicurezza informatica dell'IA devono essere trattati come **critici**.

Linee guida per lo sviluppo di sistemi di intelligenza artificiale sicuri

Le linee guida sono suddivise in quattro aree chiave all'interno del ciclo di vita dello sviluppo del sistema di IA: **progettazione sicura**, **sviluppo sicuro**, **dispiegamento sicuro** ed **esercizio e manutenzione sicuri**. Per ogni area, suggeriamo considerazioni e mitigazioni che contribuiranno a ridurre il rischio complessivo del processo di sviluppo del sistema di IA dell'organizzazione.

Le linee guida contenute in questo documento sono strettamente allineate alle pratiche del ciclo di vita dello sviluppo del software definite all'interno delle seguenti guide:

- Secure development and deployment (Sviluppo e distribuzione sicuri)dell'NCSC
- Secure Software Development Framework (Quadro di sviluppo di software sicuro) (SSDF)[6] del National Institute of Standards and Technology (NIST)



1. Progettazione sicura

Questa sezione contiene linee guida che si applicano alla fase della **progettazione** del ciclo di vita dello sviluppo del sistema di IA. Si occupa della comprensione dei rischi e della modellazione delle minacce, nonché di argomenti specifici e di compromessi da considerare nella progettazione di sistemi e modelli.

Sensibilizza il personale sui rischi e sulle minacce

I proprietari dei sistemi e gli alti dirigenti comprendono le minacce alla sicurezza dell'IA e le relative mitigazioni. I data scientist e gli sviluppatori sono consapevoli delle minacce alla sicurezza e delle modalità di fallimento e aiutano i proprietari dei rischi a prendere decisioni informate. Fornisci agli utenti indicazioni sui rischi di sicurezza unici che i sistemi di IA devono affrontare (ad esempio, come parte della formazione standard di InfoSec) e forma gli sviluppatori sulle tecniche di codifica sicura e sulle pratiche di IA sicure e responsabili.

Modella le minacce al sistema

Nell'ambito del processo di gestione del rischio, è necessario applicare un processo olistico per valutare le minacce al sistema, che comprende la comprensione dei potenziali impatti sul sistema, sugli utenti, sulle organizzazioni e sulla società in generale nel caso in cui un componente dell'IA venga compromesso o si comporti in modo inaspettato[7]. Questo processo prevede la valutazione dell'impatto delle minacce specifiche dell'IA[8] e la documentazione del processo decisionale.

Sii ben consapevole che la sensibilità e i tipi di dati utilizzati nel sistema possono influenzare il suo valore come bersaglio da parte di un aggressore. La tua valutazione dovrebbe considerare che alcune minacce potrebbero aumentare, dato che i sistemi di IA vengono sempre più considerati come obiettivi di alto valore e che le IA stesse consentono nuovi vettori di attacco automatizzati.

Progetta il sistema in funzione della sicurezza, della funzionalità e delle prestazioni.

Ritieni che il compito in questione sia affrontato nel modo più appropriato utilizzando l'IA. Una volta stabilito questo, procedi a valutare l'adeguatezza delle scelte progettuali specifiche per l'IA. Considera il modello di minaccia e le mitigazioni di sicurezza associate insieme a funzionalità, esperienza dell'utente, ambiente di distribuzione, prestazioni, garanzia, supervisione, requisiti etici e legali, oltre ad altre considerazioni. Per esempio:

- considera la sicurezza della catena di approvvigionamento quando scegli, ad esempio, se sviluppare internamente o utilizzare componenti esterni:
    - considera la scelta di addestrare un nuovo modello, di utilizzare un modello esistente (con o senza messa a punto) o di accedere a un modello tramite un'API esterna;
    - se scegli di lavorare con un fornitore esterno di modelli, dovrai effettuare una valutazione di dovuta diligenza della postura di sicurezza del fornitore;
    - se utilizzi una libreria esterna, sarà necessario completare una valutazione di dovuta diligenza (ad esempio, per assicurarsi che la libreria disponga di controlli che impediscano al sistema di caricare modelli non attendibili senza esporsi immediatamente all'esecuzione di codice arbitrario[9]);
    - implementi la scansione e l'isolamento/sandboxing quando importi modelli di terze parti o pesi serializzati, che dovrebbero essere trattati come codice di terze parti non attendibile e potrebbero consentire l'esecuzione di codice remoto;

  - se utilizzi API esterne, applichi controlli appropriati ai dati che possono essere inviati a servizi al di fuori del controllo della propria organizzazione, ad esempio richiedendo agli utenti di effettuare il login e di confermare prima di inviare informazioni potenzialmente sensibili;
  - applichi controlli appropriati ed effettui la sanificazione dei dati e degli input; questo include quando si incorpora il feedback dell'utente o i dati di apprendimento continuo nel modello, riconoscendo che i dati di addestramento definiscono il comportamento del sistema.

