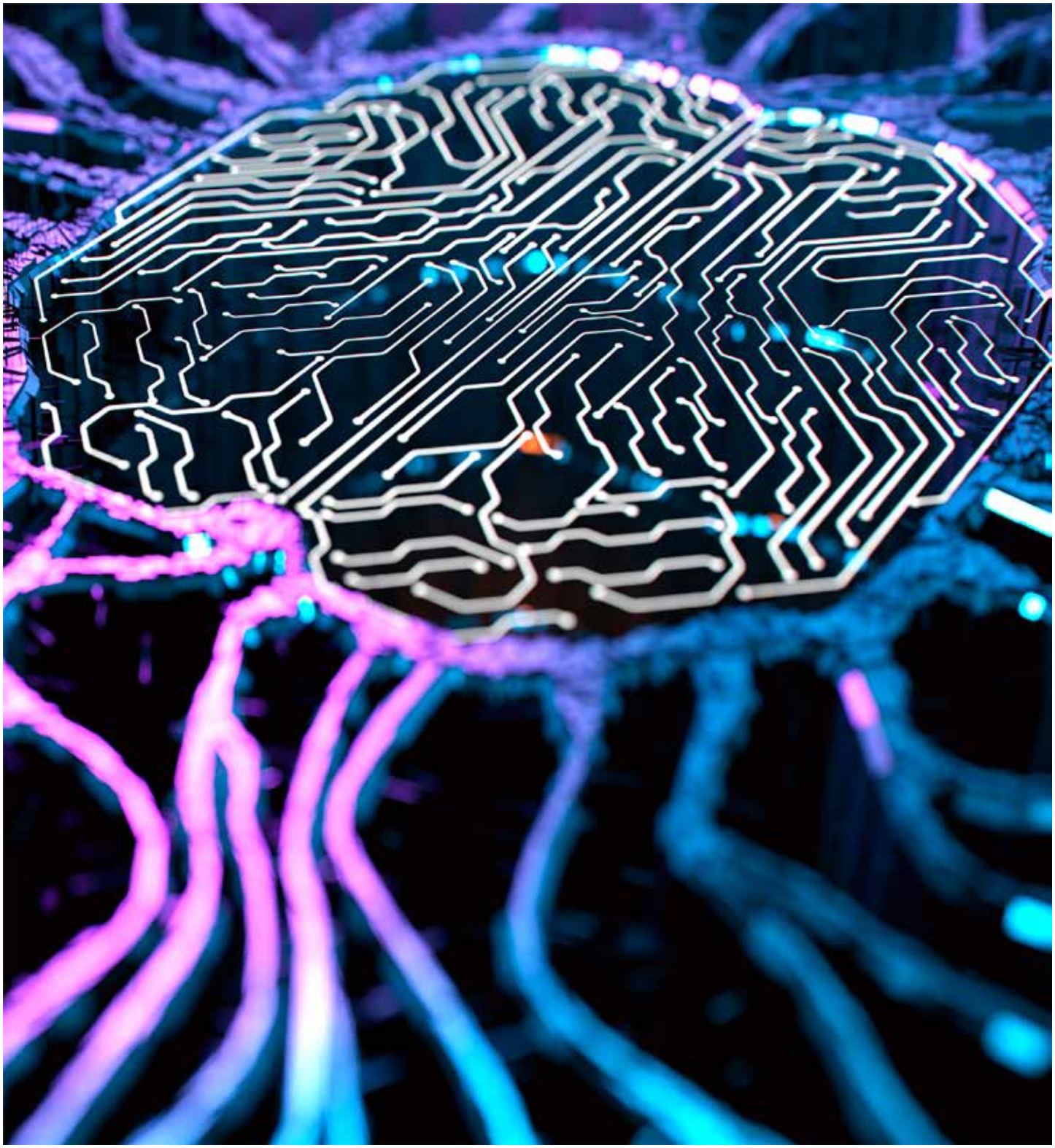
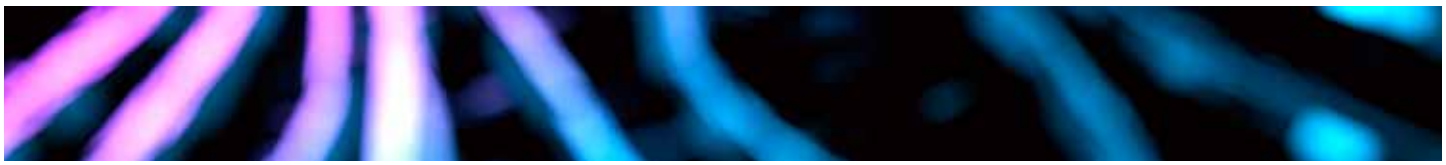
Mga patnubay para sa ligtas na paglikha sa sistema ng AI





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

Tungkol sa dokumentong ito

Ang dokumentong ito ay inilathala ng Pambansang Sentro ng Seguridad sa Cyber ng UK (NCSC), ang Ahensya ng Seguridad sa Cyber at Seguridad sa Imprastraktura ng US (CISA), at ang mga sumusunod na internasyonal na kasosyo:

- Pambansang Ahensya ng Seguridad (NSA)
- Federal Bureau of Investigations (FBI)
- Sentro ng Seguridad sa Cyber ng Australia (ACSC)
- Sentro ng Cyber Security ng Canada (CCCS)
- Pambansang Sentro ng Cyber Security ng New Zealand (NCSC-NZ)
- CSIRT ng Pamahalaang Chile
- Pambansang Ahensya ng Seguridad sa Cyber at Impormasyon ng Czechia (NUKIB)
- Information System Authority ng Estonia (RIA) at Pambansang Sentro ng Seguridad sa Cyber ng Estonia (NCSC-EE)
- Ahensya ng Seguridad sa Cyber ng Pransya (ANSSI)
- Tanggapang pang Pederal para sa seguridad ng impormasyon ng Germany (BSI)
- Pambansang Cyber Directorate ng Israel (INCD)
- Pambansang Ahensya ng Seguridad sa Cyber ng Italya (ACN)
- Pambansang Sentro ng Incident Readiness at Stratehiya sa Seguridad sa Cyber ng Japan (NISC)
- Secretariat sa Polisiya ng Science, Teknolohiya at Innovation ng Gabinete ng Japan
- Pambansang Ahensya ng Information Technology Development ng Nigeria (NITDA)
- Pambansang Sentro ng Seguridad sa Cyber ng Norway (NCSC-NO)
- Ministro ng Digital Affairs ng Poland
- NASK National Research Institute ng Poland (NASK)
- National Intelligence Service ng Republika ng Korea (NIS)
- Ahensya ng Seguridad sa Cyber ng Singapore (CSA)

Mga Pasasalamat

Ang mga sumusunod na organisasyon ay nag-ambag sa pagbuo ng mga alituntuning ito:

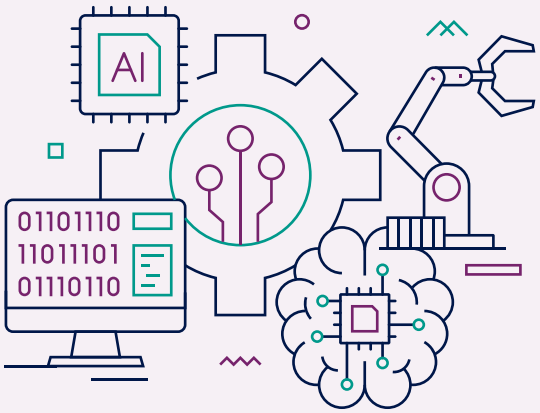
- Alan Turing Institute
- Anthropic
- Databricks
- Sentro ng Seguridad at Umuusbong na Teknolohiya ng Unibersidad ng Georgetown
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute ng Unibersidad ng Carnegie Mellon
- Sentro ng Stanford para sa Seguridad sa AI
- Programa sa Stanford Program sa Geopolitics, Technology at Governance

Pagtatatwa

Ang impormasyon sa dokumentong ito ay ibinibigay ng NCSC at ng mga organisasyong may-akda na hindi mananagot para sa anumang pagkawala, pinsala o pinsala sa anumang uri na dulot ng paggamit nito maliban kung kinakailangan ng batas. Ang impormasyon sa dokumentong ito ay hindi bumubuo o nagpapahiwatig ng pag-eendorso o rekomendasyon ng anumang third party na organisasyon, produkto, o serbisyo ng NCSC at mga ahensya ng awtorisasyon. Ang mga link at reference sa mga website at third party na materyales ay ibinibigay para sa impormasyon lamang at hindi kumakatawan sa pag-endorso o rekomendasyon ng naturang mga mapagkukunan sa iba.

Ang dokumentong ito ay makukuha sa TLP:CLEAR basis (https://www.first.org/tlp/).

Nilalaman



Executive summary

Ang dokumentong ito ay nagrerekomenda ng mga patnubay para sa mga nagbibigay ng anumang mga sistema na gumagamit ng artificial intelligence (AI), kung ang mga sistema na iyon ay ginawa mula sa simula o binuo sa ibabaw ng mga tool at serbisyo na ibinigay ng iba. Ang pagpapatupad ng mga alituntuning ito ay makakatulong sa mga provider na bumuo ng mga sistema ng AI na gumagana ayon sa nilalayon, magagamit kapag kinakailangan, at gumagana nang hindi naghahayag ng sensitibong datos sa mga hindi awtorisadong partido.

Ang dokumentong ito ay pangunahing nakatuon sa mga nagbibigay ng sistema ng AI na gumagamit ng mga modelong hino-host ng isang organisasyon, o gumagamit ng mga panlabas na application programming interface (API). Hinihimok namin ang **lahat** na stakeholder (kabilang ang mga data scientist, developer, manager, tagapasya at mga risk owner) na basahin ang mga patnubay na ito upang matulungan silang gumawa ng matalinong mga desisyon tungkol sa **disenyo**, **pag-unlad**, **deployment** at **operasyon** ng kanilang sistema ng AI.

