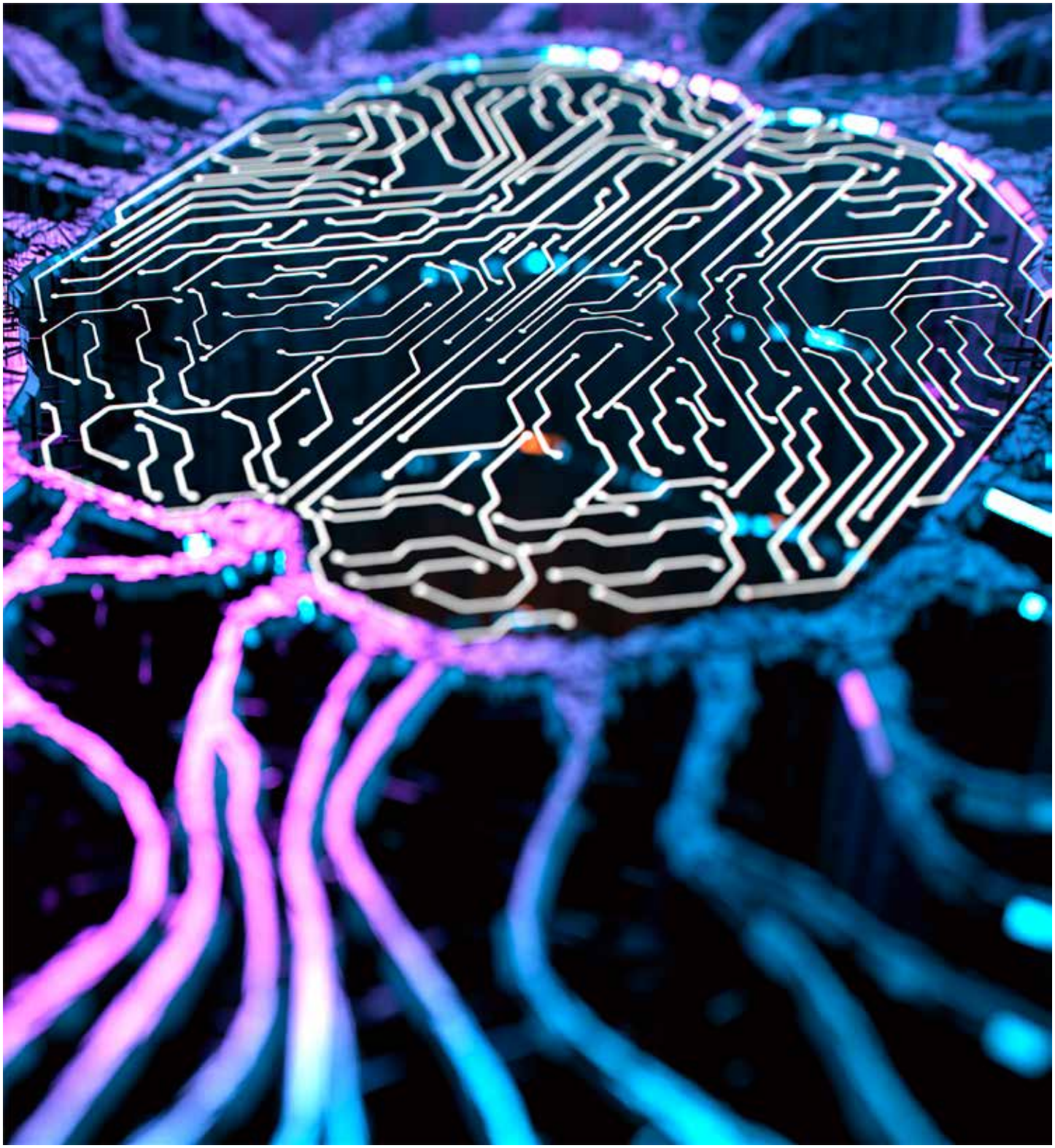
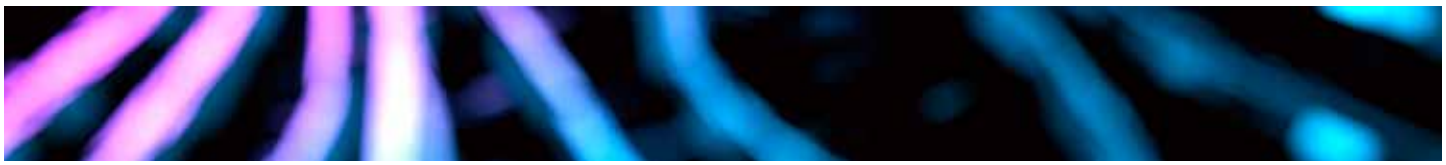
Smjernice za razvoj sigurnog sustava umjetne inteligencije





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

O ovom dokumentu

Ovaj dokument objavili su Nacionalni centar za kibernetičku sigurnost Ujedinjenog Kraljevstva (NCSC), Agencija za kibernetičku sigurnost i infrastrukturnu sigurnost SAD-a (CISA) i sljedeći međunarodni partneri:

- Agencija za nacionalnu sigurnost (NSA)
- Federalni istražni biro (FBI)
- Australski centar za kibernetičku sigurnost Australske uprave za signale (ACSC)
- Kanadski centar za kibernetičku sigurnost (CCCS)
- Nacionalni centar za kibernetičku sigurnost Novog Zelanda (NCSC-NZ)
- CSIRT vlade Čilea
- Češka nacionalna agencija za kibernetičku i informacijsku sigurnost (NUKIB)
- Uprava za informacijski sustav Estonije (RIA) i Nacionalni centar za kibernetičku sigurnost Estonije (NCSC-EE)
- Francuska agencija za kibernetičku sigurnost (ANSSI)
- Njemački savezni ured za sigurnost informacija (BSI)
- Izraelska nacionalna kibernetička uprava (INCD)
- Talijanska nacionalna agencija za kibernetičku sigurnost (ACN)
- Japanski nacionalni centar za spremnost na incidente i strategiju kibernetičke sigurnosti (NISC)
- Japansko tajništvo za znanost, tehnologiju i inovacijsku politiku, Ured kabineta
- Nigerijska nacionalna agencija za razvoj informacijske tehnologije (NITDA)
- Norveški nacionalni centar za kibernetičku sigurnost (NCSC-NO)
- Poljsko ministarstvo digitalnih poslova
- Nacionalni istraživački institut Poljske (NASK)
- Nacionalna obavještajna služba Republike Koreje (NIS)
- Agencija za kibernetičku sigurnost Singapura (CSA)

Priznanja

Sljedeće organizacije pridonijele su razvoju ovih smjernica:

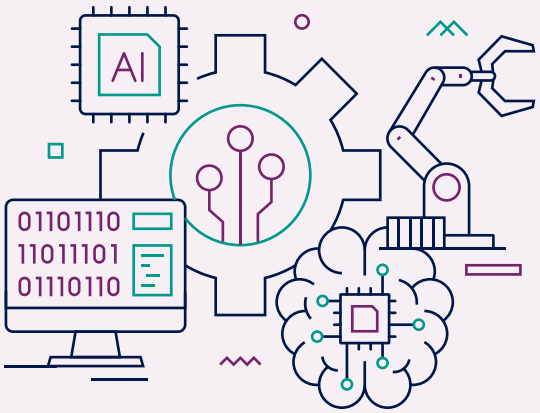
- Institut Alan Turing
- Anthropic
- Databricks
- Centar za sigurnost i nove tehnologije Sveučilišta Georgetown
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Institut za softversko inženjerstvo na Sveučilištu Carnegie Mellon
- Stanford centar za sigurnost umjetne inteligencije
- Stanford program o geopolitici, tehnologiji i upravljanju