- integri lo sviluppo del sistema software di IA nelle migliori pratiche di sviluppo e operazioni sicure esistenti; tutti gli elementi del sistema di IA sono scritti in ambienti appropriati utilizzando pratiche e linguaggi di codifica che riducono o eliminano le classi di vulnerabilità note, ove plausibile;

- se i componenti dell'IA devono attivare azioni, ad esempio modificare i file o indirizzare l'output a sistemi esterni, applichi restrizioni appropriate alle azioni possibili (se necessario, questo include IA esterne e fail-safe non IA);

- Le decisioni relative all'interazione con l'utente sono informate dai rischi specifici dell'IA, ad esempio:

  - il sistema fornisce agli utenti output utilizzabili senza rivelare inutili livelli di dettaglio a un potenziale aggressore;
  - se necessario, il sistema fornisce un'efficace protezione per gli output del modello;
  - se si offre un'API a clienti o collaboratori esterni, applichi controlli appropriati per mitigare gli attacchi al sistema di intelligenza artificiale attraverso l'API;
  - integri le impostazioni più sicure nel sistema per impostazione predefinita;
  - applichi i principi del minimo privilegio per limitare l'accesso alle funzionalità di un sistema;
  - agli utenti vengono spiegate le funzionalità più rischiose e si richiede loro di scegliere espressamente di utilizzarle; i casi d'uso vietati vengono espressamente comunicati e, ove possibile, si informano gli utenti riguardo a soluzioni alternative.

Considera i vantaggi e i compromessi in termini di sicurezza quando scegli il modello di IA

La scelta del modello di IA comporterà il bilanciamento di una serie di requisiti. Ciò include la scelta dell'architettura del modello, della configurazione, dei dati di addestramento, dell'algoritmo di addestramento e degli iperparametri. Le decisioni da prendere si basano sul modello di minaccia e vengono regolarmente rivalutate in base ai progressi della ricerca sulla sicurezza dell'intelligenza artificiale e all'evoluzione della comprensione della minaccia.

Quando si sceglie un modello di IA, le considerazioni probabilmente includono, ma non si limitano a:

- la complessità del modello che si sta utilizzando, cioè l'architettura e il numero di parametri scelti; l'architettura e il numero di parametri scelti per il modello influiscono, tra gli altri fattori, sulla quantità di dati di addestramento necessari e sulla sua resistenza alle variazioni dei dati di input quando è in uso;