Tungkol sa mga patnubay

Ang mga sistema ng AI ay may potensyal na magdala ng maraming benepisyo sa lipunan. Gayunpaman, para ganap na maisakatuparan ang mga pagkakataon ng AI, dapat itong mabuo, maitalaga at mapatakbo sa isang ligtas at responsableng paraan.

Ang mga sistema ng AI ay napapailalim sa mga bagong kahinaan sa seguridad na kailangang isaalang-alang kasama ng karaniwang mga banta sa seguridad sa cyber. Kapag ang bilis ng pag-unlad ay mataas – tulad sa kaso ng AI – ang seguridad ay kadalasang pangalawang pagsasaalang-alang. Ang seguridad ay dapat na isang pangunahing kinakailangan, hindi lamang sa yugto ng pag-unlad, kundi sa buong life cycle ng sistema.

Dahil dito, ang mga patnubay ay hinati-hati sa apat na pangunahing bahagi sa life cycle ng pagbuo ng sistema ng AI: **ligtas na disenyo**, **ligtas na pagbuo**, **ligtas na deployment**, at **ligtas na operasyon at pagmementena**. Para sa bawat seksyon, nagmumungkahi kami ng mga pagsasa-alang-alang at pagpapagaan na makakatulong na mabawasan ang pangkalahatang panganib sa isang proseso ng pagbuo ng sistemang AI ng organisasyon.

1. **Ligtas na disenyo**
   Ang seksyong ito ay naglalaman ng mga alituntunin na naaangkop sa yugto ng disenyo ng ikot ng buhay ng pagbuo ng AI system. Sinasaklaw nito ang pag-unawa sa mga panganib at pagmomodelo ng pagbabanta, pati na rin ang mga partikular na paksa at palitan na isasaalang-alang sa disenyo ng sistema at modelo.
2. **Ligtas na pagbuo**
   Ang seksyong ito ay naglalaman ng mga alituntunin na nalalapat sa yugto ng pagbuo ng sistema ng AI life cycle, kabilang ang seguridad sa uganayan sa pagtustos, dokumentasyon, at pag-aari asset at teknikal na pamamahala sa utang.
3. **Ligtas na deployment**
   Ang seksyong ito ay naglalaman ng mga alituntunin na nalalapat sa yugto ng pagtalaga ng sistema ng AI development lifecycle, kabilang ang pagprotekta sa imprastraktura at mga modelo mula sa kompromiso, pagbabanta o pagkawala, pagbuo ng mga proseso ng pamamahala ng insidente, at responsableng paglabas.
4. **Ligtas na operasyon at pagmementena**
   Ang seksyong ito ay naglalaman ng mga patnubay na nalalapat sa ligtas na operasyon at yugto ng pagpapanatili ng sistema sa AI development life cycle. Nagbibigay ito ng mga patnubay sa mga aksyon na partikular na nauugnay sa sandaling nakatalaga na ang isang sistema, kabilang ang pagkakalista at pagsubaybay, pamamahala ng pagsasapanahon (update) at pagbabahagi ng impormasyon.

Ang mga patnubay ay sumusunod sa isang 'secure by default' na diskarte, at malapit na nakahanay sa mga kasanayang tinukoy sa NCSC's Secure development and deployment guidance, NIST's Secure Software Development Framework , at 'mga prinsipyong secure by design' na inilathala ng CISA, NCSC at internasyonal na mga ahensya ng cyber. Pinaunahan nila ang:

- pag-aangkin sa responsibilidad sa mga resulta ng seguridad para sa mga kustomer
- pagyayakap sa radikal na katapatan at pananagutan
- ang pagbubuo ng istraktura ng organisasyon at pamumuno upang secure by design ang pangunahing priyoridad sa negosyo

Pagpapakilala

Ang mga sistema ng artificial intelligence (AI) ay may potensyal na magdala ng maraming benepisyo sa lipunan. Gayunpaman, para ganap na maisakatuparan ang mga pagkakataon ng AI, dapat itong mabuo, maitalaga at mapatakbo sa isang ligtas at responsableng paraan. Ang seguridad sa cyber ay isang kinakailangang paunang kondisyon para sa kaligtasan, katatagan, pribasiay, pagiging patas, bisa at pagiging maaasahan ng mga sistema ng AI.

Ngunit ang mga sistema ng AI ay napapailalim sa mga bagong kahinaan sa seguridad na kailangang isaalang-alang kasama ng mga karaniwang banta sa seguridad sa cyber. Kapag ang bilis ng pag-unlad ay mataas - tulad ng kaso sa AI - ang seguridad ay kadalasang maaaring pangalawang pagsasa-alang-alang. Ang seguridad ay dapat na isang pangunahing kinakailangan, hindi lamang sa yugto ng pag-unlad, ngunit sa buong life cycle ng sistema.

Inirerekomenda ng dokumentong ito ang mga patnubay para sa mga nagbibigay[1] ng anumang sistema na gumagamit ng AI, kung ang mga sistema na iyon ay ginawa mula sa simula o binuo sa ibabaw ng mga tool at serbisyong ibinigay ng iba pa. Ang pagpapatupad ng mga patnubay na ito ay makakatulong sa mga nagbibigay na bumuo ng mga sistema ng AI na gumagana ayon sa nilalayon, magagamit kapag kinakailangan, at gumagana nang hindi naghahayag ng sensitibong datos sa mga hindi awtorisadong partido.