Izjava o odricanju odgovornosti

Informacije u ovom dokumentu su "onakve kakve su" ih dostavili u NCSC-u i drugim autorskim organizacijama koje neće biti odgovorne za bilo kakav gubitak, ozljedu ili štetu bilo koje vrste uzrokovane njegovom uporabom, osim ako to zahtijeva zakon. Informacije u ovom dokumentu ne predstavljaju niti impliciraju podršku ili preporuku bilo koje organizacije, proizvoda ili usluge treće strane od strane NCSC-a i autorskih agencija. Poveznice i reference na web stranice i materijale trećih strana su imformacijske naravi i njima se ne odobravaju ili preporučuju navedeni resursi u odnosu na druge.

Ovaj je dokument dostupan na temelju TLP:CLEAR (https://www.first.org/tlp/).

Sadržaj



Izvršni sažetak

Ovaj dokument preporučuje smjernice za pružatelje svih sustava koji koriste umjetnu inteligenciju (AI), bilo da su ti sustavi stvoreni od nule ili izgrađeni na temelju alata i usluga koje pružaju drugi. Primjena ovih smjernica pomoći će pružateljima da izgrade sustave umjetne inteligencije koji funkcioniraju kako je predviđeno, dostupni su kada je potrebno i rade bez otkrivanja osjetljivih podataka neovlaštenim stranama.

Ovaj je dokument prvenstveno namijenjen pružateljima sustava umjetne inteligencije koji koriste organizacije modele ili vanjska sučelja za programiranje aplikacija (API). Pozivamo **sve** dionike (uključujući podatkovne znanstvenike, programere, menadžere, donositelje odluka i vlasnike rizika) da pročitaju ove smjernice kako bi im pomogli u donošenju temeljenih na dostupnim informacijama o **dizajnu**,**razvoju**, **postaavljanju** i **radu** njihovih sustava umjetne inteligencije.

O smjernicama

Sustavi umjetne inteligencije imaju potencijal donijeti mnoštvo dobrobiti društvu. Međutim, za potpuno ostvaranje potencijala umjetne inteligencije, mora ga se razviti, postaviti i njime upravljati na siguran i odgovoran način.

Sustavi umjetne inteligencije podložni su novim sigurnosnim ranjivostima koje je potrebno razmotriti uz standardne prijetnje kibernetičkoj sigurnosti. Kada je tempo razvoja visok – kao što je slučaj s umjetnom inteligencijom – sigurnost često može biti drugorazredna stvar. Sigurnost mora biti temeljni zahtjev, ne samo u fazi razvoja, već tijekom čitavog životnog ciklusa sustava.

Iz tog razloga, smjernice su podijeljene u četiri ključna područja unutar životnog ciklusa razvoja sustava umjetne inteligenije: **siguran dizajn**, **sigurni razvoj**, **sigurna implementacija**, i **siguran rad i održavanje**. Za svaki odjeljak predlažemo razmatranja i ublažavanja koja će pomoći u smanjenju cjelokupnog rizika za proces razvoja organizacijskog sustava umjetne inteligencije.

1. **Siguran dizajn**
   Ovaj odjeljak sadrži smjernice koje se odnose na fazu kreiranja životnog ciklusa sustava umjetne inteligecije. Ono obuhvaća razumijevanje rizika i modeliranje prijetnji, kao i specifične teme i kompromise koje treba uzeti u obzir prilikom kreoranja sustava i modela.

2. **Siguran razvoj**
   Ovaj odjeljak sadrži smjernice koje se odnose na razvojnu fazu životnog ciklusa razvoja sustava umjetne inteligencije, uključujući sigurnost opskrbnog lanca, dokumentaciju te upravljanje imovinom i tehničkim dugom.

3. **Sigurna implementacija**
   Ovaj odjeljak sadrži smjernice koje se primjenjuju na fazu implementacije životnog ciklusa razvoja sustava umjetne inteligencije, uključujući zaštitu infrastrukture i modela od ugrožavanja, prijetnje ili gubitka, razvoj procesa upravljanja incidentima i odgovorno oslobađanje.

4. **Siguran rad i održavanje**
   Ovaj odjeljak sadrži smjernice koje se odnose na siguran rad i fazu održavanja u životnom ciklusu razvoja AI sustava. Ono pruža smjernice o radnjama koje su posebno relevantne nakon što je sustav postavljen, uključujući bilježenje i praćenje, upravljanje ažuriranjem i dijeljenje informacija.

Smjernice slijede pristup temeljen na "zadanim postavkama" sigurnosti te su ujedno usklađene s praksama definiranim u NCSC-ovim smjernicama za siguran razvoj i implementaciju, NIST-ovim okvirom za siguran razvoj softvera i 'načelu sigurnog dizajna' koje su objavili CISA, NCSC i međunarodne agencije za kibernetiku. Oni daju prioritet:

- preuzimanju vlasništva nad sigurnosnim rezultatima za klijente
- prihvaćanju radikalne transparentnosti i odgovornosti
- izgradnji organizacijske strukture i vodstva s integriranom sigurnošću kao ključnim poslovnim prioritetom

Uvod

Sustavi umjetne inteligencije (AI) potencijalno mogu donijeti mnoštvo dobrobiti društvu. Međutim, kako bi se mogućnosti umjetne inteligencije u potpunosti iskoristile, mora se razviti, primijeniti i njome upravljati na siguran i odgovoran način. Kibernetička sigurnost nužan je preduvjet za sigurnost, otpornost, privatnost, pravednost, učinkovitost i pouzdanost sustava umjetne inteligencije.

Međutim, sustavi umjetne inteligencije podložni su novim sigurnosnim ranjivostima koje je potrebno razmotriti zajedno sa standardnim prijetnjama kibernetičkoj sigurnosti. Kad je tempo razvoja visok – kao što je slučaj s umjetnom inteligencijom – sigurnost često može biti u drugom planu. Sigurnost mora biti temeljni zahtjev, ne samo u fazi razvoja, već tijekom čitavog životnog ciklusa sustava.

Ovaj dokument preporučuje smjernice za pružatelje[1] svih sustava koji koriste umjetnu inteligenciju, bilo da su ti sustavi stvoreni od nule ili izgrađeni na temelju alata i usluga koje pružaju drugi. Primjena ovih smjernica pomoći će pružateljima da izgrade sustave umjetne inteligencije koji funkcioniraju kako je predviđeno, koji su dostupni kada je to potrebno i rade bez otkrivanja osjetljivih podataka neovlaštenim stranama.