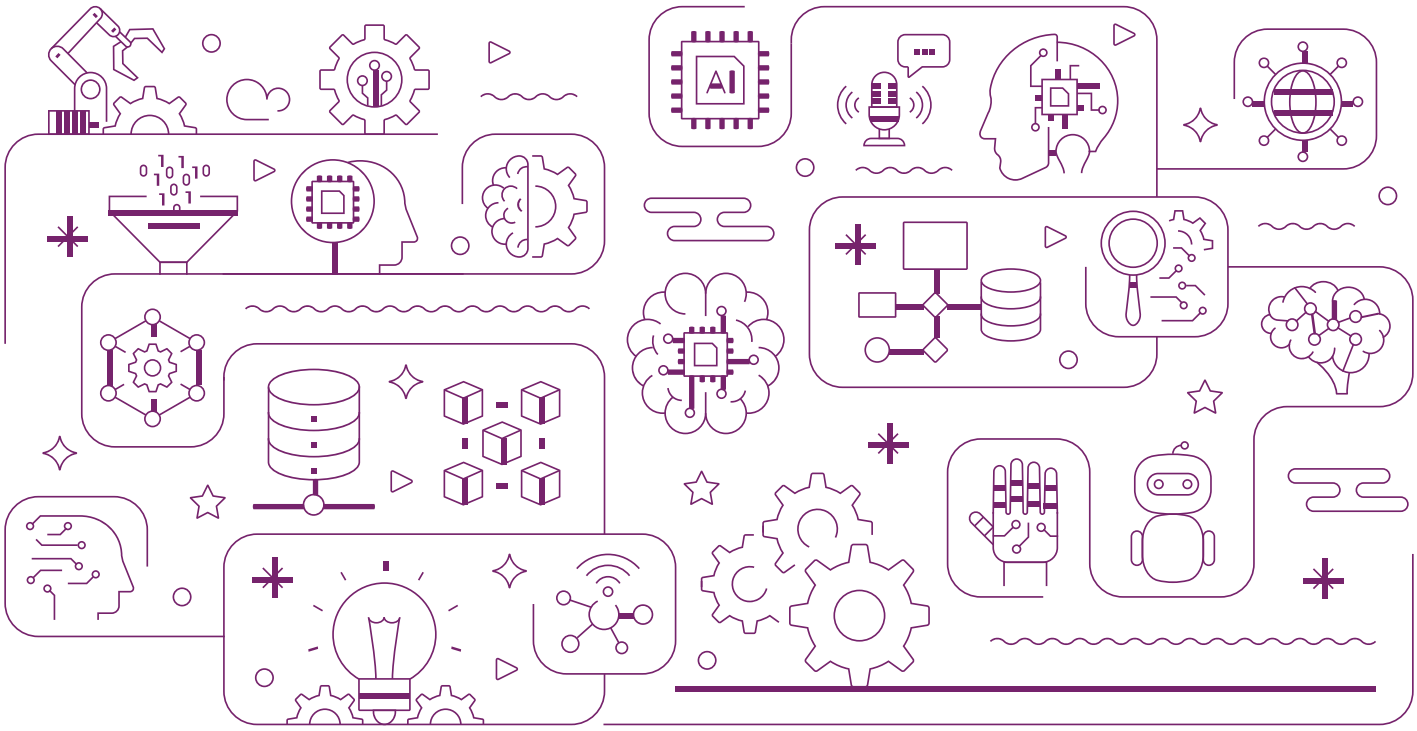
- l'adeguatezza del modello per il proprio caso d'uso e/o la fattibilità di adattamento alle proprie esigenze specifiche (ad esempio con una messa a punto);

- la capacità di allineare, interpretare e spiegare i risultati del modello (ad esempio per il debugging, l'audit o la conformità alle normative); a questo proposito può essere vantaggioso utilizzare modelli più semplici e trasparenti rispetto a modelli ampi e complessi, più difficili da interpretare;

- caratteristiche dei set di dati di addestramento, tra cui dimensioni, integrità, qualità, sensibilità, età, rilevanza e diversità;

- il valore dell'utilizzo di tecniche di hardening del modello (come l'adversarial training), di regolarizzazione e/o di rafforzamento della privacy;
- la provenienza e le catene di approvvigionamento dei componenti, tra cui il modello o il modello fondativo, i dati di formazione e gli strumenti associati.

Per ulteriori informazioni sull'impatto di molti di questi fattori sulla sicurezza, consulta i Principles for the Security of Machine Learning (Principi per la sicurezza dell'apprendimento automatico) dell'NCSC, in particolare Design for security - model architecture (Progettazione per la sicurezza - architettura del modello).



2. Sviluppo sicuro

Questa sezione contiene linee guida che si applicano alla fase dello **sviluppo** del ciclo di vita del sistema di IA, tra cui la sicurezza della catena di approvvigionamento, la documentazione e la gestione degli asset e del debito tecnico.

Proteggi la catena di approvvigionamento

Valuta e monitora la sicurezza delle catene di approvvigionamento dell'IA durante l'intero ciclo di vita di un sistema e richiedi ai fornitori di aderire agli stessi standard che la tua organizzazione applica ad altri software. Se i fornitori non possono aderire agli standard della tua organizzazione, dovrai agire in conformità con le politiche di gestione del rischio esistenti.

Laddove non siano prodotti internamente, è necessario acquisire e mantenere componenti hardware e software ben protetti e ben documentati (ad esempio, modelli, dati, librerie software, moduli, middleware, framework e API esterne) da sviluppatori commerciali, open source e di terze parti verificate, per garantire una solida sicurezza dei sistemi.

Se non vengono soddisfatti i criteri di sicurezza, bisogna essere pronti a passare a soluzioni alternative per i sistemi di importanza vitale. Puoi utilizzare risorse come la Supply Chain Guidance (Guida alla catena di approvvigionamento) dell'NCSC e quadri di riferimento come Supply Chain Levels for Software Artifacts, SLSA (Livelli della catena di fornitura per gli artefatti software)[10] per tracciare le attestazioni della catena di fornitura e i cicli di vita dello sviluppo del software .