Ang mga patnubay na ito ay dapat isaalang-alang kasabay ng itinatag na seguridad sa cyber, pamamahala sa peligro, at pinakamahusay na kasanayan sa pagtugon sa insidente. Sa partikular, hinihikayat namin ang mga provider na sundin ang mga prinsipyong 'secure by design'[2] na binuo ng Ahensya ng Seguridad sa Cyber at Imprastraktura ng US (CISA), Pambansang Sentro ng Seguridad sa Cyber ng UK (NCSC), at lahat ng aming mga internasyonal na kasosyo. Pinaunahan ng mga prinsipyo:

- Ang pag-angkin sa responsibilidad sa mga resulta ng seguridad para sa mga kustomer
- pagyakap sa radikal na katapatan at pananagutan
- pagbuo ng istraktura ng organisasyon at pamumuno upang secure by design ang pangunahing priyoridad sa negosyo.

Ang pagsunod sa mga prinsipyo ng 'secure by design' ay nangangailangan ng makabuluhang mapagkukunan sa buong life cycle ng isang sistema. Ito ay nangangahulugan na ang mga gumagawa ay dapat mamuhunan sa pagbibigay-priyoridad sa **mga katangian**, **mga mekanismo**, at **pagpapatupad** ng mga tool na nagpoprotekta sa mga kustomer sa bawat layer ng disenyo ng sistema, at sa lahat ng mga yugto ng development lifecycle. Kapag ginawa ito, maiiwasan ang magastos na muling pagdidisenyo sa ibang pagkakataon, pati na rin ang pagprotekta sa mga kustomer at ang kanilang datos sa
malapit na termino.

Bakit naiiba ang seguridad ng AI?

Sa dokumentong ito ginagamit namin ang 'AI' para partikular na sumangguni sa mga application ng machine learning (ML)[3]. Lahat ng uri ng ML ay nasa saklaw. Tinutukoy namin ang mga ML application bilang mga application na:

- kasangkot ang mga bahagi ng software (mga modelo) na nagbibigay-daan sa mga computer na kilalanin at dalhin ang konteksto sa mga pattern sa datos nang hindi kinakailangang tahasang i-program ng isang tao ang mga panuntunan.
- bumuo ng mga hula, rekomendasyon, o desisyon batay sa istatistikal na pangangatwiran

Bukod sa mga umiiral na banta sa seguridad sa cyber, ang mga sistema ng AI ay napapailalim sa mga bagong uri ng mga kahinaan. Ang terminong 'adversarial machine learning' (AML), ay ginagamit upang ilarawan ang pagsasamantala sa mga pangunahing kahinaan sa mga bahagi ng ML, kabilang ang hardware, software, mga daloy ng trabaho at mga ugnayan sa pagtustos. Ang AML ay nagbibigay-daan sa mga umaatake na magdulot ng mga hindi sinasadyang pag-uugali sa mga sistemang ML na maaaring kabilang ang:

- pagkaka-apekto sa pag-uuri o pagganap ng regression ng modelo
- na nagpapahintulot sa mga user na magsagawa ng mga hindi awtorisadong pagkilos
- pagkukuha ng sensitibong impormasyon ng modelo

Maraming paraan para makamit ang mga epektong ito, gaya ng agarang pag-atake ng injection sa domain ng large language model (LLM), o sadyang sirain ang datos ng pagsasanay o puna ng gumagamit (kilala bilang 'data poisoning').

Sino ang dapat bumasa sa dokumentong ito?

Ang dokumentong ito ay pangunahing nakatuon sa mga nagbibigay ng sistema ng AI, batay man sa mga modelong hino-host ng isang organisasyon o paggamit ng mga panlabas na application programming interface (API). Gayunpaman, hinihimok namin ang **lahat** na stakeholder (kabilang ang mga data scientist, gumagawa, manedyer, tagapagdesisyon at risk owners) na basahin ang mga patnubay na ito upang matulungan silang gumawa ng matalinong mga desisyon tungkol sa **disenyo**, **pagtatalaga** at **operasyon** ng kanilang mga machine learning na sistemang AI.

Ngunit, hindi lahat ng mga patnubay ay direktang naaangkop sa lahat ng organisasyon. Mag-iiba-iba ang antas ng pagiging sopistikado at mga paraan ng pag-atake depende sa kalaban na nagta-target sa sistema ng AI, kaya dapat isaalang-alang ang mga patnubay kasama ng mga kaso ng paggamit at profile ng pagbabanta ng iyong organisasyon.

Sino ang may pananagutan sa pagbuo ng ligtas na AI?

