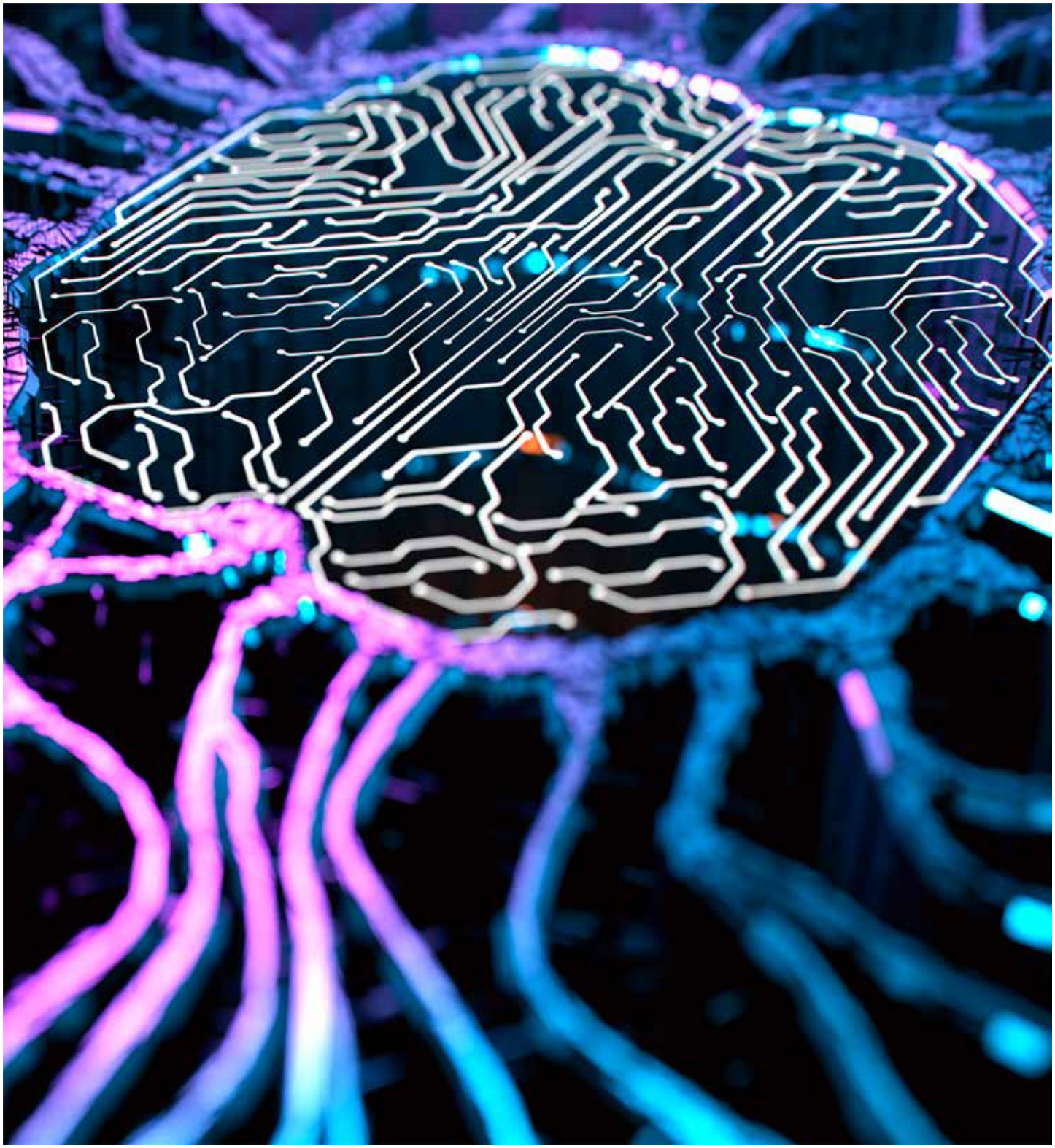
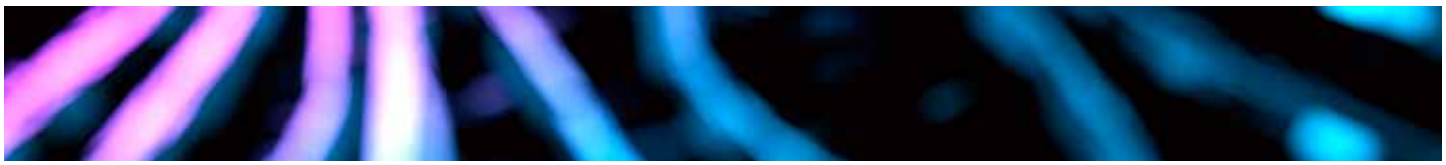
Linjigwida għall-iżvilupp sikur tas-sistema AI





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

Dwar dan id-dokument

Dan id-dokument huwa ppubblikat miċ-Ċentru Nazzjonali tas-Sigurtà Diġitali fir-Renju Unit (NCSC), l-Aġenzija tas-Sigurtà Diġitali u l-Infrastruttura fl-Istati Uniti (CISA) u l-imsieħba internazzjonali li ġejjin:

- L-Aġenzija Nazzjonali tas-Sigurtà (NSA)
- L-Uffiċċju Federali tal-Investigazzjonijiet (FBI)
- Iċ-Ċentru Awstraljan għas-Sigurtà Diġitali tad-Direttorat Awstraljan tas-Sinjali (ACSC)
- Iċ-Ċentru Kanadiż għas-Sigurtà Diġitali (CCCS)
- Iċ-Ċentru Nazzjonali għas-Sigurtà Diġitali fi New Zealand (NCSC-NZ)
- CSIRT tal-Gvern Ċilean
- L-Aġenzija Nazzjonali tas-Sigurtà Diġitali u l-Informazzjoni taċ-Ċeka (NUKIB)
- L-Awtorità tas-Sistema tal-Informazzjoni fl-Estonja u iċ-Ċentru Nazzjonali tas-Sigurtà Diġitali tal-Estonja (NCSC-EE)
- L-Aġenzija Franċiża tas-Sigurtà Diġitali (ANSSI)
- L-Uffiċċju Federali għas-Sigurtà tal-Informazzjoni tal-Ġermanja (BSI)
- Id-Direttorat Nazzjonali Diġitali tal-Iżrael (INCD)
- L-Aġenzija Nazzjonali tas-Sigurtà Diġitali tal-Italja (ACN)
- Iċ-Ċentru Nazzjonali tal-Ġappun għall-Preparazzjoni tal-Inċidenti u Strateġija għas-Sigurtà Diġitali
- Is-Segretarjat Ġappuniż għall-Policies tax-Xjenza, Teknoloġija u Innovazzjoni, fl-Uffiċju tal-Kabinett
- L-Aġenzija Nazzjonali għall-Iżvilupp tat-Teknoloġija tal-Informazzjoni fin-Niġerja (NITDA)
- Iċ-Ċentru Nazzjonali tas-Sigurtà Diġitali Norveġiż (NCSC-NO)
- Il-Ministeru Pollakk għall-Affarijiet Diġitali
- L-Istitut Nazzjonali Pollakk għar-Riċerka (NASK)
- Is-Servizz għall-Intelliġenza Nazzjonali tar-Repubblika tal-Koreja (NIS)
- L-Aġenzija tas-Sigurtà Diġitali ta' Singapore (CSA)

Rikonoxximent

Dawn l-organizzazzjonijiet li ġejjin ikkontribwew għall-Iżvilupp ta' dawn il-linjigwida:

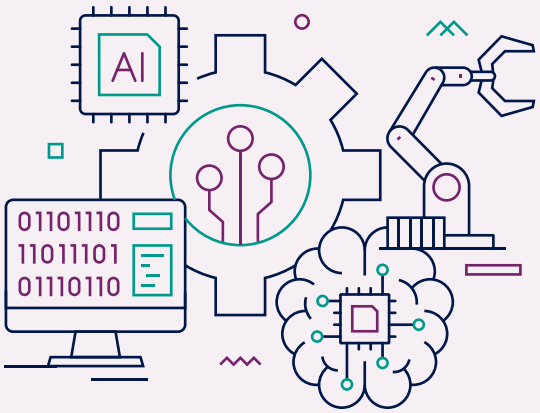
- L-Istitut Alan Turing
- Anthropic
- Databricks
- Iċ-Ċentru għas-Sigurtà u Teknoloġija Emerġenti fi ħdan l-Università ta' Georgetown
- Google
- Google DeepMind
- IBM
- Imbue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- L-Istitut tal-Inġinerija tas-Software fl-Università Carnegie Mellon
- Iċ-Ċentru ta' Stanford għall-Ħarsien tal-AI
- Il-Programm dwar il-Ġeopolitika, Teknoloġija u Governanza ta' Stanford

Rinunzja tad-dritt legali

L-informazzjoni f'dan id-dokument hija provduta 'kif inhi' mill-NCSC u l-organizzazzjonijiet awturi li mhumiex responsabbli ta' l-ebda telf, deni jew ħsara ta' kull tip ikkawżat mill-użu tiegħu ħlief jekk ikun mitlub mill-liġi. L-informazzjoni f'dan id-dokument ma jikkostitwix jew jimplika approvazzjoni jew rakkomandazzjoni ta' kwalunkwe organizzazzjoni terzi parti, prodott, jew servizz mill-NCSC u agenziji awtriċi. Konnessjonijiet u riferenzi għal websites u materjali ta' terzi persuni huma provduti bħala informazzjoni biss u ma jirrapreżentawx approvazzjoni jew rakkommandazzjoni ta' dawn ir-riżorsi flok oħrajn.