Identifica, traccia e proteggi le risorse

Comprendi il valore per l'organizzazione delle risorse legate all'IA, inclusi modelli, dati (compresi i feedback degli utenti), suggerimenti, software, documentazione, registri e valutazioni (comprese le informazioni sulle capacità potenzialmente non sicure e sulle modalità di guasto), riconoscendo dove rappresentano un investimento significativo e dove l'accesso a tali risorse consente a un aggressore di poter agire. Gestisci i log come dati sensibili e implementa controlli per proteggerne la riservatezza, l'integrità e la disponibilità.

Sii a conoscenza di dove risiedono gli asset e valuta e accetta i rischi associati. Hai a disposizione processi e strumenti per tracciare, autenticare, controllare le versioni e proteggere le risorse, e sei inoltre in grado di ripristinare un buono stato conosciuto in caso di compromissione.

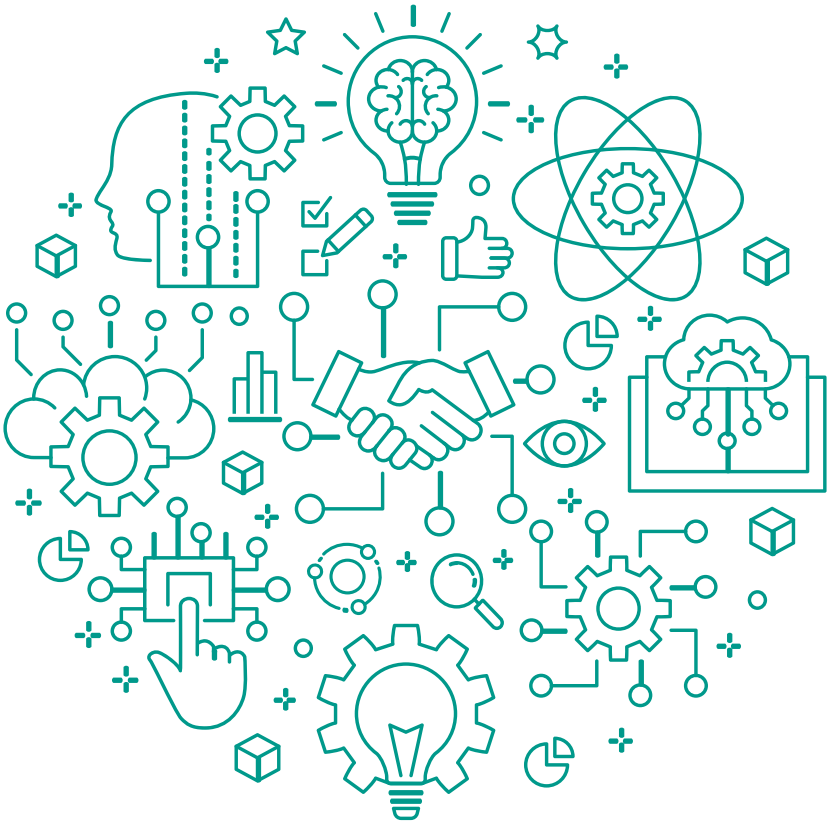
Metti a disposizione processi e controlli per gestire i dati a cui i sistemi di IA possono accedere e per gestire i contenuti generati dall'IA in base alla loro sensibilità (e alla sensibilità degli input che li hanno generati).

Documenta i dati, i modelli e i suggerimenti

Documenta la creazione, l'esercizio e la gestione del ciclo di vita di qualsiasi modello, insieme di dati e metaprodotti o sistema. La documentazione include informazioni rilevanti per la sicurezza, come le fonti dei dati di addestramento (compresi i dati di messa a punto e il feedback umano o di altri operatori), l'ambito e i limiti previsti, i guardrail, gli hash o le firme crittografiche, il tempo di conservazione, la frequenza di revisione suggerita e le potenziali modalità di errore. Tra le strutture utili a tal fine vi sono le schede modello, le schede dati e le distinte base del software (SBOM). La produzione di una documentazione esaustiva favorisce la trasparenza e la responsabilità[11].