Kadalasan mayroong maraming mga aktor sa modernong AI supply chain. Ipinapalagay ng isang simpleng diskarte ang dalawang entity:

- ang 'provider' na responsable para sa data curation, algorithmic development, disenyo, deployment at pagmementena
- ang 'user', na nagbibigay ng mga input at tumatanggap ng mga output

Bagama't ginagamit ang diskarte ng provider-user na ito sa maraming application, nagiging hindi karaniwan[4], dahil maaaring tingnan ng mga provider na isama ang software, datos, modelo at/o malayuang serbisyong ibinibigay ng mga third party sa kanilang sariling mga sistema. Ang mga kumplikadong supply chain na ito ay nagpapahirap para sa mga end user na maunawaan kung saan nakasalalay ang responsibilidad para sa ligtas na AI.

Karaniwang walang sapat na pananaw at/o kadalubhasaan ang mga gumagamit (maging 'end user', o tagapagbigay na may kasamang external na bahagi ng AI[5]) upang lubos na maunawaan, suriin o matugunan ang mga panganib na nauugnay sa mga sistema na ginagamit nila. Dahil dito, alinsunod sa mga prinsipyo ng 'secure by design', **dapat na panagutin ng mga provider ng AI component ang mga resulta ng seguridad ng mga gumagamit sa ibaba ng ugnayan sa pagtustos.**

Dapat ipatupad ng mga nagbibigay ang mga kontrol sa seguridad at pagpapagaan kung posible sa loob ng kanilang mga modelo, pipeline at/o sistema, at kung saan ginagamit ang mga setting, ipatupad ang pinakaligtas na opsyon bilang default. Kung saan ang mga panganib ay hindi maaaring mabawasan, ang nagbibigay ay dapat na responsable para sa:

- pagpapaalam sa mga user sa ibaba ng ugnayan sa pagtustos ng mga panganib na tinatanggap nila at (kung naaangkop) ng sarili nilang mga gumagamit
- pagpapayo sa kanila kung paano gamitin ang bahagi nang ligtas

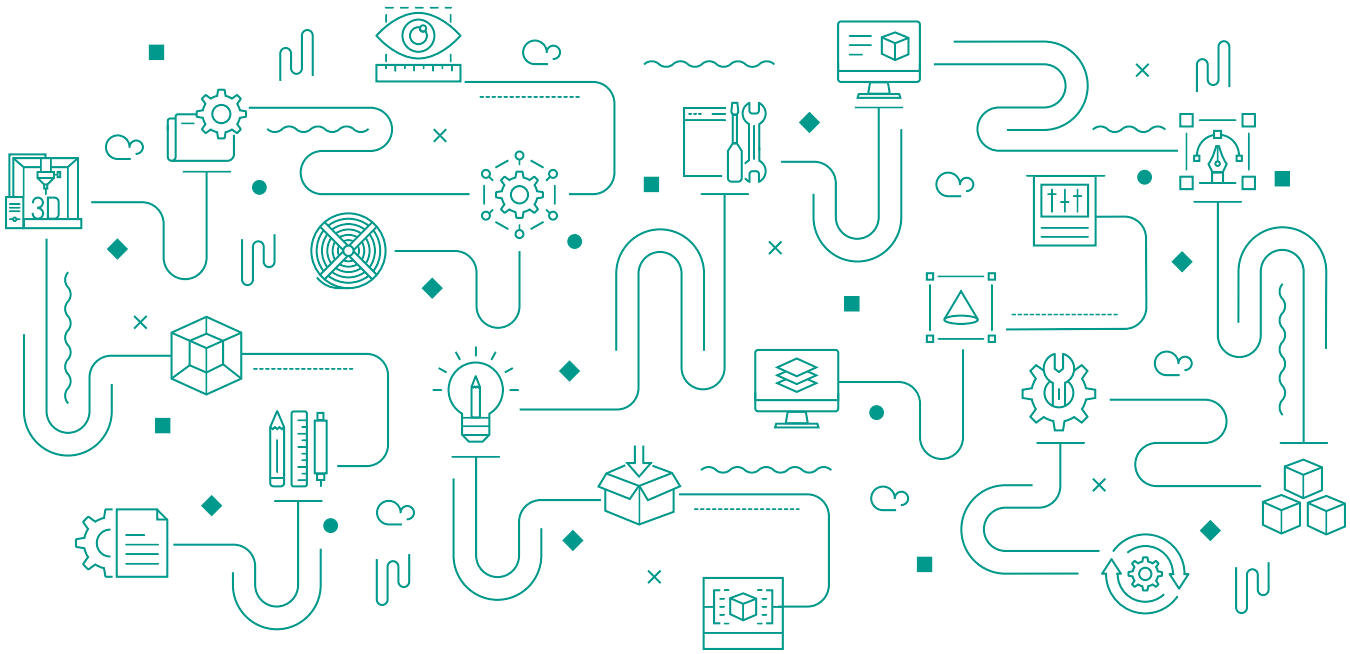
Kung ang kompromiso sa sistema ay maaaring humantong sa nakikita o malawakang pisikal o reputasyon na pinsala, malaking pagkawala ng mga operasyon ng negosyo, pagtagas ng sensitibo o kumpidensyal na impormasyon at/o mga legal na implikasyon, ang mga panganib sa seguridad sa cyber ng AI ay dapat ituring bilang **kritikal.**

Mga patnubay para sa ligtas na pagbuo ng mga sistema ng AI

Ang mga patnubay ay hinati-hati sa apat na pangunahing bahagi sa loob ng life cycle ng pagbuo ng sistema ng AI: **ligtas na disenyo**, **ligtas na pagbuo**, **ligtas na pagtalaga**, at **ligtas na operasyon at pagmementena**. Para sa bawat lugar, nagmumungkahi kami ng mga pagsasaalang-alang at pagpapagaan na makakatulong na mabawasan ang pangkalahatang panganib sa proseso ng pagbuo ng sistemang AI ng organisasyon.