Dan id-dokument huwa provdut fuq bażi TLP: CLEAR (https://www.first.org/tlp/).

Kontenut



Sommarju Eżekuttiv

Dan id-dokument jirrakkommanda linjigwida għall-provvedituri ta' kull sistema li tuża l-intelliġenza artifiċjali (AI), kemm jekk dawn is-sistemi kienu maħluqa mill-bidu nett jew mibnija fuq għodda u servizzi provduti minn ħaddieħor. L-Implimentazzjoni ta' dawn il-linjigwida se jgħinu provvedituri jibnu sistemi tal-AI li jiffunzjonaw kif huma intenzjonati li jiffunzjonaw, disponibbli meta huma meħtieġa, u jaħdmu mingħajr ma jiżvelaw informazzjoni sensittiva lill-persuni mhux awtorizzati.

Dan id-dokument huwa primarjament immirat lejn provvedituri tas-sistemi tal-AI li qed jużaw mudelli offruti minn xi organizzazzjoni, jew interfaces esterni tal-applikazzjoni tal-programmi (API's). Inħeġġu lill-istakeholders **kollha** (inklużi xjentisti tal-informazzjoni, żviluppaturi, maniġers, dawk li jieħdu deċiżjonijiet u riskji) biex jaqraw dawn il-linjigwida biex jgħinuhom jagħmlu deċiżjonijiet infurmati dwar id-**disinn**, **l-iżvilupp**, **l-ingaġġ** u **l-operazzjoni** tas-sistemi tal-AI tagħhom.

Dwar il-linjigwida

Is-sistemi tal-AI għandhom il-potenzjal li jġibu bosta benefiċċji fis-soċjetà. Intant, biex l-opportunitajiet tal-AI jkunu realizzati kompletament, għandha tkun żviluppata, ingaġġata u operata b'mod sikur u responsabbli.

Is-sistemi tal-AI huma soġġetti għall-vulnerabbilitajiet ġodda tas-sigurtà li għandhom bżonn ikunu kkunsidrati mat-theddid normali tas-sigurtà diġitali. Meta r-ritmu tal-iżvilupp huwa għoli - bħalma huwa l-każ tal-AI - is-sigurtà tistà tkun kunsiderazzjoni sekondarja. Is-sigurtà għandha tkun rekwiżit ċentrali, mhux biss fil-fażi tal-iżvilupp, imma fiċ-ċiklu tal-ħajja kollha tas-sistema.

Għal din ir-raġuni, il-linjigwida huma maqsuma f'erba' oqsma prinċipali fiċ-ċiklu tal-ħajja tal-iżvilupp tas-sistema tal-AI: **disinn sikur**, **żvilupp sikur**, **ingagg sikur**, and **operazzjoni u manutenzjoni sikuri**. Għal kull sezzjoni nissuġġerixxu kunsiderazzjonijiet u mitigazzjonijiet li jgħinu biex inaqqsu r-riskju inġenerali li jkollu l-proċess tal-iżvilupp ta' sistema tal-AI organizzattiva.

1. **Disinn sikur**
   Din is-sezzjoni tinkludi linjigwida li japplikaw għal-istadju tad-disinn taċ-ċiklu tal-ħajja tal-iżvilupp tas-sistema tal AI. Tkopri edukazzjoni fuq il-mudelli tar-riskji u theddid, kif ukoll suġġetti speċifiċi u għażliet flok oħrajn li wieħed jikkunsidra dwar id-disinn tas-sistema u tal-mudell.

2. **Żvilupp sikur**
   Din is-sezzjoni fiha linjigwida li japplikaw għall-istadju tal-iżvilupp taċ-ċiklu tal-ħajja tal-iżvilupp tas-sistema tal-AI inklużi sigurtà fil-provvista, dokumentazzjoni, u assi u immaniġġjar ta' dejn tekniku.

3. **Ingaġġ sikur**
   Din is-sezzjoni fiha linjigwida li japplikaw għall-istadju tal-ingaġġ taċ-ċiklu tal-ħajja tal-iżvilupp tas-sistema tal-AI, inkluż il-protezzjoni tal-infrastruttura u l-mudelli mill-kompromess, theddid jew telf, żvilupp tal-proċessi tal-immaniġġjar tal-inċidenti, u implimentazzjoni responsabbli.

4. **Operazzjoni u manutenzjoni sikuri**
   Din is-sezzjoni fiha linjigwida li japplikaw għall-istadju tal-operazzjoni u manutenzjoni taċ-ċiklu tal-ħajja tal-iżvilupp tas-sistema tal-AI. Tipprovdi linjigwida fuq azzjonijiet li huma partikolarment rilevanti ladarba sistema tkun ġiet ingaġġata, inkluż illogjar, moniteraġġ, maniġġjar aġġornat u qsim ta' informazzjoni

Il-linjigwida jsegwu perspettiva ta''secure by default' u huma f'alinjament qrib ma' prattiċi definiti fil-gwida tal-iżvilupp u l-ingaġġ tal-NCSC, l-Istruttura Sigura tal-Iżvilupp tas-Software tal-NIST ,u prinċipji ddisinjati b'sigurtà, 'ippubblikati minn CISA, l-NCSC u aġenziji internazzjonali diġitali. Jipprijoritizzaw:

- li jkunu responsabbli tar-riżultati tas-sigurtà għall-klijenti
- li jħaddnu trasparenza u akkontabilità radikali
- li jibnu strutturi u tmexxija organizzazzjonali sabiex is-sigurtà bħala parti mid-disinn tkun prijorita' prinċipali fin-negozju

Introduzzjoni

Is-sistemi tal-Intelliġenza Artifiċjali għandhom il-potenzjal li jġibu bosta benefiċċji fis-soċjetà. Intant, biex l-opportunitajiet tal-AI jkunu realizzati kompletament, għandha tkun żviluppata, ingaġġata u operata b'mod sikur u responsabbli. Is-sigurtà diġitali hija prekundizzjoni neċessarja għall-ħarsien, qawwa, privatezza, imparzjalità, effikaċja u dipendibiità tas-sistemi tal-AI.

Intant, is-sistemii tal-AI huma soġġetti għall-vulnerabbilitajiet ġodda tas-sigurtà li għandhom bżonn ikunu kkunsidrati mat-theddid normali tas-sigurtà diġitali. Meta r-ritmu tal-iżvilupp huwa għoli - bħalma huwa l-każ tal-AI - is-sigurtà tistà tkun kunsiderazzjoni sekondarja. Is-sigurtà għandha tkun rekwiżit ċentrali, mhux biss fil-fażi tal-iżvilupp, imma fiċ-ċiklu tal-ħajja kollha tas-sistema.

Dan id-dokument jirrakkommanda linjigwida għall-provvedituri [1] ta' kull sistema li tuża l-intelliġenza artifiċjali (AI), kemm jekk dawn is-sistemi kienu maħluqa mill-bidu nett jew mibnija fuq għodda u servizzi provduti minn ħaddieħor. L-Implimentazzjoni ta' dawn il-linjigwida se jgħinu provvedituri jibnu sistemi tal-AI li jiffunzjonaw kif huma intenzjonati li jiffunzjonaw, disponibbli meta huma meħtieġa, u jaħdmu mingħajr ma jiżvelaw informazzjoni sensittiva lill-persuni mhux awtorizzati.