Gestisci il debito tecnico

Come per qualsiasi sistema software, è necessario identificare, tracciare e gestire il "debito tecnico" durante il ciclo di vita di un sistema di intelligenza artificiale (il debito tecnico è quello che riguarda le decisioni ingegneristiche che non rispettano le migliori pratiche per ottenere risultati a breve termine, a scapito dei benefici a lungo termine). Come il debito finanziario, anche il debito tecnico non è intrinsecamente negativo, ma deve essere gestito fin dalle prime fasi dello sviluppo[12]. Sii ben consapevole che agire in questo senso può essere più impegnativo in un contesto di IA rispetto a un software standard e che i livelli di debito tecnico saranno probabilmente elevati a causa dei rapidi cicli di sviluppo e della mancanza di protocolli e interfacce consolidati. Assicurati che i piani del ciclo di vita (compresi i processi di dismissione dei sistemi di IA) valutino, riconoscano e riducano i rischi per i futuri sistemi simili.



3. Dispiegamento sicuro

Questa sezione contiene linee guida che si applicano alla fase del **dispiegamento** del ciclo di vita dello sviluppo del sistema di IA, compresa la protezione dell'infrastruttura e dei modelli da compromissioni, minacce o perdite, lo sviluppo di processi di gestione degli incidenti e il rilascio responsabile.

Metti in sicurezza l'infrastruttura

Applica buoni principi di sicurezza per le infrastrutture all'infrastruttura utilizzata in ogni parte del ciclo di vita del sistema. Applica controlli di accesso appropriati alle tue API, ai tuoi modelli e dati e alle loro pipeline di formazione ed elaborazione, sia in fase di ricerca e sviluppo che di implementazione. Ciò include un'adeguata segregazione degli ambienti che contengono codice o dati sensibili. Questo aiuterà anche a mitigare gli attacchi standard di sicurezza informatica che mirano a rubare un modello o a danneggiarne le prestazioni.

Proteggi il modello in maniera continua

Gli aggressori possono essere in grado di ricostruire la funzionalità di un modello[13] o i dati su cui è stato addestrato[14], accedendo a un modello direttamente (acquisendo i pesi del modello) o indirettamente (interrogando il modello tramite un'applicazione o un servizio). Gli aggressori possono anche manomettere i modelli, i dati o i prompt durante o dopo l'addestramento, rendendo il risultato non affidabile.

Puoi proteggere il modello e i dati dall'accesso diretto e indiretto, rispettivamente, mediante:

- l'implementazione delle migliori pratiche di sicurezza informatica standard;
- l'implementazione di controlli sull'interfaccia di interrogazione per rilevare e prevenire i tentativi di accesso, modifica ed esfiltrazione di informazioni riservate.

Per garantire che i sistemi di consumo possano convalidare i modelli, devi calcolare e condividere hash crittografici e/o firme dei file del modello (ad esempio, i pesi del modello) e dei set di dati (compresi i checkpoint) non appena il modello viene addestrato. Come sempre nella crittografia, una buona gestione delle chiavi è essenziale[15].

L'approccio alla mitigazione del rischio di riservatezza dipenderà notevolmente dal caso d'uso e dal modello di minaccia. Alcune applicazioni, ad esempio quelle che coinvolgono dati molto sensibili, possono richiedere garanzie teoriche che possono essere difficili o costose da applicare. Se opportuno, si possono usare tecnologie di miglioramento della privacy (come la privacy differenziale o la crittografia omomorfa) per esplorare o assicurare i livelli di rischio associati ai consumatori, agli utenti e agli aggressori che hanno accesso ai modelli e agli output.

Sviluppa procedure di gestione degli incidenti

L'inevitabilità di incidenti di sicurezza che colpiscono i sistemi di intelligenza artificiale si riflette nei piani di risposta, escalation e rimedio agli incidenti. I piani riflettono diversi scenari e vengono regolarmente rivalutati in base all'evoluzione del sistema e della ricerca in generale. Le risorse digitali critiche dell'azienda sono conservate in backup offline. Il personale specializzato è stato addestrato per valutare e affrontare gli incidenti legati all'IA. Fornisci log di controllo di alta qualità e altre funzioni o informazioni di sicurezza a clienti e utenti senza alcun costo aggiuntivo, per rendere possibili i loro processi di risposta agli incidenti.

Rilascia l'IA in modo responsabile

Rilascia i modelli, le applicazioni o i sistemi solo dopo averli sottoposti a un'adeguata ed efficace valutazione della sicurezza, come il benchmarking e il red teaming (oltre ad altri test che esulano dall'ambito di queste linee guida, come la sicurezza o l'equità), e comunica chiaramente agli utenti le limitazioni note o le potenziali modalità di guasto. I dettagli sulle librerie di test di sicurezza open-source sono riportati nella sezione Letture di approfondimento alla fine di questo documento.