Ang mga patnubay itinakda sa dokumentong ito ay malapit na nakahanay sa mga kasanayan sa life cycle ng pagbuo ng software na tinukoy sa:

- Ligtas na pag-unlad at gabay sa pagtatalaga ng NCSC
- ang National Institute of Standards and Technology (NIST) Secure Software Development Framework (SSDF)[6]



1. Ligtas na disenyo

Naglalaman ang seksyong ito ng mga patnubay na naaangkop sa yugto ng **disenyo** ng life cycle ng pagbuo ng sistemang AI. Sinasaklaw nito ang pag-unawa sa mga panganib at pagmomodelo ng pagbabanta, pati na rin ang mga partikular na paksa at palitan na dapat isaalang-alang sa disenyo ng sistema at modelo.

Itaas ang kamalayan ng kawani sa mga banta at panganib

Nauunawaan ng mga may-ari ng sistema at nakatataas na pinuno ang mga banta upang maging ligtas ang AI at ang kanilang mga pagpapagaan. Ang iyong mga data scientist at developer ay nagpapanatili ng kamalayan sa mga nauugnay na banta sa seguridad at mga mode ng pagkabigo at tumutulong sa mga risk owner na gumawa ng matalinong mga desisyon. Nagbibigay ka sa mga user ng gabay sa mga natatanging panganib sa seguridad na kinakaharap ng mga sistema ng AI (halimbawa, bilang bahagi ng karaniwang pagsasanay sa InfoSec) at sanayin ang mga developer sa mga ligtas na diskarte sa coding at ligtas at responsableng mga kasanayan sa AI.

I-modelo ang mga banta sa iyong sistema

Bilang bahagi ng iyong proseso ng pamamahala sa peligro, naglalapat ka ng isang holistic na proseso upang masuri ang mga banta sa iyong sistema, na kinabibilangan ng pag-unawa sa mga potensyal na epekto sa sistema, mga user, organisasyon, at mas malawak na lipunan kung ang isang bahagi ng AI ay nakompromiso o kumikilos nang hindi inaasahan[7]. Kasama sa prosesong ito ang pagtatasa sa epekto ng mga banta na partikular sa AI[8] at pagdodokumento sa iyong paggawa ng desisyon.

Kinikilala mo na ang sensitivity at mga uri ng datos na ginagamit sa iyong sistema ay maaaring makaimpluwensya sa halaga nito bilang isang target sa isang umaatake. Dapat isaalang-alang ng iyong pagtatasa na maaaring lumaki ang ilang banta habang ang sistema ng AI ay lalong tumitingin bilang mga target na matataas ang halaga, at dahil ang AI mismo ay nagbibigay-daan sa mga bago, mga awtomatikong attack vector.

Idisenyo ang iyong sistema para sa seguridad pati na rin ang functionality at performance

Ikaw ay may tiwala na ang gawain ay pinakaangkop na tinutugunan gamit ang AI. Kapag natukoy na ito, tinatasa mo ang pagiging angkop ng iyong mga pagpipilian sa disenyo na partikular sa AI. Isinasaalang-alang mo ang iyong modelo ng pagbabanta at nauugnay na pagpapagaan ng seguridad kasama ng functionality, karanasan ng user, kapaligiran sa pag-deploy, pagganap, pagtitiyak, pangangasiwa, etikal at legal na mga kinakailangan, bukod sa iba pang mga pagsasaalang-alang. Halimbawa:

- isinasaalang-alang mo ang seguridad ng supply chain kapag pumipili kung bubuo in-house o gagamit ng mga panlabas na bahagi, halimbawa:
    - ang iyong piniling magsanay ng bagong modelo, gumamit ng kasalukuyang modelo (mayroon man o walang fine-tuning) o mag-access ng modelo sa pamamagitan ng panlabas na API ay naaangkop sa iyong mga kinakailangan
    - ang iyong piniling magtrabaho kasama ang isang panlabas na tagapagbigay ng modelo ay may kasamang isang angkop na pagsusuri sa sariling postura ng seguridad ng provider na iyon
    - kung gumagamit ng panlabas na library, kumpletuhin mo ang due diligence evaluation (halimbawa, upang matiyak na ang library ay may mga kontrol na pumipigil sa sistema na mag-load ng mga hindi pinagkakatiwalaang modelo nang hindi kaagad inilalantad ang kanilang mga sarili sa arbitrary na pagpapatupad ng code[9])
    - nagpapatupad ka ng pag-scan at paghihiwalay/sandboxing kapag nag-import ng mga third-party na modelo o serialized na timbang, na dapat ituring bilang hindi pinagkakatiwalaang third-party na code at maaaring paganahin ang remote code execution