Ove bi se smjernice trebale razmotriti zajedno s uspostavljenom kibernetičkom sigurnošću, upravljanjem rizikom i najboljom praksom odgovora na incidente. Konkretno, pozivamo pružatelje da slijede načela "sigurnosti po dizajnu"[2] koja su razvili Agencija za kibernetičku sigurnost i sigurnost američke infrastrukture (CISA), Nacionalni centar za kibernetičku sigurnost Ujedinjenog Kraljevstva (NCSC) i svi naši međunarodni partneri. Ova načela daju prioritet:

- preuzimanju vlasništva nad sigurnosnim rezultatima za klijente
- prihvaćanju radikalne transparentnosti i odgovornosti
- izgradnji organizacijske strukture i vodstva s integriranom sigurnošću kao ključnim poslovnim prioritetom.

Slijeđenje načela 'sigurnosti po dizajnu' zahtijeva značajne resurse tijekom životnog ciklusa sustava. To znači da programeri moraju ulagati u davanje prioriteta **značajkama**, **mehanizmima** i **implementaciji** alata koji štite kupce na svakom sloju dizajna sustava i u svim fazama životnog ciklusa razvoja. Time će se spriječiti potreba za daljnjim skupim redizajnima te će se ujedno zaštititi klijenti i njihovi podaci u bliskoj budućnosti.

Zašto je sigurnost umjetne inteligencije drukčija?

U ovom dokumentu koristimo 'AI' isključivo za označavanje aplikacija strojnog učenja (ML) [3]. Sve vrste ML-a su obuhvaćene. ML aplikacije definiramo kao aplikacije koje:

- uključuju softverske komponente (modele) koji omogućuju računalima da prepoznaju i daju kontekst obrascima u podacima bez izričite potrebe da pravila programiraju ljudi
- generiraju predviđanja, preporuke ili odluke na temelju statističkog zaključivanja

Kao i postojeće prijetnje kibernetičkoj sigurnosti, sustavi umjetne inteligencije podložni su novim vrstama ranjivosti. Izraz suparničko strojno učenje - 'adversarial machine learning' (AML) koristi se za opisivanje iskorištavanja temeljnih ranjivosti u komponentama ML-a, uključujući hardver, softver, tijekove rada i lance opskrbe. AML omogućuje napadačima da izazovu neželjena ponašanja u ML sustavima koja mogu uključivati:

- utjecanje na izvedbu klasifikacije ili regresije modela
- dopuštanje korisnicima izvođenje neovlaštenih radnji
- izdvajanje osjetljivih informacija o modelu

Postoji mnogo načina za postizanje ovih učinaka, kao što su brzi napadi ubrizgavanjem u domenu velikog jezičnog modela (LLM) ili namjerno oštećenje podataka o obuci ili povratnih informacija korisnika (poznato kao 'trovanje podataka').

Tko bi trebao pročitati ovaj dokument?

Ovaj je dokument prvenstveno namijenjen pružateljima AI sustava, bilo da se temelje na modelima koje hostira organizacija ili koriste vanjska programska sučelja aplikacija (API). Međutim, potičemo **sve** dionike (uključujući znanstvenike podataka, programere, upravitelje, donositelje odluka i vlasnike rizika) da pročitaju ove smjernice kako bi im pomogle u donošenju informiranih odluka o **dizajnu**, **uvođenju** i **radu** njihovih sustava strojnog učenja umjetne inteligencije.

Ipak, neće sve smjernice biti izravno primjenjive na sve organizacije. Razina sofisticiranosti i metode napada razlikovat će se ovisno o protivniku koji cilja na sustav umjetne inteligencije, tako da bi smjernice trebalo razmotriti uz slučajeve upotrebe i profil prijetnji vaše organizacije.

Tko je odgovoran za razvoj sigurne umjetne inteligencije?

U modernim opskrbnim lancima umjetne inteligencije često postoji mnoštvo sudionika. Jednostavan pristup pretpostavlja dva entiteta:

- ‘pružatelj’ koji je odgovoran za čuvanje podataka, algoritamski razvoj, dizajn, implementaciju i održavanje
- 'korisnik', koji daje ulaze i prima izlaze

Dok se ovaj pristup pružatelj-korisnik koristi u mnogim aplikacijama, on postaje sve neuobičajeniji[4], budući da pružatelji mogu nastojati ugraditi u svoje vlastite sustave softver, podatke, modele i/ili udaljene usluge koje pružaju treće strane. Ovi složeni opskrbni lanci otežavaju krajnjim korisnicima da utvrde odgovornost za sigurno korištenje umjetne inteligencije.

Korisnici (bilo 'krajnji korisnici' ili pružatelji usluga koji uključuju vanjsku AI komponentu[5]) obično nemaju dovoljno vidljivosti i/ili stručnosti da u potpunosti razumiju, procijene ili riješe rizike povezane sa sustavima koje koriste. Kao takvi, u skladu s načelima integrirane sigurnosti ', **pružatelji komponenti umjetne inteligencije trebali bi preuzeti odgovornost za sigurnosne rezultate korisnika u nižim dijelovima opskrbnog lanca.**

Pružatelji bi trebali implementirati sigurnosne kontrole i mjere ublažavanja gdje je to moguće unutar svojih modela, i/ili sustava, a gdje se koriste postavke, implementirati najsigurniju opciju kao zadanu. Tamo gdje se rizici ne mogu umanjiti, pružatelj bi trebao biti odgovoran za:

- informiranje korisnika u nižim dijelovima opskrbnog lanca o rizicima koje oni i (ako je primjenjivo) njihovi korisnici prihvaćaju
- savjetovanje o sigurnoj upotrebi komponente