Dawn il-linjigwida għandhom ikunu kkunsidrati flimkien ma' sigurtà diġitali stabbilita, immaniġġjar tar-riskju u l-aħjar prattika tar-rispons tal-inċidenti. Partikolarment, inħeġġu lill-provvedituri biex isegwu l-prinċipji ta''secure by design' [2] żviluppati mill-Aġenzija tas-Sigurtà Diġitali u s-Sigurta Infrastrutturali tal-Istati Uniti (CISA), ċ-Ċentru Nazzjonali tas-Sigurtà Diġitali fir-Renju Unit (NCSC), u l-imsieħba internazzjonali kollha tagħna. Il-prinċipji jipprijoritizzaw:

- li jkunu responsabbli tar-riżultati tas-sigurtà għall-klijenti
- li jħaddnu trasparenza u akkontabilità radikali
- li jibnu strutturi u tmexxija organizzazzjonali sabiex is-sigurtà bħala parti mid-disinn tkun prijorita' prinċipali fin-negozju

Biex ikunu segwiti l-prinċipji 'secure by design' hemm bżonn ta riżorsi sinjifikanti matul iċ-ċiklu tal-ħajja tas-sistema. Dan ifisser li l-iżviluppaturi għandhom jinvestu billi jipprijoritizzaw **fatturi**, **mekkaniżmi**, u **implimentazzjoni** ta' għodda li jipproteġu lill-klijenti f'kull livell tad-disinn tas-sistema, u f'kull stadju tal-iżvilupp fiċ-ċiklu tal-ħajja. Dan jevita disinni oħra fil-futur għolja fil-prezz. kif ukoll jipproteġi klijenti u l-informazzjoni tagħhom fiż- żmien qarib.

Għaliex hija differenti s-sigurtà tal-AI?

F'dan id-dokument nużaw 'AI' biex nirreferu speċifikamemt fl-applikazzjonijiet tat-tagħlim tal-magni (ML)[3]. Kull tip ta' ML huwa fl-iskop Niddefinixxu applikazzjonijiet ta' ML bħala applikazzjonijiet li:

- jinvolvu biċċiet (mudelli) tas-software li jippermettu kompjuters biex jirrikonoxxu u jġibu kuntest fit-tfassil tal-informazzjoni mingħajr ma jkun hemm bżonn li r-regoli jkunu pprogrammati b'mod espliċitu minn uman
- jiġġeneraw tbassir, rakkommandazzjonijiet, jew deċiżjonijiet ibbażati fuq raġunar statistiku

Jeżisti theddid diġitali kurrenti, kif ukoll sistemi tal-AI huma suġġetti għal tipi ġodda ta' vulnerabbilitajiet. It-terminu 'adversarial machine learning' (AML), huwa użat biex jiddeskrivi sfruttament ta' vulnerabbilitajiet fundamentali fil-partijiet tal-ML, inkluż hardware, software, workflows u supply chains. AML jiffaċilita attakkanti li jikkawżaw imġieba mhux b'intenzjoni fis-sistemi ML li tista' tinkludi:

- effett fuq il-klassifikazzjoni tal-mudell jew azzjoni rigressiva
- permess lil min jużah biex iwettaq azzjonijiet mhux awtorizzati
- l-akkwist ta' informazzjoni sensittiva tal-mudell

Hemm bosta modi biex takkwista dawn l-effetti, bħal attakki malajr ta' injezzjoni fl-isfera tal-mudell kbir tal-lingwa (LLM), jew il-korruzzjoni deliberata tal-informazzjoni dwar it-taħriġ jew feedback (magħruf bħala 'avvelenament tal-informazzjoni')

Min għandu jaqra dan id-dokument?

Dan id-dokument huwa primarjament immirat lejn provvedituri tas-sistemi tal-AI li qed jużaw mudelli offruti minn xi organizzazzjoni, jew interfaces esterni tal-applikazzjoni tal-programmi (API's). Intant inħeġġu lill-istakeholders **kollha** (inklużi xjentisti tal-informazzjoni, żviluppaturi, maniġers, dawk li jieħdu deċiżjonijiet u riskji) biex jaqraw dawn il-linjigwida biex jgħinuhom jagħmlu deċiżjonijiet infurmati dwar **id-disinn**, **l-iżvilupp**, **l-ingaġġ** u **l-operazzjoni** tas-sistemi tal-AI tagħhom.

Intant, mhux il-linjigwida kollha se jkunu applikabbli direttament għall-organizzazzjonijiet kollha. Dan il-livell ta' sofistikazzjoni u il-metodi tal-attakk ivarjaw skond l-avversarju li jattakka s-sistema tal-AI, għalhekk il-linjigwida għandhom ikunu kkunsidrati flimkien mal-użu tal każijiet u l-profil ta' theddid tal-organizzazzjoni tiegħek.

Min huwa responsabbli biex jiżviluppa AI sikur?

Normalment hemm ħafna atturi fis-supply chains tal-Intelliġenza Artifiċjali moderna. Angolu sempliċi jassumi żewġ entitajiet:

- 'il-provveditur' li huma responsabbli għal-kustodja tal-iinformazzjoni, l-iżvilupp algoritmiku, disinn, ingaġġ u manutenzjoni
- l-utent, li jipprovdi kontribut fi dħul u jirċievi kontribut fi ħruġ

Filwaqt li dan l-angolu ta' provveditur-utent huwa użat f'bosta applikazzjonijiet, qiegħed isir inqas komuni kull ma jmur [4], meta provvedituri qed ifittxu li jinkorporaw software, informazzjoni, mudelli jew/u servizzi remoti provduti minn terzi persuni fis-sistemi tagħhom stess. Dawn is-supply chains kumplessi jagħmluha iktar diffiċli għall-utenti biex jifhmu fejn taqa' r-responsabbilita' ta' AI sikura.

L-utenti (kemm jekk end-users, jew provvedituri li jinkorporaw partijiet esterni tal-AI[5]) tipikament m'għandhomx viżibilità sufiċjenti u/jew kompetenza biex jifhmu kompletament, jevalwaw jew jindirizzaw riskji marbutin mas-sistemi li jkunu qed jużaw. Għalhekk, skond il-prinċipji ta''secure by design' **provvedituri tal-partijiet tal-AI għandhom jieħdu responsabbilta tar-riżultati sikuri tal-utenti iktar l-isfel fis-supply chain.**

Il-provvedituri għandhom jimplimentaw kontrolli u mitigazzjonijiet tas-sigurtà fejn huwa possibbli fil-mudelli tagħhom, pipelines u/jew sistemi, u fejn hemm issettjar, jimplimentaw l-iktar għażla sikura bħala 'default'. Fejn ir-riskji ma jistgħux ikunu mitigati, il-provveditur għandu jkun responsabbli li:

- jinforma lill-utenti l-iktar isfel fis-supply chain dwar ir-riskji li huma u (jekk japplika) l-utenti tagħhom stess qegħdin jaċċettaw
- javżahom dwar kif jużaw il-partijiet b'mod sikur

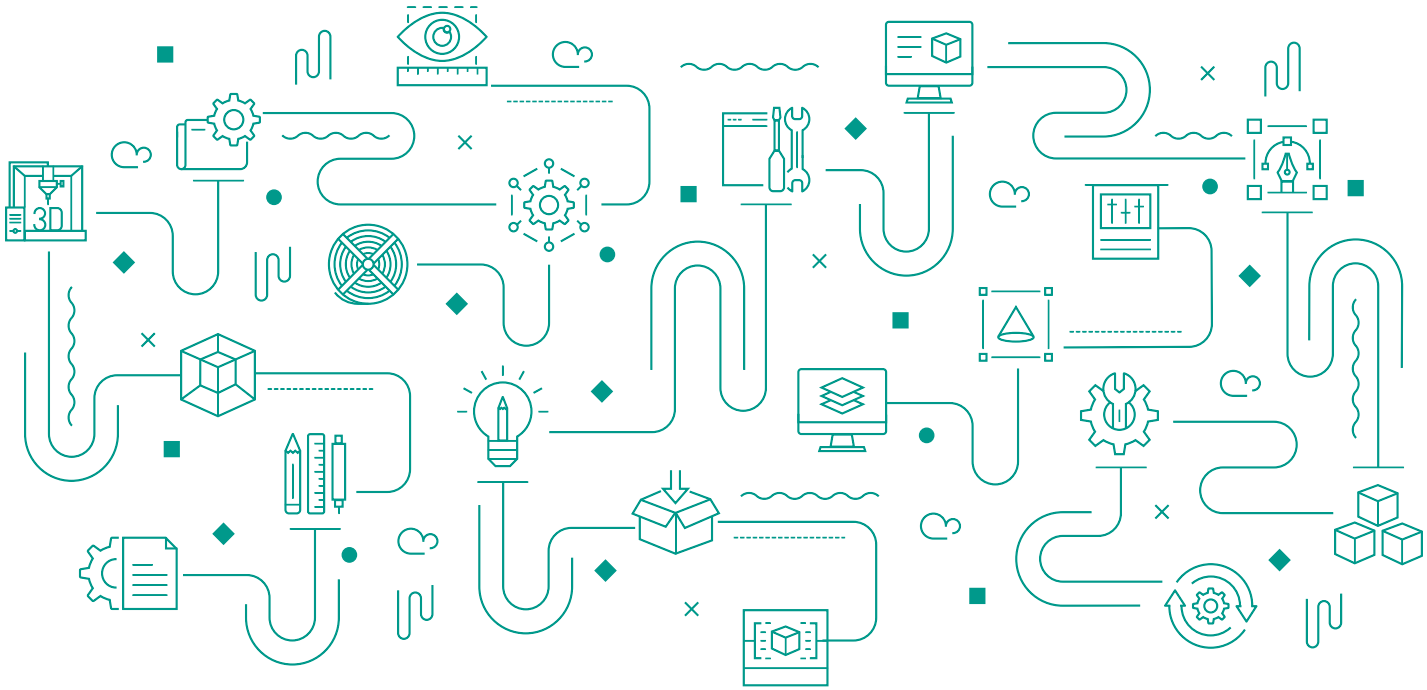
Fejn kompromess tas-sistema jista' jġib ħsara tanġibbli u wiesgħa fiżikament jew fil-fama, telf sinjifikanti fl-operat tan-negozju, kxif ta informazzjoni sensittiva jew kunfidenzjali u/jew implikazzjonijiet legali, r-riskji fis-sigurtà diġitali tal-AI għandhom ikunu trattati bħala **kritiċi**.