Agevola gli utenti nel fare le cose giuste

Sii ben consapevole che ogni nuova impostazione o opzione di configurazione deve essere valutata in relazione ai vantaggi aziendali che ne derivano e agli eventuali rischi per la sicurezza che introduce. Idealmente, l'impostazione più sicura sarà integrata nel sistema come unica opzione. Quando la configurazione è necessaria, l'opzione predefinita dovrebbe essere ampiamente sicura contro le minacce più comuni (cioè, sicura per impostazione predefinita). Applica controlli per impedire l'uso o il dispiegamento del sistema in modi dannosi.

Fornisci agli utenti indicazioni sull'uso appropriato del modello o sistema, evidenziando anche i limiti e le potenziali modalità di guasto. Indica chiaramente agli utenti quali sono gli aspetti della sicurezza di cui sono responsabili e sii trasparente su dove (e come) i loro dati potrebbero essere utilizzati, consultati o archiviati (ad esempio, se vengono utilizzati per la riqualificazione dei modelli o se vengono esaminati da dipendenti o partner).

4. Esercizio e manutenzione sicuri

Questa sezione contiene linee guida che si applicano alla fase di **esercizio e manutenzione sicuri** del ciclo di vita dello sviluppo del sistema d'intelligenza artificiale. Questa sezione fornisce linee guida sulle azioni particolarmente rilevanti una volta che un sistema è stato implementato, tra cui la registrazione e il monitoraggio, la gestione degli aggiornamenti e la condivisione delle informazioni.

Monitora il comportamento del sistema

Misura gli output e le prestazioni del modello e del sistema in modo da poter osservare i cambiamenti improvvisi e graduali del comportamento che influiscono sulla sicurezza. Puoi tenere conto delle potenziali intrusioni e compromissioni e identificarle, nonché della naturale deriva dei dati.

Monitora gli input del sistema

In linea con i requisiti in materia di privacy e protezione dei dati, monitora e registra gli input al sistema (come le richieste di inferenza, le interrogazioni o i prompt) per consentire il rispetto degli obblighi di conformità, audit, indagini e rimedi in caso di compromissione o uso improprio. Ciò potrebbe includere il rilevamento esplicito di input fuori distribuzione e/o avversari, compresi quelli che mirano a sfruttare le fasi di preparazione dei dati (come il ritaglio e il ridimensionamento delle immagini).

Segui un approccio di sicurezza fin dalla progettazione per gli aggiornamenti

Includi aggiornamenti automatici di default in ogni prodotto e utilizza procedure di aggiornamento sicure e modulari per distribuirli. I processi di aggiornamento (compresi i regimi di test e valutazione) rispecchiano il fatto che le modifiche ai dati, ai modelli o alle richieste possono portare a cambiamenti nel comportamento del sistema (ad esempio, gestisci gli aggiornamenti principali come nuove versioni). Supporta gli utenti per valutare e rispondere alle modifiche del modello (ad esempio, fornendo l'accesso in anteprima e API con versioni).

Raccogli e condividi le lezioni apprese

Partecipa alle comunità di condivisione delle informazioni, collaborando con l'ecosistema globale dell'industria, del mondo accademico e dei governi per condividere le migliori pratiche, ove opportuno. Mantieni linee di comunicazione aperte per il feedback riguardante la sicurezza del sistema, sia internamente che esternamente alla tua organizzazione, compreso il consenso ai ricercatori di sicurezza per la ricerca e la segnalazione delle vulnerabilità. Quando è necessario, segnala i problemi a una comunità più ampia, ad esempio pubblicando bollettini che rispondono alle rivelazioni di vulnerabilità, compresa un'enumerazione dettagliata e completa delle vulnerabilità comuni. Intraprendi delle azioni per mitigare e risolvere i problemi in modo rapido e appropriato.

Letture di approfondimento

Sviluppo dell'intelligenza artificiale

Principles for the security of machine learning (Principi per la sicurezza dell'apprendimento automatico)
Guida dettagliata dell'NCSC sullo sviluppo, il dispiegamento o l'esercizio di un sistema con una componente di apprendimento automatico.