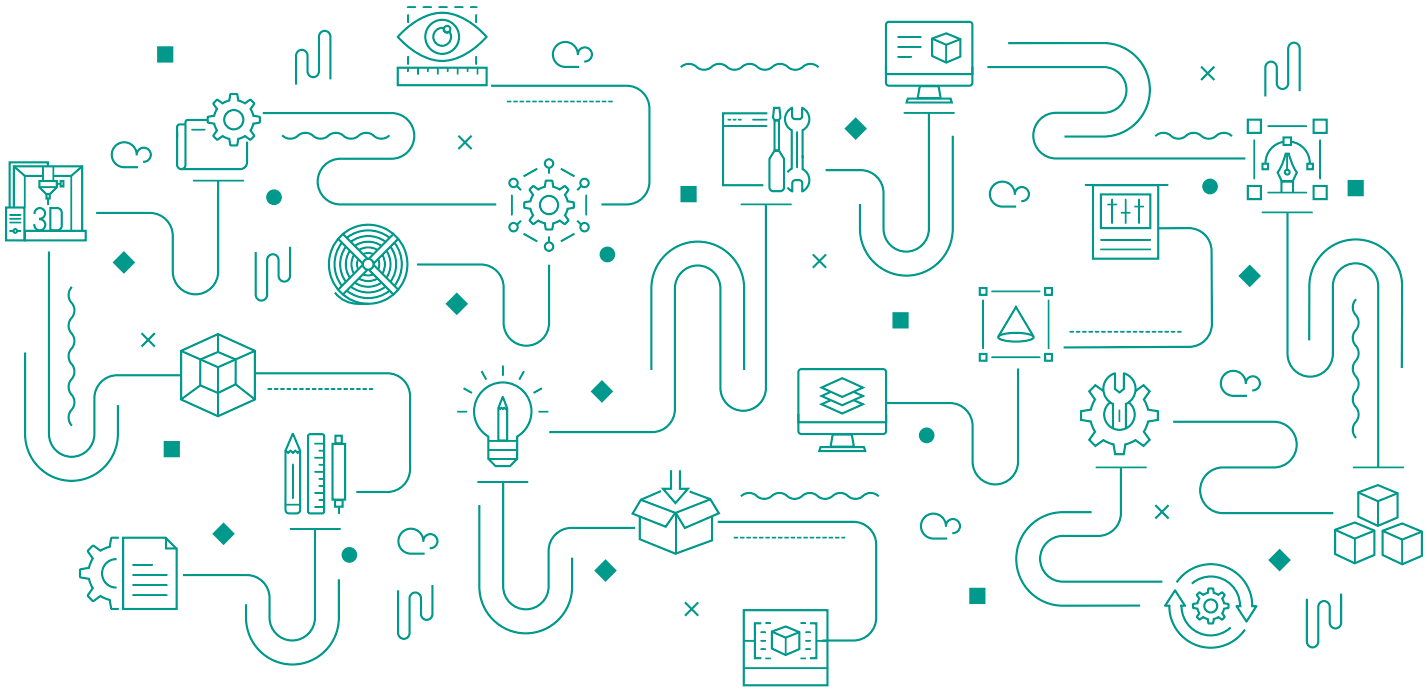
U slučaju kompromitacije sustava koja može prouzrokovati opipljivu ili široko rasprostranjenu fizičku štetu ili štetu po nečiji ugled, značajan gubitak poslovnih operacija, curenje osjetljivih ili povjerljivih informacija i/ili pravnih implikacija, rizike kibernetičke sigurnosti umjetne inteligencije treba tretirati kao **kritične.**

Smjernice za siguran razvoj sustava umjetne inteligencije

Smjernice su podijeljene u četiri ključna područja unutar životnog ciklusa razvoja AI sustava: **siguran dizajn**, **siguran razvoj**, **sigurna implementacija** te **siguran rad i održavanje** . Za svako područje predlažemo razmatranja i ublažavanje rizika što će pomoći u smanjenju cjelokupnog rizika u procesu razvoja organizacijskog sustava umjetne inteligencije.

Smjernice navedene u ovom dokumentu usko su usklađene s praksama životnog ciklusa razvoja softvera definiranim u:

- NCSC-ovim smjernicama za siguran razvoj i implementaciju
- Okviru za razvoj sigurnog softvera (SSDF)[6]pri Nacionalnom institutu za standarde i tehnologiju (NIST)



1. Siguran dizajn

Ovo područje sadrži smjernice koje se odnose na fazu **kreiranja** životnog ciklusa sustava umjetne inteligencije. On pokriva razumijevanje rizika i modeliranje prijetnji, kao i specifične teme i kompromise koje treba uzeti u obzir prilikom izrade sustava i modela.

Podignite svijest osoblja o prijetnjama i rizicima

Vlasnici sustava i viši rukovoditelji razumiju prijetnje sigurnosti sustavu umjetne inteligencije i njihova ublažavanja. Vaši znanstvenici i programeri podataka održavaju svijest o relevantnim sigurnosnim prijetnjama i načinima kvarova te pomažu vlasnicima rizika u donošenju promišljenih odluka. Korisnicima dajete smjernice o jedinstvenim sigurnosnim rizicima s kojima se suočavaju AI sustavi (na primjer, kao dio standardne InfoSec obuke) i obučavate programere o sigurnim tehnikama kodiranja te sigurnim i odgovornim praksama umjetne inteligencije.

Modelirajte prijetnje vašem sustavu

U sklopu procesa upravljanja rizikom, primjenjujete holistički postupak za procjenu prijetnji vašem sustavu, što uključuje razumijevanje potencijalnih utjecaja na sustav, korisnike, organizacije i šire društvo ako je AI komponenta ugrožena ili se ponaša na neočekivan način[7] Ovaj postupak uključuje procjenu utjecaja prijetnji specifičnih za umjetnu inteligenciju[8] i dokumentiranje vaših odluka.

Shvaćate da osjetljivost i vrste podataka koji se koriste u vašem sustavu mogu utjecati na njegovu vrijednost kao mete napadaču. Vaša bi procjena trebala uzeti u obzir povećanje određenih prijetnji dok se sustavi umjetne inteligencije sve više smatraju ciljevima visoke vrijednosti, a kako sama umjetna inteligencija omogućuje nove, automatizirane vektore napada.

Dizajnirajte svoj sustav za sigurnost, kao i funkcionalnost i performanse

Uvjereni ste da se zadatak koji je pred vama najprikladnije rješava pomoću umjetne inteligencije. Nakon što ste to utvrdili, procjenjujete prikladnost svojih izbora dizajna specifičnih za umjetnu inteligenciju. Uzimate u obzir svoj model prijetnje i povezane sigurnosne mjere ublažavanja zajedno s funkcionalnošću, korisničkim iskustvom, okruženjem za implementaciju, izvedbom, osiguranjem, nadzorom, etičkim i pravnim zahtjevima, između ostalih razmatranja. Na primjer:

- uzimate u obzir sigurnost opskrbnog lanca prilikom odabira unutarnjeg razvoja ili korištenja vanjskih komponenti primjerice:
    - vaš izbor za obuku novog modela, korištenje postojećeg modela (sa ili bez finog podešavanja) ili pristup modelu putem vanjskog API-ja prikladan je vašim zahtjevima
    - vaš izbor za rad s vanjskim pružateljem modela uključuje procjenu dubinske analize sigurnosnog položaja tog pružatelja
    - ako koristite vanjsku biblioteku, dovršite procjenu dubinske analize (na primjer, kako biste osigurali da biblioteka ima kontrole koje sprječavaju sustav da učitava nepouzdane modele bez da se odmah izlažu proizvoljnom izvršavanju koda[9])
    - implementirate skeniranje i izolaciju/sandboxing prilikom uvoza modela treće strane ili serijalizirane težine, koje bi trebalo tretirati kao od nepouzdane kod treće strane, a koje bi moglo omogućiti daljinsko izvršavanje koda

  - ako koristite vanjske API-je, primjenjujete odgovarajuće kontrole na podatke koji se mogu slati uslugama izvan kontrole vaše organizacije, kao što je zahtijevanje od korisnika da se prijave i potvrde prije slanja potencijalno osjetljivih informacija
  - primjenjujte odgovarajuće provjere i čišćenje podataka i unosa; to uključuje uključivanje povratnih informacija korisnika ili podataka o kontinuiranom učenju u vaš model, shvaćajući da podaci o obuci definiraju ponašanje sustava u koji

- integrirate razvoj softverskog sustava umjetne inteligencije u postojeće najbolje prakse sigurnog razvoja i operacija; svi elementi AI sustava napisani su u odgovarajućim okruženjima koristeći prakse kodiranja i jezike koji smanjuju ili eliminiraju poznate klase ranjivosti gdje je to moguće