Linjigwida għall-iżvilupp sikur tas-sistema tal-AI

Il-linjigwida huma maqsuma f'erba' oqsma prinċipali fiċ-ċiklu tal-ħajja tal-iżvilupp tas-sistema tal-AI: **disinn sikur**, **żvilupp sikur**, **ingagg sikur**, and **operazzjoni u manutenzjoni sikuri.** Għal kull qasam, nissuġġerixxu konsiderazzjonijiet u mitigazzjonijiet li se jgħinu r-riskju inġenerali tal-proċess tal-iżvilupp tal-organizzazzjoni tas-sistema tal-AI.

Il-linjigwida mniżżla f'dan id-dokument huma qrib sewwa tal-prattiċi taċ-ċiklu tal-ħajja tal-iżvilupp tas-software definiti fil-:

- Il-Gwida għall-Iżvilupp u Ingaġġ Sikur tal-NCSC
- Il-Qafas dwar l-Iżvilupp Sikur tas-Software tal-Istitut Nazzjonali tal-Istandards u t-Teknoloġija (NIST)(SSDF)[6]



1. Disinn Sikur

Din il-parti fiha linjigwida li japplikaw għall-istadju **tad-disinn** tal-iżvilupp taċ-ċiklu tal-ħajja tas-sistema tal-AI. Tkopri edukazzjoni dwar mudellar tar-riskji u theddid, kif ukoll suġġetti speċifiċi u għażliet biex wieħed jikkunsidra dwar id-disinn tas-sistema u l-mudell.

Qanqal l-għarfien tal-istaff tat-theddid u riskji

Is-sidien tas-sistema u mexxejja għolja jifhmu t-theddid fis-sigurtà tal-AI u t-tnaqqis ta' dan. Ix-xjentisti u l-iżviluppaturi tad-data jibqgħu jkunu konxji tat-theddid relevanti għas-sigurtà u tipi ta' falliment u jgħinu 'l dawk fir-risku biex jieħdu deċiżjonijiet infurmati. Inti tipprovdi lill-utenti bi gwida dwar ir-riskji uniċi fis-sigurtà li jiffaċjaw sistemi tal-AI (per eżempju, bħala parti ta' taħriġ standard InfoSec) u titrenja lill-iżviluppaturi fuq it-teknika sikura tal-kowding u prattiċi sikuri u responsabbli tal-AI.

Immudella t-theddid għas-sistema tiegħek

Bħala parti mill-proċess tiegħek tal-immaniġġjar tar-riskju, applika proċess ħolistiku biex tassessja t-theddid tas-sistema tiegħek, li tinkludi li tifhem l-impatt potenzjali fuq is-sistema, l-utenti, organizzazzjonijiet, u s-soċjetà iktar wiesgħa jekk parti mill-AI tiġi f'kompromess jew taġixxi b'mod mhux mistenni[7]. Dan il-proċess jinvolvi assessjar tal-impatt ta' theddid speċifiku tal-AI[8] u dokumentazzjoni tad-deċiżjonijiet tiegħek.

Inti tirrikonoxxi li s-sensitività u tipi ta' data użata fis-sistema tiegħek tista' tinfluwenza l-valur tagħha bħala mira ta' xi attakkant. L-assessjar tiegħek għandu jikkunsidra li xi theddid jista' jikber hekk kif s-sistemi tal-AI isiru iktar ikkunsidrati bħala miri ta' valur għoli, u hekk kif l-AI stess tiffaċilita vektori ġodda awtomatiċi ta' attakki.

Iddisinja s-sistema tiegħek għas-sigurtà kif ukoll il-funzjonalità u l-operat

Kun kunfidenti li l-biċċa xogħol kurrenti hija l-iktar waħda adattata biex tkun indirizzata bl-użu tal-AI. Meta tiddetermina dan, tassessja jekk tkunx għażilt id-disinn speċifiku tal-AI tiegħek it-tajjeb. Inti tikkunsidra l-mudell tad-theddid diġitali u l-mitigazzjonijiet assoċjati tas-sigurtà mal-funzjonalità, l-esperjenza tal-utent, l-ambjent tal-ingaġġ, l-operat, iċ-ċertezza, moniteraġġ, ir-rekwiżiti etiċi u legali, fost konsiderazzjonijiet oħra. Per eżempju:

- tikkunsidra s-sigurtà tas-supply chain meta tagħżel jekk tiżviluppax internament jew tuża partijiet esterni, per eżempju:
    - l-għażla tiegħek li tħarreġ mudell ġdid, tuża mudell kurrenti (bi jew mingħajr fine-tuning) jew taċċessa mudell permezz ta API estern li jkun tajjeb għall-bżonnijiet tiegħek
    - l-għażla tiegħek li taħdem ma' provveditur estern ta' mudell tinkludi evalwazzjoni diliġenti tal-istat tas-sigurtà tal-provveditur
    - jekk tuża librerija esterna, tagħmel evalwazzjoni diliġenti (per eżempju, biex taċċerta li l-librerija għandha kontrolli li ma jħallux lis-sistema tillowdja mudelli mhux fdati mingħajr ma jesponu immedjatament lilhom infushom għall-eżekuzzjoni arbitrarja tal-kodiċi[9])
    - Timplimenta scanning u iżolazzjoni/sandboxing meta timporta mudelli ta' terzi parti jew piż serjalizzat, li għandu jiġi trattat bħala kodiċi mhux li tafdah ta' terzi persuni u li jista' jiffaċilita l-eżekuzzjoni remota ta' kodiċi

- Jekk tuża API's esterni, applika il-kontrolli t-tajba għall-informazzjoni li tista' tintbagħat għas-servizzi barra mill-kontroll tal-organizzazzjoni bħal per eżempju li jitolbu lill-utent biex jilloggja u jikkonferma qabel ma tintbagħat informazzjoni potenzjalment sensittiva
- tiċċekkja u tnaddaf l-informazzjoni u kontribut ieħor; dan jinkludi meta tinkorpora feedback minn dawk li jużaw s-servizz jew informazzjoni kontinwa ta' tagħlim fil-mudell tiegħek, billi tirrikonoxxi li l-informazzjoni ta' taħriġ tiddefinixxi l-imġieba tas-sistema

- tintegra żvilupp fis-sistema tas-software tal-AI fl-aħjar prattiċi kurrenti tal-iżvilupp u operazzjonijiet sikuri; l-elementi kollha tas-sistema tal-AI huma miktubin f'ambjent kif xieraq bl-użu ta' prattiċi tal-kodiċi u lingwi li jnaqqsu jew jeliminaw klassijiet magħrufa ta' vulnerabilitajiet fejn huwa plawsibbli

- Jekk il-partijiet tal-AI ikollhom bżonn iġegħlu li tibda xi ħaġa, per eżempju tiswija ta' files jew direzzjoni ta' materjal għall-sistemi esterni, applika r-restrizzjonijiet adatti għall-azzjonijiet li jistgħu jiġru (dan jinkludi fail-safes tal-AI estern jew mhux AI jekk hemm bżonn)

- deċiżjonijiet dwar interazzjoni mal-utent huma infurmati minn riskji speċifiċi tal-AI, per eżempju:

  - is-sistema tiegħek tipprovdi lill-utenti b'materjal li jista' jintuża mingħajr ma jiżvela livelli mhux neċessarji ta' dettal lil xi persuna li tista' tattakka sistemi
  - Jekk hemm bżonn, is-sistema tiegħek tipprovdi sigurtà effettiva għall-materjal li joħroġ mill-mudelli.
  - Jekk qed toffri API lill-klijenti u kollaboraturi esterni, applika kontrolli xierqa li jnaqqsu mill-attakki fuq is-sistema tal-AI permezz tal-API
  - Integra l-iktar issettjar sikur fis-sistema 'by default'
  - Applika l-iktar prinċipji mhux privileġġati biex tillimita l-aċċess għall-funzjonalità ta' sistema
  - Spjega l-kapaċitajiet iktar riskjużi lill-utenti u itlobhom li jgħażlu li jużawhom; tikkomunika każijiet ta' projbizzjoni, u fejn huwa possibbli, tinforma lill-utenti b'soluzzjonijiet alternattivi

Ikkunsidra l-benefiċċji tas-sigurtà u trade-offs meta tagħżel il-mudell tal-AI tiegħek

L-għażla tiegħek tal-mudell tal-AI se tkun tinvolvi bbilanċjar ta' serje ta' rekwiżiti. Dan jinkludi għażla ta' arkitettura tal-mudell, konfigurazzjoni, informazzjoni dwar taħriġ, l-algoritmu tat-taħriġ u hyperparameters. Id-deċiżjonijiet tiegħek huma nfurmati mill-mudell tiegħek tad-theddid diġitali, u huma assessjati mill-ġdid kif ir-riċerka tas-sigurtà tal-AI tkompli tavvanza u l-għarfien dwar it-theddid jevolvi.

Meta tuża mudell tal-AI, il-konsiderazzjonijiet tiegħek probabbli jinkludu, imma ma jkunux limitati għal:

- kumplessità tal-mudell li qed tuża, jiġifieri, l-arkitettura li tuża u n-numru ta' fatturi li jillimitaw; il-mudell li tagħżel u n-numru ta' limitazzjonijiet, se jaffetwaw fost fatturi oħra, kemm hemm bżonn ta' informazzjoni dwar taħriġ u kemm hija f'saħħitha fil-konfront tal-bidliet fl-informazzjoni li tidħol meta tkun tintuża

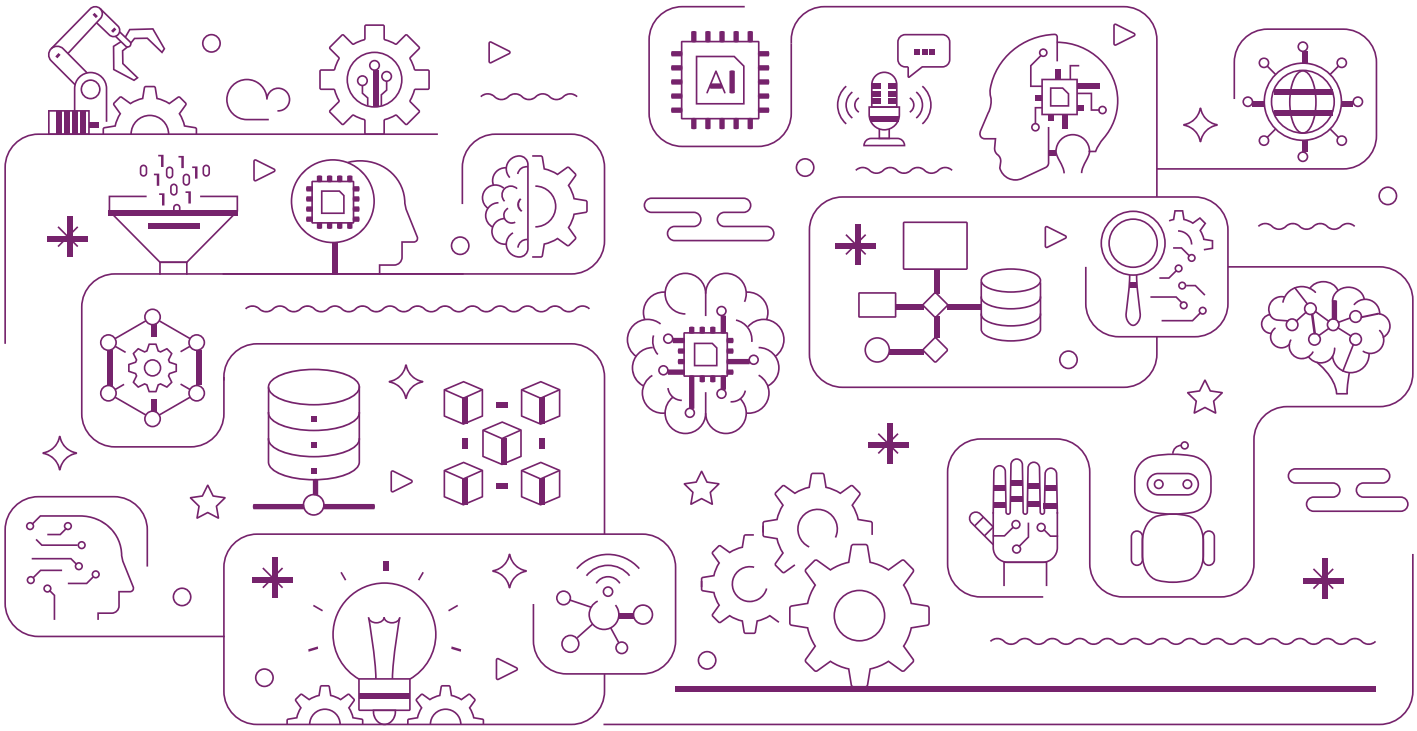
- kemm huwa adattat tajjeb il-mudell tiegħek għall-użu tal-każ tiegħek u/jew il-probabbilità li tadattah għall-bżonn speċifiku tiegħek (per eżempju 'fine-tuning)

- il-kapaċità li iġġib f'livell wieħed, tinterpreta u tispjega l-kontribut li joħroġ mill-mudell tiegħek (per eżempju għal debugging, reviżjoni jew konformità regolatorja); jista' jkun hemm benefiċċji billi tuża mudelli iktar sempliċi u trasparenti minflok mudelli kbar u kumplessi li huma iktar diffiċli biex tinterpretahom

- karatteristiċi tat-taħriġ ta' settijiet ta' informazzjoni, inklużi d-daqs, l-integrità, il-kwalità, sensittività, ż-żmien, rilevanza u diversità

- il-valur fl-użu tat-tisħiħ tal-mudell (bħat-taħriġ avversarju), regolarizzazzjoni u/jew teknika li ttejjeb il-privatezza
- il-provenjenza u supply chains tal-partijiet inklużi l-mudell jew il-mudell fundatur, data tat-taħriġ u għodda assoċjati mas-sistema

Għal iktar informazzjoni dwar kemm minn dawn il-fatturi jkollhom impatt fuq x'jiġri mis-sigurtà, irreferi għall-Prinċipji dwar is-Sigurtà fit-Tagħlim tal-Magni', (ML) partikolarment Disinn tas-Sigurtà (mudell tal-arkitettura).



2. Żvilupp sikur

Din il-parti fiha linjigwida li japplikaw għall-istadju **tal-iżvilupp** taċ-ċiklu tal-ħajja tas-sistema tal-AI, inkluż sigurtà fis-supply chain, dokumentazzjoni, u immaniġġjar tad-dejn tal-assi u teknoloġiku.

Assigura s-supply chain tiegħek

Assessja u issorvelja s-sigurtà tas-supply chains tal-AI tiegħek matul iċ-ċiklu tal-ħajja tas-sistema, u itlob lis-suppliers biex jimxu mal-istess standards li l-organizzazzjoni tiegħek tapplika għall-software ieħor. Jekk is-suppliers ma jistgħux jimxu mal-istandards tal-organizzazzjoni tiegħek, aġixxi skond il-policies kurrenti tiegħek dwar l-immaniġġjar tar-riskji.

Meta ma jkunux prodotti internament, akkwista u żomm hardware u software assigurat u dokumentat tajjeb (per eżempju, mudelli, data, libreriji ta' software, modules, middleware, oqsfa, u API's esterni) minn sorsi verifikati kummerċjali u pubbliċi u żviluppaturi terzi parti biex tassigura sigurtà f'saħħitha fis-sistemi tiegħek.

Int tkun lest li tfalli biex issib soluzzjonijiet alternattivi għal-sistemi b'missjoni kritika, jekk il-kriterji tas-sigurtà ma jkunux sodisfati. Tuża riżorsi bħal Gwida tas-Supply Chain tal-NCSC u oqsfa bħas Supply Chain Levels for Software Artifacts (SLSA)[10] għall-attestazzjoni ta' tracking tas-supply chain u ċikli tal-ħajja tal-iżvilupp tas-software.