Secure by Design - Shifting the Balance of Cybersecurity Risk (Secure by Design - Lo spostamento dell'equilibrio del rischio di sicurezza informatica): Principles and Approaches for Secure by Design Software (Principi e approcci per la progettazione sicura di software)
Questa guida, redatta congiuntamente dal CISA, dall'NCSC e da altre agenzie, descrive come i produttori di sistemi software, compresa l'IA, debbano adottare misure per integrare la sicurezza nella fase di progettazione dello sviluppo del prodotto e fornire prodotti sicuri fin dall'inizio.

AI Security Concerns in a Nutshell (Pillole di problemi che riguardano la sicurezza dell'intelligenza artificiale)
Realizzato dall'Ufficio federale tedesco per la sicurezza informatica (BSI), questo documento fornisce un'introduzione ai possibili attacchi ai sistemi di apprendimento automatico e alle potenziali difese contro tali attacchi.

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems (Principi guida internazionali per le organizzazioni che sviluppano sistemi di IA avanzati del Processo di Hiroshima) e Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (Codice internazionale di condotta per le organizzazioni che sviluppano sistemi avanzati di IA del Processo di Hiroshima)
Questi documenti, prodotti nell'ambito del Processo di Hiroshima sull'IA del G7, forniscono una guida per le organizzazioni che sviluppano i sistemi di IA più avanzati, compresi i modelli di base più avanzati e i sistemi di IA generativi, con l'obiettivo di promuovere un'IA sicura, protetta e affidabile a livello mondiale.

AI Verify (Verifica IA)
Quadro di test per la governance dell'IA di Singapore e kit di strumenti software che convalidano le prestazioni dei sistemi IA rispetto a una serie di principi riconosciuti a livello internazionale attraverso test standardizzati.

Multilayer Framework for Good Cybersecurity Practices for AI (Quadro multistrato per le buone pratiche di sicurezza informatica per l'IA) — ENISA (europa.eu)
Un quadro di riferimento per guidare le autorità nazionali competenti e gli stakeholder dell'IA in materia di misure da seguire per mettere in sicurezza i propri sistemi, operazioni e processi di IA.

ISO 5338: AI system life cycle processes (Processi del ciclo di vita dei sistemi di IA - in corso di revisione)
Un insieme di processi e concetti associati per descrivere il ciclo di vita dei sistemi di IA basati sull'apprendimento automatico e sui sistemi euristici.

AI Cloud Service Compliance Criteria Catalogue (Catalogo dei criteri di conformità dei servizi cloud IA) (AIC4)
Il Catalogo dei Criteri di Conformità dei Servizi Cloud IA di BSI fornisce criteri specifici per l'IA, che consentono di valutare la sicurezza di un servizio di IA durante il suo ciclo di vita.

NIST IR 8269 (bozza) A Taxonomy and Terminology of Adversarial Machine Learning (Una tassonomia e una terminologia dell'apprendimento automatico antagonista)
Un insieme di processi e concetti associati per descrivere il ciclo di vita dei sistemi di IA basati sull'apprendimento automatico e sui sistemi euristici.

MITRE ATLAS
Base di conoscenza delle tattiche, delle tecniche e dei casi di studio antagonisti per i sistemi di apprendimento automatico, modellata e collegata al framework MITRE ATT&CK.

An Overview of Catastrophic AI Risks (Una panoramica dei rischi catastrofici dell'IA) (2023)
Realizzato dal Center for AI Safety, questo documento illustra le aree di rischio poste dall'IA.

Modelli linguistici di grandi dimensioni: Opportunities and Risks for Industry and Authorities (Opportunità e rischi per l'industria e le autorità)
Documento prodotto da BSI per le aziende, le autorità e gli sviluppatori che vogliono saperne di più sulle opportunità e sui rischi dello sviluppo, del dispiegamento e/o dell'utilizzo dei modelli linguistici di grandi dimensioni.

Tra i progetti open-source che aiutano gli utenti a testare i modelli di IA in termini di sicurezza vi sono:

- Adversarial Robustness Toolbox (IBM)
- CleverHans (Università di Toronto)
- TextAttack (Università della Virginia)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft)
- AI Verify (Infocomm Media Development Authority, Singapore)

Sicurezza informatica

CISA's Cybersecurity Performance Goals (Obiettivi di prestazione della sicurezza informatica della CISA)
Un insieme comune di protezioni che tutte le entità di infrastrutture critiche dovrebbero implementare per ridurre significativamente la probabilità e l'impatto dei rischi noti e delle tecniche antagonistiche.