- ako AI komponente trebaju pokrenuti radnje, na primjer promijeniti datoteke ili preusmjeriti izlaze na vanjske sustave, treba primijeniti odgovarajuća ograničenja na moguće radnje (ovo uključuje vanjske AI sustave i one sustave koji ne spadaju u sustave umjetne inteligenecije, no koji imaju integriranu zaštitu od kvarova ukoliko je to potrebno)

- odluke o interakciji s korisnikom temeljene su na rizicima specifičnim za AI, na primjer:

  - vaš sustav korisnicima pruža korisne rezultate bez otkrivanja nepotrebnih razina detalja potencijalnom napadaču
  - u slučaju potrebe, vaš sustav osigurava učinkovite zaštitne ograde oko izlaza modela
  - ako nudite API vanjskim kupcima ili suradnicima, primjenjujete odgovarajuće kontrole koje ublažavaju napade na AI sustav putem API
  - prema zadanim postavkama integrirate najsigurnije postavke u sustav
  - primjenjujete načela najmanjih privilegija kako biste ograničili pristup funkcionalnosti sustava
  - korisnicima objašnjavate riskantnije mogućnosti i tražite od korisnika da se odluče za njihovu upotrebu; priopćavate zabranjene slučajeve upotrebe i, gdje je to moguće, informirate korisnike o alternativnim rješenjima

Razmotrite sigurnosne prednosti i kompromise pri odabiru svog AI modela

Vaš izbor modela umjetne inteligencije uključivat će balansiranje niza zahtjeva. To uključuje izbor arhitekture modela, konfiguracije, podatke za obuku, algoritme za obuku i hiperparametre. Vaše odluke temeljene su na vašem modelu prijetnje i one se, s napretkom istraživanja sigurnosti sustava umjetne inteligencije i razumijevanja prijetnji, redovito podvrgavaju procjenama.

Prilikom odabira modela, vaša razmatranja će vjerojatno uključivati, ali nisu ograničena na:

- složenost modela koji koristite, odnosno odabranu arhitekturu i broj parametara; odabrana arhitektura vašeg modela i broj parametara će, između ostalih čimbenika, utjecati na to koliko podataka za obuku zahtijeva kao i na otpornost modela na promjene u ulaznim podacima prilikom korištenja

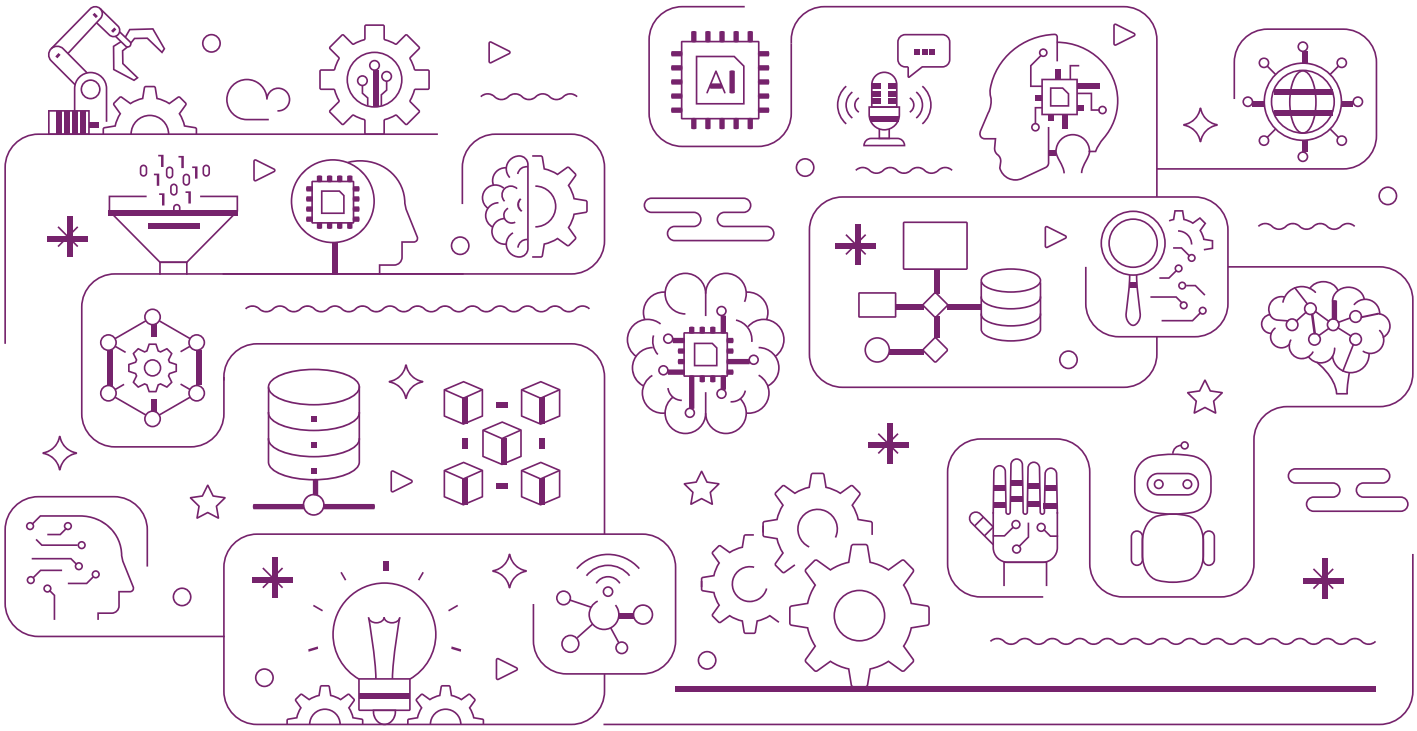
- prikladnost modela za vaš slučaj upotrebe i/ili izvedivost njegove prilagodbe vašim specifičnim potrebama (na primjer finim podešavanjem)

- sposobnost usklađivanja, tumačenja i objašnjenja rezultata vašeg modela (na primjer za otklanjanje pogrešaka, reviziju ili usklađenost s propisima); možda postoje prednosti korištenja jednostavnijih, transparentnijih modela u odnosu na velike i složene modele koje je teže interpretirati

- karakteristike skupova podataka za obuku, uključujući veličinu, cjelovitost, kvalitetu, osjetljivost, dob, relevantnost i raznolikost

- vrijednost korištenja učvršćivanja modela (kao što je 'adversarial training'), regulacija i/ili tehnika za poboljšanje privatnosti
- podrijetlo i opskrbni lanac komponenti uključujući model ili temeljni model, podatke o obuci i povezane alate

Za više informacija o tome koliko ovih čimbenika utječe na sigurnosne ishode, pogledajte NCSC-ova 'Načela za sigurnost strojnog učenja', posebno Dizajn za sigurnost (arhitektura modela).



2. Siguran razvoj

Ovaj odjeljak sadrži smjernice koje se primjenjuju na fazu **razvoja** životnog ciklusa razvoja sustava umjetne inteligencije, uključujući sigurnost opskrbnog lanca, dokumentaciju te upravljanje imovinom i tehničkim dugom.

Zaštitite svoj lanac opskrbe

Vi procjenjujete i nadzirete sigurnost svojih opskrbnih lanaca umjetne inteligencije tijekom životnog ciklusa sustava i zahtijevate od dobavljača da se pridržavaju istih standarda koje vaša vlastita organizacija primjenjuje na drugi softver. Ako se dobavljači ne mogu pridržavati standarda vaše organizacije, postupate u skladu s postojećim politikama upravljanja rizikom.