Identifika, issorvelja u ipproteġi l-assi tiegħek

Int tifhem il-valur li għandhom l-assi tiegħek tal-AI għall-organizzazzjoni tiegħek, inklużi mudelli, data (inkluż feedback mill-utenti tiegħek), prompts, software, dokumentazzjoni, logs u assessjar (inkluża informazzjoni dwar kapaċitajiet li huma potenzjalment mhux sikuri u modi ta' falliment) waqt li tirrikonoxxi fejn jirrapreżentaw investiment sinjifikattiv u fejn l-aċċess tagħhom jiffaċilita xi attakk. Titratta logs bħala data sensittiva u timplimenta kontrolli biex tipproteġi il-privatezza, l-integrità u l-aċċessibilità tagħhom.

Inti taf fejn l-assi tiegħek jinsabu u assessjajt u aċċettajt kull riskju assoċjat. Għandek proċessi u għodda x'tissorvelja, tawtentika, tikkontrolla l-verżjoni tagħhom u tassigura l-assi tiegħek, u tista' terġa' lura għall- istat tajjeb fil-każ ta' kompromess.

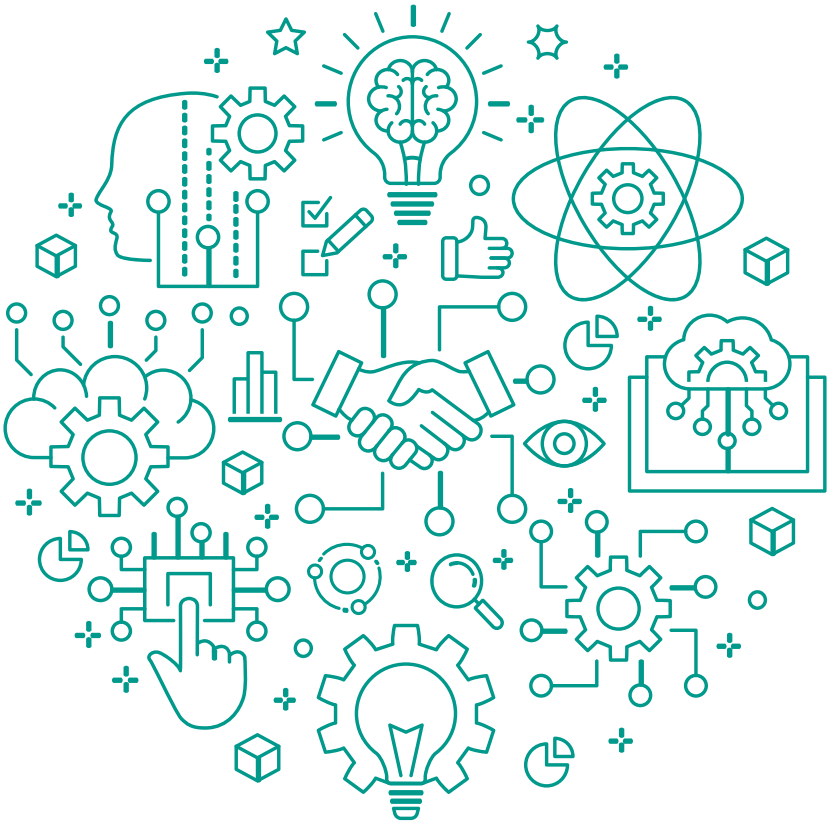
Għandek proċessi u kontrolli lesti biex timmaniġġja d-data li jistgħu jaċċessaw is-sistemi tal-AI, u biex timmaniġġja kontenut iġġenerat mill-AI skond is-sensitività tiegħu (u s-sensitività tal-informazzjoni li tidħol li ġġeneratha)

Iddokumenta l-informazzjoni, mudelli u prompts tiegħek

Tiddokumenta l-ħolqien, l-operat u l-immaniġġjar taċ-ċiklu tal-ħajja ta' kwalunkwe mudell, settijiet ta' data u prompts tas-sistema jew tal-meta. Id-dokumentazzjoni tiegħek tinkludi informazzjoni rilevanti fis-sigurtà bħal sorsi ta' data tat-taħriġ (inkluż data fid-dettal u feedback mill-persuni jew operazzjonali), l-iskop intenzjonat u limitazzjonijiet, protezzjonijiet, hashes kriptografiċi jew firem, żmien, suġġerimenti dwar il-frekwenza tar-reviżjoni u tipi potenzjali ta' fallimenti). Strutturi utli li jgħinu dan jinkludu cards tal-mudelli, tad-data u SBOM's (software bills of materials). Il-produzzjoni tad-dokumentazzjoni komprensiiva tissapportja t-trasparenza u l-akkontabilità[11].

Immaniġġja d-dejn tekniku tiegħek

Bħal f'kull sistema tas-software, identifika, issorvelja u immaniġġja 'd-dejn tekniku' tiegħek matul iċ-ċiklu tal-ħajja tas-sistema tal-AI (dejn tekniku huwa fejn id-deċiżjonijiet tal-inġinerija jonqsu mill-aħjar prattiċi biex jakkwistaw riżulatati għaż-żmien qasir, spejjeż tal-benefiċċji fit-tul). Bħad-dejn finanzjarju, d-dejn tekniku mhux minnu nnifsu ħażin iżda għandu jkun immaniġġjat mill-istadji bikrija tal-iżvilupp[12]. Tirrikonoxxi li jekk tagħmel dan jista' jkun iktar diffiċli fil-kuntest tal-AI milli f'software komuni, u li l-livelli tad-dejn tekniku se jkunu probabli għolja minħabba ċikli magħġġla tal-iżvilupp u nuqqas ta' protokolli u interfaces stabbiliti tajjeb. Inti taċċerta li l-pjanijiet taċ-ċiklu tal-ħajja tas-sistema tiegħek (inkluż il-proċessi tal-irtirar ta' sistemi tal-AI) jassessjaw, jirrikonoxxu u jnaqqsu r-riskji għal sistemi simili fil-futur.



3. Ingaġġ Sikur

Din il-parti fiha linjigwida li japplikaw għall-istadju tal-**ingaġġ** taċ-ċiklu tal-ħajja tas-sistema tal-AI, inkluż il-protezzjoni tal-infrastruttura u l-mudelli mill-kompromess, theddid u telf, żvilupp tal-proċessi tal-immaniġġjar tal-inċidenti, u l-ħruġ responsabbli fil-berah.

Assigura l-infrastruttura tiegħek

Inti tapplika prinċipji tal-infrastruttura tajbin għall-infrastruttura użata f'kull parti taċ-ċiklu tal-ħajja tas-sistema tiegħek. Inti tapplika kontrolli xierqa għall-aċċess tal-API's tiegħek, mudelli u data, u fis-sistemi tat-taħriġ u l-ipproċessar, fir-riċerka u żvilupp kif ukoll fl-ingaġġ. Dan jinkludi segregazzjoni kif suppost tal-ambjenti li jżommu fihom kowd jew data sensittiva. Dan jgħin ukoll biex jonqsu l-attakki standard fuq is-sigurtà diġitali li għandhom l-għan li jisirqu mudell jew jagħmlu ħsara fl-operat tiegħu.

Ipproteġi l-mudell tiegħek kontinwament

L-attakkanti jistgħu jkunu kapaċi jerġgħu jibnu mill-ġdid il-funzjonalità ta' mudell[13] jew id-data li fuqha kien imħarreġ[14], billi jaċċessaw mudell direttament (billi jakkwistaw piżijiet mudelli) jew indirettament (billi jinvestigaw il-mudell permezz ta' applikazzjoni jew servizz). Min jattakka jista' wkoll ibagħbas bil-mudelli, informazzjoni u suġġerimenti matul jew wara t-taħriġ, li jirrenei dak li joħroġ bla fiduċja.

Ipproteġi l-mudell u l-informazzjoni mill-aċċess dirett u indirett, ripettivament, billi:

- timplimenta l-aħjar prattiċi standard tas-sigurtà diġitali
- timplimenta kontrolli fuq il-paġna tat-tiftix (query interface) biex tinvestiga u tevita attentati ta' aċċess, modifika, u filtrazzjoni ta' informazzjoni kunfidenzjali

Taċċerta li sistemi tal-konsumazzjoni jistgħu jivvalidaw mudelli, tikkalkola u taqsam hashes kriptografiċi u/jew firem ta' files ta' mudellli (per eżempju, piż) u settijiet ta' data (inklużi checkpoints) hekk kif il-mudell ikun imħarreġ. Bħal dejjem bil-kriptografija, immaniġġjar prinċipali tajjeb huwa essenzjali[15].