NCSC CAF Framework (Quadro di riferimento CAF NCSC)
Il Cyber Assessment Framework (CAF) fornisce una guida alle organizzazioni responsabili di servizi e attività di vitale importanza.

MITRE's Supply Chain Security Framework (Quadro di riferimento per la sicurezza della catena di approvvigionamento del MITRE)
Un quadro di riferimento per la valutazione dei fornitori e dei prestatori di servizi all'interno della catena di fornitura.

Gestione del rischio

NIST AI Risk Management Framework (AI RMF) (Quadro di riferimento per la gestione del rischio dell'intelligenza artificiale del NIST)
L'AI RMF illustra come gestire i rischi socio-tecnici per gli individui, le organizzazioni e la società associati all'IA.

ISO 27001: Sicurezza delle informazioni, sicurezza informatica e protezione della privacy
Questo standard fornisce alle organizzazioni una guida per la creazione, l'implementazione e il mantenimento di un sistema di gestione della sicurezza delle informazioni.

ISO 31000: Gestione del rischio
Uno standard internazionale che fornisce alle organizzazioni linee guida e principi per la gestione del rischio all'interno delle organizzazioni.

Guida NCSC alla gestione del rischio (NCSC Risk Management Guidance)
Questa guida aiuta i professionisti del rischio in materia di sicurezza informatica a comprendere e gestire meglio i rischi di sicurezza informatica che interessano le loro organizzazioni.

Note

1. Qui definito come una persona, un'autorità pubblica, un'agenzia o un altro organismo che sviluppa un sistema di IA (o che commissiona lo sviluppo un sistema di IA) e lo immette sul mercato o lo mette in servizio con il proprio nome o marchio
2. Per maggiori informazioni sulla sicurezza fin dalla progettazione (secure by design), è possibile consultare la pagina web di CISA Secure by Design e la guida Shifting the Balance of Cybersecurity Risk (Lo spostamento dell'equilibrio del rischio di sicurezza informatica): Principi e approcci per la sicurezza del software fin dalla progettazione.
3. Rispetto agli approcci di IA non relativi all'apprendimento automatico come i sistemi basati su regole.
4. Il CEPS descrive sette diversi tipi di interazione per lo sviluppo dell'IA nella sua pubblicazione "Reconciling the AI Value Chain with the EU's Artificial Intelligence Act" (Riconciliare la catena del valore dell'IA con la legge sull'intelligenza artificiale dell'UE).
5. ISO/IEC 22989:2022(en) lo definisce come "un elemento funzionale che costruisce un sistema di IA".
6. Il NIST ha il compito di produrre linee guida (e di intraprendere altre azioni) per promuovere lo sviluppo e l'uso sicuro, protetto e affidabile dell'intelligenza artificiale (IA). È possibile consultare le responsabilità del NIST ai sensi dell'Ordine Esecutivo (Executive Order) del 30 ottobre 2023.
7. Maggiori informazioni sulla modellazione delle minacce sono disponibili presso la OWASP Foundation.
8. Si può inoltre consultare Adversarial Machine Learning 101 (Apprendimento automatico antagonista di base) di MITRE ATLAS.
9. GitHub: RCE PoC for Tensorflow using a malicious Lambda layer (RCE PoC per Tensorflow che utilizza un livello Lambda dannoso).
10. SLSA: Safeguarding artifact integrity across any software supply chain (Salvaguardare l'integrità degli artefatti in tutta la catena di approvvigionamento del software).
11. METI (Ministero giapponese dell'Economia, del Commercio e dell'Industria, 2023), Guide of Introduction of Software Bill of Materials (SBOM) for Software Management (Guida all'introduzione della distinta base del software (SBOM) per la gestione del software).
12. Ricerca Google: Apprendimento automatico: The High Interest Credit Card of Technical Debt (La carta di credito ad alto interesse del debito tecnico).
13. Tramèr et al 2016, Stealing Machine Learning Models via Prediction APIs (Furto di modelli di apprendimento automatico tramite API di predizione).
14. Boenisch, 2020, Attacks against Machine Learning Privacy - Part 1 (Attacchi alla privacy dell'apprendimento automatico - Parte 1): Model Inversion Attacks with the IBM-ART Framework (Modelli di attacchi di inversione con il framework IBM-ART).
15. Centro nazionale per la sicurezza informatica, 2020, Design and build a privately hosted Public Key Infrastructure (Progettare e costruire un'infrastruttura a chiave pubblica ospitata privatamente).