Ako se ne proizvodi interno, stječete i održavate dobro osigurane i dobro dokumentirane hardverske i softverske komponente (na primjer, modele, podatke, softverske biblioteke, module, međuprograme, okvire i vanjske API-je) iz provjerenih komercijalnih, otvorenih kodova, i drugih programera trećih strana kako biste osigurali snažnu pouzdanost vaših sustava.

Spremni ste prijeći na alternativna rješenja za kritične sustave ako sigurnosni kriteriji nisu zadovoljeni. Koristite resurse kao što su NCSC Smjernice za lanac opskrbe i okvire kao što su Razine lanca opskrbe za softverske artefakte (SLSA)[10] za praćenje atestiranja lanca opskrbe i životnih ciklusa razvoja softvera.

Identificirajte, pratite i zaštitite svoju imovinu

Razumijete vrijednost vaše imovine povezane s umjetnom inteligencijom za vašu organizaciju, uključujući modele, podatke (uključujući povratne informacije korisnika), upute, softver, dokumentaciju, zapisnike i procjene (uključujući informacije o potencijalno nesigurnim mogućnostima i načinima kvarova), prepoznajući gdje oni predstavljaju značajna ulaganja i gdje se pristup njima omogućuje napadaču. Zapisnike tretirate kao osjetljive podatke i implementirate kontrole za zaštitu njihove povjerljivosti, integriteta i dostupnosti.

Znate gdje se nalazi vaša imovina te ste procijenili i prihvatili sve povezane rizike. Imate procese i alate za praćenje, provjeru autentičnosti, kontrolu verzija i osiguranje svoje imovine, te ih možete vratiti u poznato dobro stanje u slučaju kompromitacije.

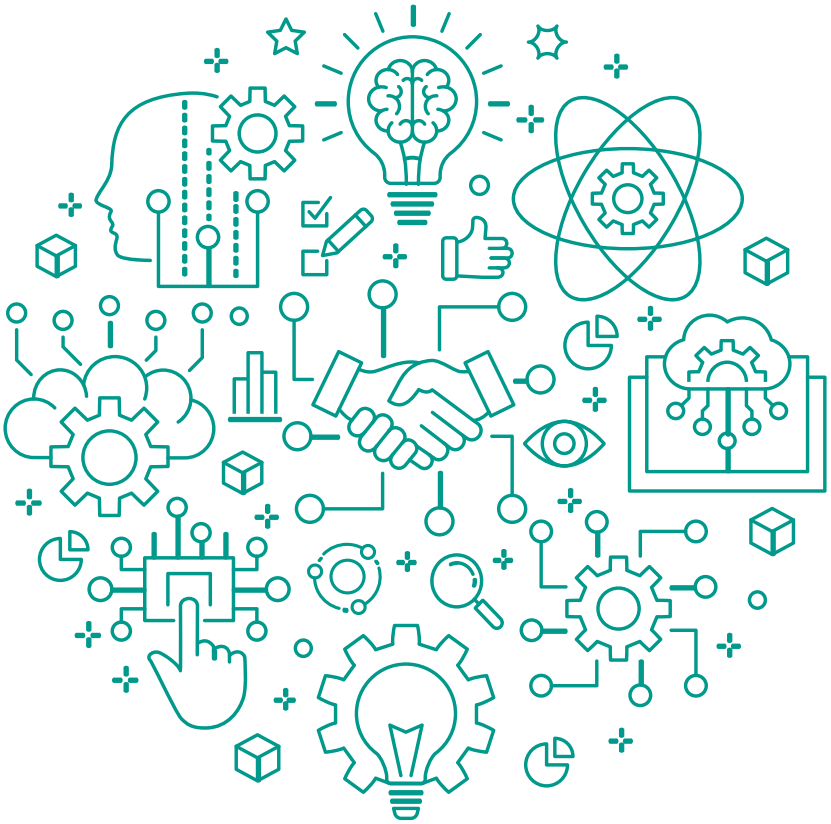
Imate procese i kontrole za upravljanje kojim podacima AI sustavi mogu pristupiti i za upravljanje sadržajem koji generira AI u skladu s njegovom osjetljivošću (i osjetljivošću inputa koji su ušli u njegovo generiranje).

Dokumentirajte svoje podatke, modele i upite

Dokumentirate stvaranje, rad i upravljanje životnim ciklusom svih modela, skupova podataka i meta ili sistemskih upita. Vaša dokumentacija uključuje informacije relevantne za sigurnost kao što su izvori podataka o obuci (uključujući podatke o finom podešavanju i ljudske ili druge operativne povratne informacije), predviđeni opseg i ograničenja, zaštitne ograde, kriptografske kontrolne identifikacijske brojeve ili potpise, vrijeme zadržavanja, predloženu učestalost pregleda i moguće načine kvara. Korisne strukture koje pomažu u tome uključuju kartice modela, podatkovne kartice i softverske popise materijala (SBOM). Izrada sveobuhvatne dokumentacije podupire transparentnost i odgovornost[11].

Upravljajte svojim tehničkim dugom

Kao i kod bilo kojeg softverskog sustava, identificirate, pratite i upravljate svojim 'tehničkim dugom' tijekom životnog ciklusa sustava umjetne inteligencije (tehnički dug je mjesto gdje se donose inženjerske odluke koje nisu u skladu s najboljom praksom za postizanje kratkoročnih rezultata, nauštrb dugotrajnijih beneficija). Poput financijskog duga, tehnički dug nije sam po sebi loš, ali njime treba upravljati od najranijih faza razvoja[12]. Shvaćate da to može biti veći izazov u kontekstu umjetne inteligencije nego za standardni softver i da će vaša razina tehničkog duga vjerojatno biti visoka zbog brzih razvojnih ciklusa i nedostatka dobro uspostavljenih protokola i sučelja. Osiguravate da vaši planovi životnog ciklusa (uključujući procese za stavljanje van pogona sustava umjetne inteligencije) procjenjuju, priznaju i umanjuju rizike za buduće slične sustave.



3. Sigurna implementacija

Ovaj odjeljak sadrži smjernice koje se primjenjuju na fazu **implementacije** životnog ciklusa razvoja AI sustava, uključujući zaštitu infrastrukture i modela od ugrožavanja, prijetnje ili gubitka, razvoj procesa upravljanja incidentima i odgovorno oslobađanje.

Osigurajte svoju infrastrukturu

Primijenite načela dobre sigurnosti na infrastrukturu koja se koristi u svakom dijelu životnog ciklusa vašeg sustava. Primjenjujete odgovarajuće kontrole pristupa svojim API-jima, modelima i podacima te njihovim obukama i procesnim cjevovodima, kako prilikom istraživanju i razvoja, tako i prilikom implementacije. To uključuje odgovarajuću segregaciju okruženja koja sadrže osjetljivi kod ili podatke. To će također pomoći u ublažavanju standardnih napada na kibernetičku sigurnost čiji je cilj ukrasti model ili naštetiti njegovoj izvedbi.