L-attitudni tiegħek lejn il-mitigazzjoni fir-riskju tal-privatezza tiddependi konsiderevolment fuq il-każ tal-użu u il-mudell tat-theddid diġitali. Xi applikazzjonijiet, per eżempju dawk li jinvolvu informazzjoni sensittiva ħafna, jistgħu jirrikjedu garanziji tijoretiċi li jistgħu jkunu diffiċli jew għolja biex tapplika. Jekk hu xieraq, teknoloġiji li jsaħħu l-privatezza (bħal privatezza differenzjali jew encryption homomorfiku) jistgħu jintużaw jew jassiguraw livelli ta' riskju assoċjat mal-konsumatur, l-utent u l-attakanti li jkollhom aċċess għall-mudelli u l-informazzjoni li toħroġ.

Żviluppa proċeduri għall-Immaniġġjar tal-Inċidenti

L-inevitabbilità tal-inċidenti tas-sigurtà li jaffetwaw is-sistemi tal-AI hija riflessa fir-risposta tiegħek tal-inċidenti, pjanijiet tal-eskalazzjoni u rimedji tal-problemi. Il-pjanijiet tiegħek jirriflettu xenarji differenti u huma regolarment assessjati kif tkompli tevolvi s-sistema u r-riċerka iktar wiesgħa. Taħżen riżorsi diġitali kritiċi tal-kumpanija f'backups offline. Min jirreaġixxi huwa mħarreġ biex jassessja u jindirizza inċidenti li għandhom x'jaqsmu mal-AI. Tipprovdi logs tar-riviżonijiet ta' kwalità u fatturi oħra tas-sigurtà jew informazzjoni lill-klijenti u utenti mingħajr ħlas addizzjonali, biex tiffaċilita l-proċessi tagħhom tar-rispons tal-inċidenti.

Erħi l-AI fil-pubbliku b'mod responsabbli

Inti toħroġ fil-pubbliku mudelli, applikazzjonijiet u sistemi biss wara li dawn ikunu suġġetti ta' evalwazzjoni xierqa u effettiva tas-sigurtà bħal 'benchmarking' u 'red teaming' (kif ukoll testijiet oħra li huma barra mill-iskop ta' dawn il-linjigwida, bħal sigurtà u sens ta' ġustizzja), u tkun ċar mal-utenti tiegħek dwar limiti magħrufa u tipi ta' fallimenti potenzjali. Dettalji ta' libreriji miftuħa għall-pubbliku għall-ittestjar tas-sigurtà jinsabu fis-sezzjoni ta' iktar qari fl-aħħar ta' dan id-dokument.

Agħmilha faċli għall-utenti biex jagħmlu li huwa xieraq

Tirrikonoxxi li kull setting jew għażla ta' konfigurazzjoni ġodda għandha tkun assessjata konġuntament mal-beneficcju kummerċjali li jkollha, u kwalunkwe riskju ta' sigurtà li tintroduċi. Idealment, l-iktar setting sikur ikun integrat fis-sistema bħala l-uniku għażla. Meta l-konfigurazzjoni hija neċessarja, l-għażla 'default' għandha tkun sikura kontra theddid komuni (jiġifieri sikura 'by default'). Inti tapplika kontrolli biex tevita l-użu jew l-ingaġġ tas-sistema tiegħek b'mod malizzjuż.

Tipprovdi lill-utenti bi gwida dwar użu tajjeb tal-mudell jew sistema tiegħek, li tinkludi attenzjoni ikbar fuq il-limitazzjonijiet u modi possibbli ta' falliment. Tgħid ċarament lill-utenti liema aspetti ta' sigurtà huma responsabbli għalihom, u tkun trasparenti dwar fejn (u kif) l-informazzjoni tagħhom tista' tiġi użata, aċċessata jew maħżuna (per eżempju, jekk tkun użata bħala mudell għat-taħriġ mill-ġdid, jew riveduta minn impjegati jew imsieħba).

4. Operazzjoni u manutenzjoni sikuri

Din il-parti fiha linjigwida li japplikaw għall-istadju tal-**operazzjoni u manutenzjoni sikuri** tal-iżvilupp taċ-ċiklu tal-ħajja tas-sistema tal-AI. Tipprovdi linjigwida fuq azzjonijiet partikolarment rilevanti ladarba s-sistema tiġi ingaġġata, inkluż illoggjar u moniteraġġ, immaniġġjar tal-aġġornamenti u il-qsim tal-informazzjoni.

Issorvelja l-imġieba tas-sistema tiegħek

Inti tqis l-outputs u kif jopera l-mudell u s-sistema tiegħek biex tkun tista' tosserva tibdil f'daqqa u gradwali fl-imġieba li taffetwa s-sigurtà. Inti tista' żżomm kont u tidentifika l-possibilità ta' indħil u kompromessi, kif ukoll diffikultajiet naturali rigward data.

Issorvejla d-dħul tal-informazzjoni fis-sistema tiegħek

F'konformità mar-rekwiżiti tal-privatezza u l-protezzjoni tad-data, inti tissorvelja u tilloggja inputs fis-sistema tiegħek (bħal rikjesti ta' inferenza, mistoqsijiet jew prompts) li jiffaċilitaw l-obbligu tal-konformità, reviżjoni, investigazzjoni u rimedju fil-każ ta' kompromess u użu ħażin. Dan jista' jinkludi l-osservazzjoni espliċita ta' inputs barra mid-distribuzzjoni u/jew avversarji, inklużi dawk li għandhom l-għan li jisfruttaw il-passi fil-preparazzjoni tad-data (bħal qtugħ jew tibdil fil-qisien tal-istampi) (cropping and resizing for images).

Segwi sigurtà 'by design' fl-aġġornamenti

Tinkludi aġġornamenti awtomatiċi 'by default' f'kull prodott u tuża proċeduri tal-aġġornament sikuri, modulari għad-distribuzzjoni tagħhom. Il-proċessi tal-aġġornament tiegħek (inklużi sistemi tal-ittestjar u evalwazzjoni) jirriflettu l-fatt li t-tibdil fid-data, mudelli u prompts jistgħu imexxu għal-tibdil fl-imġieba tas-sistema (per eżempju, li titratta aġġornamenti kbar bħala verżjonijiet ġodda). Tissaportja lill-utenti biex jevalwaw u jirrispondu għat-tibdil fil-mudell) per eżempju billi tipprovdilhom aċċess bikri u API's verżjonati).

Iġbor u aqsam dak li titgħallem

Tipparteċipa f'kommunitajiet li jaqsmu informazzjoni ma' xulxin, u b'hekk tikkollabora mal-ekosistema globali tas-settur, l-akkademja u l-gvernijiet biex taqsam l-aħjar prattiċi kif hu xieraq. Iżżomm miftuħa l-kommunikazzjoni għall-feedback rigward is-sigurtà tas-sistema, kemm internament kif esternament fl-organizzazzjoni tiegħek, inkluż li tipprovdi l-kunsens lir-riċerkaturi biex jagħmlu riċerka u jirrapurtaw vulnerabbilitajiet. Fejn hemm bżonn, tiġbed l-attenzjoni tal-kommunità iktar wiesgħa, per eżempju tippubblika bullettini li jkunu jindirizzaw vulnerabbilitajiet żvelati, inklużi enumerazzjonijiet dettaljati u kompluti komuni tal-vulnerabbilitajiet. Tieħu azzjoni biex tnaqqas u tirrimedja problemi malajr u kif xieraq.

Aqra iktar

Żvilupp tal-AI

Il-Prinċipji tas-sigurtà fit-tagħlim tal-magni (ML)
Il-gwida dettaljata tal-NCSC dwar l-iżvilupp, l-ingaġġ jew l-operat ta' sistema b'fattur tal-ML.

Sigurtà mid-Disinn inniffsu - Bidla fil-Bilanċ tar-Riskju fis-Sigurtà Diġitali. Prinċipji u Perspettivi għal Software li huwa Sigur mid-Disinn Innifsu
Miktub konġunt minn CISA, NCSC u aġenziji oħra, din il-gwida tiddeskrivi kif il-manifatturi tas-sistemi tas-software, l-AI inkluża, għandha tieħu passi biex tikkunsidra s-sigurtà fl-istadju tad-disinn fl-iżvilupp tal-prodott, u tbiegħ prodotti li jkunu siguri kif jinxtraw.

Tħassib fis-sigurtà tal-AI fil-Qosor
Miġjub mill-Uffiċju Federali Ġermaniż għas-Sigurtà fl-Informazzjoni (BSI), dan id-dokument jipprovdi introduzzjoni fl-attakki potenzjali fuq is-sistemi taħt-tagħlim tal-magni u difiżi potenzjali kontra dawn l-attakki.