Kontinuirano štitite svoj model

Napadači bi mogli rekonstruirati funkcionalnost modela[13] ili podataka na kojima je obučen[14] izravnim pristupom modelu (stjecanjem težine modela) ili neizravnim (upitom o modelu putem aplikacije ili usluge). Napadači također mogu neovlašteno intervenirati s modelima, podacima ili upitima tijekom ili nakon obuke, čineći izlaz nepouzdanim.

Model i podatke štitite od izravnog i neizravnog pristupa:

- implementacijom standardnih najboljih praksi kibernetičke sigurnosti
- implementacijom kontrola na sučelju upita za otkrivanje i sprječavanje pokušaja pristupa, izmjene i izvlačenja povjerljivih informacija

Kako biste osigurali da potrošački sustavi mogu potvrditi modele, izračunavate i dijelite kriptografske hashove i/ili potpise datoteka modela (na primjer, težine modela) i skupova podataka (uključujući kontrolne točke) čim se model uvježba. Kao i uvijek s kriptografijom, dobro upravljanje ključem je suštinski bitno[15].

Vaš pristup smanjenju rizika povjerljivosti uvelike će ovisiti o slučaju upotrebe i modelu prijetnje. Neke aplikacije, na primjer one koje uključuju vrlo osjetljive podatke, mogu zahtijevati teoretska jamstva koja mogu biti teška ili skupa za primjenu. Ako je prikladno, tehnologije za poboljšanje privatnosti (kao što je diferencijalna privatnost ili homomorfna enkripcija) mogu se koristiti za istraživanje ili osiguranje razina rizika povezanih s potrošačima, korisnicima i napadačima koji imaju pristup modelima i rezultatima.

Razvijte postupke upravljanja incidentima

Neizbježnost sigurnosnih incidenata koji utječu na vaše sustave umjetne inteligencije odražava se u vašem odgovoru na incidente, eskalaciji i planovima sanacije. Vaši planovi odražavaju različite scenarije i redovito se ponovno procjenjuju s daljnjim razvijanjem sustava i širег istraživanja. Pohranjujete ključne digitalne resurse tvrtke u izvanmrežne sigurnosne kopije. Odgovorni su obučeni za procjenu i rješavanje incidenata povezanih s umjetnom inteligencijom. Kupcima i korisnicima pružate visokokvalitetne revizijske zapise i druge sigurnosne značajke ili informacije bez dodatnih troškova kako biste omogućili njihove procese odgovora na incidente.

Otpustite umjetnu inteligenciju odgovorno

Modele, aplikacije ili sustave puštate u promet tek nakon što ih podvrgnete odgovarajućoj i učinkovitoj sigurnosnoj procjeni kao što je benchmarking i red teaming (kao i drugim testovima koji su izvan opsega ovih smjernica, poput testa sigurnosti ili pravednost), i jasno upozoravate korisnike na poznata ograničenja ili moguće kvarove. Pojedinosti o bibliotekama za testiranje sigurnosti otvorenog koda dane su u odjeljku za dodatno čitanje na kraju ovog dokumenta.

Olakšajte korisnicima da rade ispravne stvari

Shvaćate da svaku novu postavku ili konfiguracijsku opciju treba procijeniti u skladu s poslovnom koristi koja proizlazi iz nje i svim sigurnosnim rizicima koje ona predstavlja. U idealnom slučaju, najsigurnija postavka bit će integrirana u sustav kao jedina opcija. Kada je potrebna konfiguracija, zadana opcija trebala bi biti općenito sigurna protiv uobičajenih prijetnji (to jest, sigurna prema zadanim postavkama). Primjenjujete kontrole kako biste spriječili korištenje ili implementaciju vašeg sustava na zlonamjerne načine.

Korisnicima dajete smjernice o prikladnoj upotrebi vašeg modela ili sustava, što uključuje isticanje ograničenja i mogućih kvarova. Korisnicima jasno navodite za koje su aspekte sigurnosti odgovorni i na transparentan način ih obavještavate o tome gdje (i kako) se njihovi podaci mogu koristiti, kako im se može pristupati ili pohranjivati (na primjer, ako se koriste za ponovnu obuku modela ili ih pregledavaju zaposlenici ili partneri).

4. Siguran rad i održavanje

Ovaj odjeljak sadrži smjernice koje se odnose na fazu **sigurnog rada i održavanja** životnog ciklusa razvoja sustava umjetne inteligencije. Ono pruža smjernice o radnjama koje su posebno relevantne nakon što je sustav postavljen, uključujući bilježenje i praćenje, upravljanje ažuriranjem i razmjenu informacija.

Pratite ponašanje svog sustava

Mjerite rezultate i performanse svog modela i sustava tako da možete promatrati iznenadne i postupne promjene u ponašanju koje utječu na sigurnost. Možete uzeti u obzir i identificirati potencijalne upade i kompromitacije, kao i prirodni pomak podataka.

Pratite unose u vaš sustav

U skladu sa zahtjevima za privatnošću i zaštitom podataka, nadzirete i bilježite unose u svoj sustav (kao što su zahtjevi za zaključivanjem, upiti ili upute) kako biste omogućili obveze usklađenosti, revizije, istrage i sanacije u slučaju ugrožavanja ili zlouporabe. To bi moglo uključivati eksplicitno otkrivanje ulaza izvan distribucije i/ili kontradiktornih unosa, uključujući one koji imaju za cilj iskorištavanje koraka pripreme podataka (kao što je obrezivanje i promjena veličine slika).

Slijedite pristup integrirane sigurnosti prilikom ažuriranja

Automatska ažuriranja prema zadanim postavkama uključujete u svaki proizvod i koristite sigurne, modularne postupke ažuriranja za njihovu distribuciju. Vaši procesi ažuriranja (uključujući režime testiranja i evaluacije) odražavaju činjenicu da promjene podataka, modela ili upita mogu dovesti do promjena u ponašanju sustava (na primjer, tretirate velika ažuriranja kao nove verzije). Podržavate korisnike u procjenama i odgovoroma na promjene modela (na primjer pružanjem pristupa pregledu i verzijama API-ja).

Prikupite i podijelite naučene lekcije

Sudjelujete u zajednicama za razmjenu informacija, surađujući u globalnom industrijskim ekosustavom, akademskom zajednicom i vladom kako biste na odgovarajući način podijelili najbolju praksu. Održavate otvorene linije komunikacije za povratne informacije o sigurnosti sustava, kako interno tako i eksterno za vašu organizaciju, uključujući davanje pristanka istraživačima sigurnosti za istraživanje i prijavu ranjivosti. Kada je to potrebno, eskalirate probleme široj zajednici, na primjer objavljivanjem biltena koji odgovaraju na otkrivanje ranjivosti, uključujući detaljno i potpuno nabrajanje zajedničkih ranjivosti. Poduzimate radnje za ublažavanje i ispravljanje problema brzo i na odgovarajući način.

Dodatna literatura

Razvoj umjetne inteligencije

Načela za sigurnost strojnog učenja
Detaljne smjernice NCSC-a o razvoju, implementaciji ili radu sustava s ML komponentom.

Integrirana sigurnost - promjena ravnoteže rizika kibernetičke sigurnosti: Načela i pristupi za integriranu sigurnost softvera
U koautorstvu CISA-e, NCSC-a i drugih agencija, ove smjernice opisuju kako bi proizvođači softverskih sustava, uključujući umjetnu inteligenciju, trebali poduzeti korake da ugrade sigurnost u fazu dizajna razvoja proizvoda i isporuku proizvoda koji sigurno dolaze iz kutije.

Ukratko o sigurnosnim problemima umjetne inteligencije
Izradio Njemački savezni ured za informacijsku sigurnost (BSI), ovaj dokument pruža uvod u moguće napade na sustave strojnog učenja i potencijalne obrane od tih napada.

Međunarodna vodeća načela procesa Hirošima za organizacije koje razvijaju napredne sustave umjetne inteligencije i Međunarodni kodeks ponašanja za organizacije koje razvijaju napredne sustave umjetne inteligencije procesa iz Hirošime
Ovi dokumenti proizvedeni kao dio izjave čelnika skupine G-7 o procesu umjetne inteligencije iz Hirošime, daju smjernice za organizacije koje razvijaju najnaprednije AI sustave, uključujući najnaprednije temeljne modele i generativne AI sustave s ciljem promicanja sigurnih i pouzdanih AI sustava diljem svijeta.

AI Verify
Singapurski okvir za testiranje upravljanja umjetnom inteligencijom i softverski alati koji potvrđuju izvedbu AI sustava prema skupu međunarodno priznatih načela putem standardiziranih testova.

Višeslojni okvir za dobre prakse kibernetičke sigurnosti za umjetnu inteligenciju — ENISA (europa.eu)
Okvir za usmjeravanje nacionalnih nadležnih tijela i dionika umjetne inteligencije o koracima koje trebaju slijediti kako bi osigurali svoje sustave umjetne inteligencije, operacije i procese.

ISO 5338: Procesi životnog ciklusa sustava umjetne inteligencije (u reviziji)
Skup procesa i povezanih koncepata za opisivanje životnog ciklusa sustava umjetne inteligencije temeljem strojnog učenja i heurističkih sustava.

Katalog kriterija usklađenosti AI usluga u oblaku (AIC4)
BSI-jev katalog kriterija usklađenosti usluga u oblaku AI pruža kriterije specifične za AI koji omogućuju procjenu sigurnosti AI usluge tijekom njenog životnog ciklusa.

NIST IR 8269 (nacrt) Taksonomija i terminologija kontradiktornog strojnog učenja
Skup procesa i povezanih koncepata za opisivanje životnog ciklusa AI sustava koji se temelje na strojnom učenju i heurističkim sustavima.

MITRE ATLAS
Baza znanja o protivničkim taktikama, tehnikama i studijama slučaja za sustave strojnog učenja (ML), po uzoru na MITER ATT&CK s kojim je povezana.

Opis rizika od katastrofe uzrokovane umjetnom inteligencijom (2023)
Ovaj dokument koji je izradio Centar za sigurnost umjetne inteligencije, utvrđuje područja rizika koje predstavlja umjetna inteligencija.

Veliki jezični modeli: Mogućnosti i rizici za industriju i nadležna tijela
Dokument koji je izradio BSI za tvrtke, nadležna tijela i programere koji žele naučiti više o prilikama i rizicima razvoja, implementacije i/ili korištenja LLM-a.

Projekti otvorenog koda za pomoć korisnicima u sigurnosnom testiranju AI modela uključuju:

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

Kibernetička sigurnost

Ciljevi uspješnosti kibernetičke sigurnosti CISA-e
Zajednički skup zaštita koje bi svi entiteti kritične infrastrukture trebali implementirati kako bi značajno smanjili vjerojatnost i utjecaj poznatih rizika i neprijateljskih tehnika.

NCSC CAF Framework
Cyber Assessment Framework (CAF) pruža smjernice za organizacije odgovorne za vitalno važne usluge i aktivnosti.

MITRE's Supply Chain Security Framework
Okvir za procjenu dobavljača i pružatelja usluga unutar opskrbnog lanca.

Upravljanje rizicima

NIST AI Risk Management Framework (AI RMF)
AI RMF opisuje kako upravljati socio-tehničkim rizicima za pojedince, organizacije i društvo koji su jedinstveno povezani s AI.

ISO 27001: Informacijska sigurnost, kibernetička sigurnost i zaštita privatnosti
Ova norma pruža organizacijama smjernice za uspostavljanje, implementaciju i održavanje sustava upravljanja informacijskom sigurnošću.

ISO 31000: Upravljanje rizikom
Međunarodni standard koji organizacijama daje smjernice i načela za upravljanje rizikom unutar organizacija.

NCSC smjernice za upravljanje rizikom
Ove smjernice pomažu stručnjacima za kibernetičku sigurnost da bolje razumiju i upravljaju rizicima kibernetičke sigurnosti koji utječu na njihove organizacije.

Bilješke

1. Ovdje definirana kao osoba, javno tijelo, agencija ili drugo tijelo koje razvija sustav umjetne inteligencije (ili ima razvijen sustav umjetne inteligencije) i stavlja taj sustav na tržište ili ga stavlja u službu pod svojim imenom ili zaštitnim znakom
2. Za više informacija o dizajnu sigurnosti pogledajte CISA-ovu web stranicu Secure by Design i smjernice Pomicanje ravnoteže rizika kibernetičke sigurnosti: Načela i pristupi računalnom programu s integriranom sigurnošću
3. Za razliku od pristupa umjetne inteligencije koji ne uključuju strojno učenje kao što su sustavi zasnovani na pravilima
4. CEPS u svojoj publikaciji opisuje sedam različitih vrsta interakcija u sklopu razvoja umjetne inteligencije 'Usklađivanje lanca vrijednosti umjetne inteligencije s EU-ovim Zakonom o umjetnoj inteligenciji'
5. ISO/IEC 22989:2022(en) definira ovo kao „funkcionalni element koji konstruira AI sustav"
6. NIST je zadužen za izradu smjernica (i poduzimanje drugih radnji) za unaprjeđenje sigurnog, zaštićenog i pouzdanog razvoja i korištenja umjetne inteligencije (AI). Pogledajte koje su odgovornosti NIST-a prema izvršnoj uredbi od 30. listopada 2023
7. Više informacija o modeliranju prijetnji dostupno je od OWASP Foundation
8. Pogledajte MITER ATLAS Adversarial Machine Learning 101
9. GitHub: RCE PoC za Tensorflow koji koristi zloćudni Lambda sloj
10. SLSA: "Očuvanje integriteta artefakta u bilo kojem lancu nabave softvera"
11. METI (Japansko ministarstvo gospodarstva, trgovine i industrije, 2023.), "Priručnik o uvođenju dokumentacije za upravljanje softverom (SBOM)'
12. Istraživanje Google-a: Strojno učenje: Visokokamatna kreditna kartica tehničkog duga
13. Tramèr et al 2016, Krađa modela strojnog učenja putem predviđanja API-a
14. Boenisch, 2020, Napadi na privatnost strojnog učenja (1. dio): Napadi inverzijom modela s IBM-ART okvirom
15. Nacionalni centar za kibernetičku sigurnost, 2020., Projektiranje i izgradnja privatne infrastrukture javnog ključa