Il-Prinċipji Internazzjonali ta' Hiroshima li jiggwidaw il-Proċess (Hiroshima Process International Guiding Principles) għall-Organizzazzjonijiet li Jiżviluppaw Sistemi tal-AI Avvanzati u l-Kodiċi tal-Kondotta tal- Proċess Internazzjonali Hiroshima għall-Organizzazzjonijiet li Jiżviluppaw Sistemi tal-AI Avvanzati
Dawn id-dokumenti prodotti bħala parti mill Proċess tal-AI ta' Hiroshima G7, jipprovdu gwida għall-organizzazzjonijiet li jipprovdu l-iktar sistemi avvanzati tal-AI, inklużi l-iktar mudelli fundaturi avvanzati u sistemi ġenerattivi tal-AI bil-għan li jippromwovu globalment sistemi tal-AI siguri u ta' min jafdahom.

AI Verifika
L-għodda informattiva ta' Singapore tal-Qafas u Software tal-Ittestjar tal-Governanza tal-AI li jivvalida l-operat tas-sistemi tal-AI kontra sett ta' prinċipji rikonoxxuti globalment permezz ta' testijiet standard.

Qafas ta' Bosta Livelli għall-Prattiċi Tajba fis-Sigurtà Diġitali tal-AI — ENISA (europa.eu)
Qafas li jiggwida l-Awtoritajiet Nazzjonali Kompetenti u Stakeholders fl-AI dwar il-passi li għandhom isegwu biex jassiguraw is-sistemi tal-AI tagħhom, l-operazzjonijiet u il-proċessi.

ISO 5338: Proċessi taċ-ċiklu tal-ħajja tas-sistema tal-AI (Tat analiżi)
Sett ta' proċessi u kunċetti assoċjati għad-deskrizzjoni taċ-ċiklu tal-ħajja tas-sistemi tal-AI ibbażati fuq tagħlim tal-magni u sistemi ewristiċi.

AI Cloud Service Compliance Criteria Catalogue (AIC4)
Katalog ta' BSI li jipprovdi kriterji speċifiċi tal-AI, li jiffaċilità l-evalwazzjoni tas-sigurtà ta' servizz tal-AI matul iċ-ċiklu tal-ħajja tagħha.

NIST IR 8269 (Draft) Tassonomija u Terminoloġija ta'Tagħlim tal-Magni Avversarju
Sett ta' proċessi u kunċetti assoċjati għad-deskrizzjoni taċ-ċiklu tal-ħajja tas-sistemi tal-AI ibbażati fuq tagħlim tal-magni u sistemi ewristiċi.

MITRE ATLAS
Bażi ta'Tagħrif dwar tattiċi u teknika avversarji, u studji ta' każi għas-sistemi taħt-tagħlim tal-magni (ML), mudellati u marbutin mal-Qafas MITRE ATT&CK.

Ħarsa ġenerali lejn Riskji Katastrofiċi għall-AI (2023)
Miġjub miċ-Ċentru għas-Sigurtà tal-AI, dan id-dokument jissettja oqsma tar-riskju fir-rigward tal-AI.

Mudelli Kbar fil-Lingwa: Opportunitajiet u Riskji għall-Industrija u l-Awtoritajiet
Dokument miġjub minn BSI għal kumpaniji, awtoritajiet u żviluppaturi li jridu jitgħallmu iktar dwar l-opportunitajiet u riskji fl-iżvilupp, ingaġġ u/jew użu ta' LLMs.

Proġetti minn sorsi pubbliċi li jgħinuk tittestja s-sigurtà tal-mudelli tal-AI jinkludu:

- Adversarial Robustness Toolbox (IBM)
- CleverHans (Università ta' Toronto)
- TextAttack (Università ta' Virginja)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft)
- AI Verifika (Infocomm Awtorità tal-Iżvilupp tal-Midja, Singapore)

Sigurtà Diġitali

L-Għanijiet ta' CISA fl-Operat tas-Sigurtà Diġitali
Sett komuni ta' protezzjonijiet li l-entitajiet kollha kritiċi tal-infrastruttura għandhom jimplimentaw biex inaqqsu b'mod sew il-probbabilità u l-impatt ta' riskji magħrufa u teknika avversarja.

Qafas NCSC CAF
Qafas NCSC CAF Il-Qafas tal-Assessjar Diġitali (Cyber Assessment Framework) (CAF) jipprovdi gwida lill-organizzazzjonijiet dwar servizzi u attivitajiet li huma essenzjalment importanti.

Il-Qafas tas-Supply Chain tas-Sigurtà MITRE
Qafas għall-evalwazzjoni tal-forniturі u provvedituri tas-servizzi fis-supply chain.

Immaniġġjar tar-Riskju

Qafas minn NIST għall-Immaniġġjar tar-Riskju tal-AI (AI RMF)
Dan jipproponi kif wieħed jimmaniġġja riskji soċjo-tekniċi għall-individwu, organizzazzjonijiet, u soċjetajiet assoċjati b'mod uniku ma' AI.

ISO 27001: Sigurtà tal-Informazzjoni, sigurtà diġitali u protezzjoni tal-privatezza
Dan l-istandard jipprovdi gwida lill-organizzazjonijiet dwar l-istabbiliment, implimentazzjoni u manutenzjoni ta' sistema ta' immaniġġjar tas-sigurtà tal-informazzjoni.

ISO 31000: Immaniġġjar tar-Riskji
Standard internazzjonali li jipprovdi organizzazzjonijiet b'linjigwida u prinċipji għall-immaniġġjar tar-riskji ġewwa l-organizzazzjonijiet.

NCSC - Gwida fl-Immaniġġjar tar-Riskji
Din il-gwida tgħin dawk li jieħdu ħsieb is-sigurtà diġitali biex jifhmu u jimmaniġġjaw aħjar ir-riskji fis-sigurtà diġitali li jaffetwaw l-organizzazzjonijiet tagħhom.

Noti

1. Hawn definit bħala persuna, awtorità pubblika, aġenzija jew korp ieħor li jiżviluppa sistema ta' AI (jew li għandhom sistema ta' AI żviluppata) u li jpoġġi dik is-sistema fis-suq jew servizz taħt isimha jew marka kummerċjali
2. Għal iktar informazzjoni dwar sigurtà, ara Secure by Design ta' CISA paġna tal-web u gwida Bidla fil-Bilanċ tar-Riskju fis-Sigurtà Diġitali: Prinċipji u Perspettivi għal Software li huwa Secure by Design
3. Kontra ta' perspettivi tal-AI li mhumiex ML bħal sistemi bbażati fuq regoli
4. CEPS jiddeskrivi sebà tipi differenti ta' interazzjoni fl-iżvilupp tal-AI fil-pubblikazzjoni tagħhom 'Rikonċiljament tal- AI Value Chain mal-"Att tal-Intelliġenza Artifiċjali" tal- EU
5. ISO/IEC 22989:2022(en) jiddefinixxi dan bħala 'element funzjonali li jibni sistema ta' AI'
6. NIST huwa ikarigat biex jipproduċi linjigwida (u jagħmel azzjonijiet oħra) biex javvanza l-iżvilupp u l-użu sikur u fiduċjuż tal-Intelliġenza Artifiċjali (AI). Ara r-Responsabbilitajiet ta' NIST fl-Ordni Eżekuttiva ta' Ottubru 30, 2023
7. Iktar informazzjoni fuq il-mudellar fit-theddid diġitali jinsab fi ħdan il-Fundazzjoni OWASP
8. Ara MITRE ATLAS Adversarial Machine Learning 101
9. GitHub: RCE PoC for Tensorflow li jużż saff malizzjuż Lambda
10. SLSA: 'Protezzjoni tal-integrità tal-prodott artifiċjali mas-supply chain ta' kwalunkwe software'
11. METI (Japanese Ministry of Economy, Trade and Industry, 2023), 'Gwida ta' Introduzzjoni tal-Abboz ta' materjali tas-Software (SBOM) għall-Immaniġġjar tas-Software'
12. Riċerka Google Tagħlim tal-Magni (ML) L-Imgħax għoli tal-Karta tal-Kreditu tad-Dejn Tekniku
13. Tramèr et al 2016, Serq ta' Mudelli tat-Tagħlim tal-Magni permezz ta' API's imbassra
14. Boenisch, 2020, Attakkii kontra il-Privatezza tat-Tagħlim tal-Magni (L-ewwel parti): Model Inversion Attakki ta' Inversjoni tal-Mudelli permezz tal-Qafas IBM-ART
15. Ċentru Nazzjonali tas-Sigurtà Diġitali, Iddisinja u ibni Infrastruttura Prinċipali Pubblika mizmuma privatament