- kung gumagamit ng mga panlabas na API, ilalapat mo ang mga naaangkop na kontrol sa datos na maaaring ipadala sa mga serbisyo sa labas ng kontrol ng iyong organisasyon, tulad ng pag-aatas sa mga user na mag-log in at magkumpirma bago magpadala ng potensyal na sensitibong impormasyon
- naglalapat ka ng mga naaangkop na pagsusuri at paglilinis ng datos at mga input; kabilang dito ang kapag nagsasama ng feedback ng user o patuloy na pag-aaral ng datos sa iyong modelo, na kinikilala na ang datos ng pagsasanay ay tumutukoy sa gawi ng sistema

- isinasama mo ang pagbuo ng AI software system sa kasalukuyang ligtas na pag-unlad at mga pinakamahuhusay na kasanayan sa pagpapatakbo; lahat ng elemento ng sistema ng AI ay nakasulat sa naaangkop na mga kapaligiran gamit ang mga kasanayan sa coding at mga wika na nagbabawas o nag-aalis ng mga kilalang klase ng mga kahinaan kung saan posible

- kung kailangan ng mga bahagi ng AI na mag-trigger ng mga aksyon, halimbawa ang pag-amyenda sa mga file o pagdidirekta ng output sa mga panlabas na sistema, maglalapat ka ng naaangkop na mga paghihigpit sa mga posibleng aksyon (kabilang dito ang external AI at hindi AI na fail-safe kung kinakailangan)

- ang mga desisyon sa pakikipag-ugnayan ng user ay ipinaaalam ng mga panganib na partikular sa AI, halimbawa:

  - ang iyong sistema ay nagbibigay sa mga user ng magagamit na mga output nang hindi nagpapakita ng mga hindi kinakailangang antas ng detalye sa isang potensyal na umaatake
  - kung kinakailangan, ang iyong sistema ay nagbibigay ng mga epektibong guardrail sa paligid ng mga output ng modelo
  - kung nag-aalok ng API sa mga external na kustomer o collaborator, ilalapat mo ang mga naaangkop na kontrol na nagpapagaan ng mga pag-atake sa sistema ng AI sa pamamagitan ng API
  - isinasama mo ang pinakaligtas na mga setting sa sistema bilang default
  - ilalapat mo ang mga prinsipyo ng hindi bababa sa pribilehiyo upang limitahan ang paggamit sa functionality ng isang sistema
  - ipinapaliwanag mo ang mga mas mapanganib na kakayahan sa mga user at hinihiling sa mga user na mag-opt in upang gamitin ang mga ito;
    nagsasabi ka ng mga ipinagbabawal na kaso ng paggamit, at, kung posible, ipaalam sa mga gumagamit ang mga alternatibong solusyon

Isaalang-alang ang mga benepisyo sa seguridad at trade-off kapag pinipili ang iyong modelo ng AI

Ang iyong pagpili ng modelo ng AI ay kasangkot sa pagbabalanse ng isang hanay ng mga kinakailangan. Kabilang dito ang pagpili ng arkitektura ng modelo, pagsasaayos, datos ng pagsasanay, algorithm ng pagsasanay at mga hyperparameter. Ang iyong mga desisyon ay nababatid ng iyong modelo ng pagbabanta, at regular na muling sinusuri habang umuunlad ang pananaliksik sa seguridad ng AI at umuusbong ang pag-unawa sa pagbabanta.

Kapag pumipili ng modelo ng AI, ang mga ito ang malamang na kasama sa iyong mga pagsasa-alang-alang, ngunit hindi limitado sa:

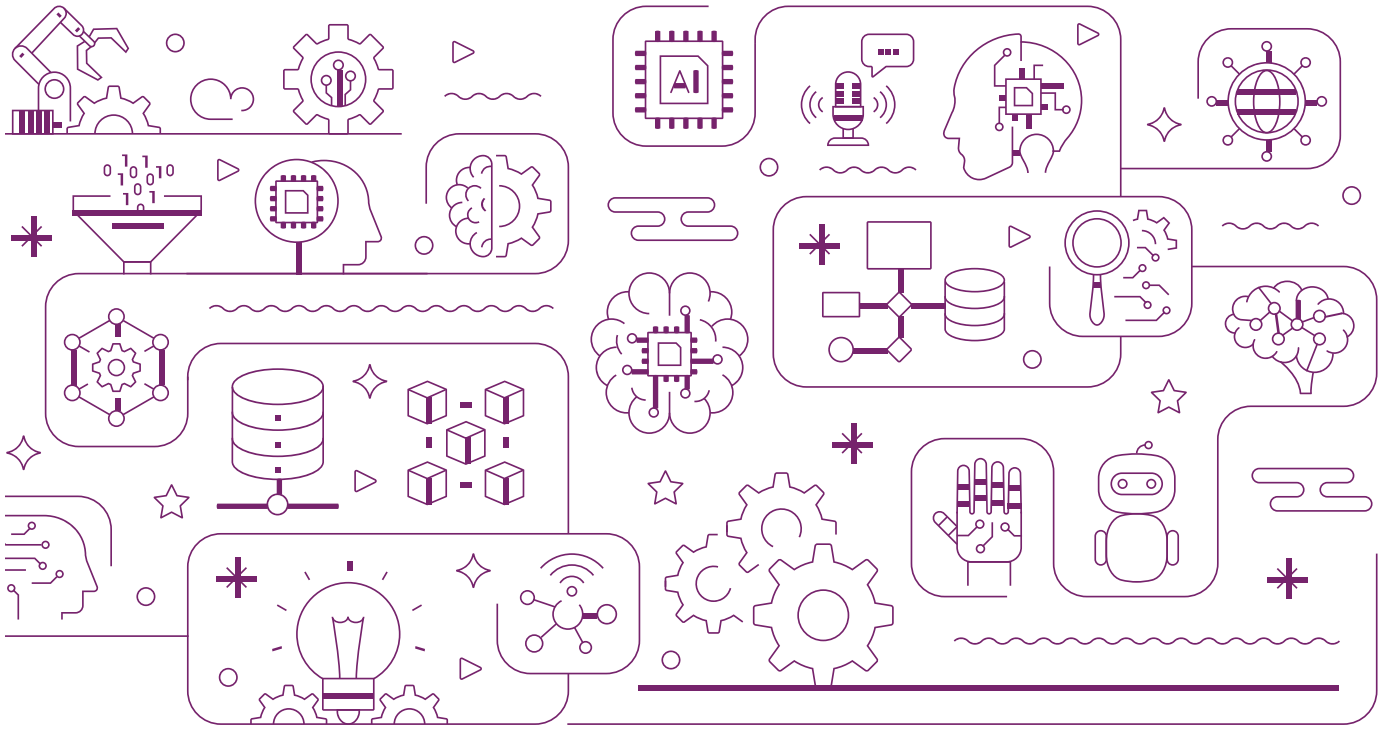
- pagiging kumplikado ng modelo na iyong ginagamit, iyon ay ang napiling arkitektura at bilang ng mga parameter; bukod sa iba pang mga salik, ang napiling arkitektura ng iyong modelo at bilang ng mga parameter ay makakaapekto sa kung gaano karaming datos ng pagsasanay ang kinakailangan nito at kung gaano ito katatag sa mga pagbabago sa input data kapag ginagamit

- ang kaangkupan ng modelo para sa iyong use case at/o pagiging posible na iakma ito sa iyong partikular na pangangailangan (halimbawa sa pamamagitan ng fine-tuning)

- ang kakayahang ihanay, bigyang-kahulugan at ipaliwanag ang mga output ng iyong modelo (halimbawa para sa pag-debug, pag-audit o pagsunod sa regulasyon); maaaring may mga benepisyo sa paggamit ng mas simple, mas transparent na mga modelo kaysa sa malaki at kumplikadong mga modelo na mas mahirap bigyang-kahulugan

- mga katangian ng (mga) dataset ng pagsasanay, kabilang ang laki, integridad, kalidad, sensitivity, edad, kaugnayan at pagkakaiba-iba
- ang halaga ng paggamit ng pagpapatigas ng modelo (tulad ng adversarial na pagsasanay), regularisasyon at/o mga diskarte sa pagpapahusay ng pagka-pribado
- ang pinagmulan at ugnayan ng pagtustos ng mga bahagi kabilang ang modelo o modelo ng pundasyon, datos ng pagsasanay at mga nauugnay na tool

Para sa higit pang impormasyon tungkol sa kung gaano karami sa mga salik na ito ang nakakaapekto sa mga resulta ng seguridad, sumangguni sa 'Mga Prinsipyo para sa Seguridad ng Machine Learning' ng NCSC, sa partikular Disenyo para sa seguridad (arkitektura ng modelo).



2. Ligtas na pagbuo

Naglalaman ang seksyong ito ng mga patnubay na naaangkop sa yugto ng **pagbuo** ng development lifecycle ng sistema ng AI, kabilang ang seguridad ng uganayan sa pagtustos, dokumentasyon, at pamamahala ng asset at teknikal na utang.

Gawing ligtas ang iyong supply chain

Sinusuri at sinusubaybayan mo ang seguridad ng iyong mga uganayan sa pagtustos ng AI sa buong life cycle ng isang sistema, at hinihiling mo sa mga supplier na sumunod sa parehong mga pamantayang inilalapat ng sarili mong organisasyon sa ibang software. Kung hindi makasunod ang mga supplier sa mga pamantayan ng iyong organisasyon, kumilos ka alinsunod sa iyong umiiral na mga patakaran sa pamamahala sa peligro.

Kung saan hindi ginawa in-house, nakakakuha ka at nagpapanatili ng ligtas at mahusay na dokumentado na mga bahagi ng hardware at software (halimbawa, mga modelo, datos, software library, module, middleware, frameworks, at panlabas na API) mula sa napatunayan na komersyal, open source, at iba pang mga third-party na gumagawa upang matiyak ang matatag na seguridad sa iyong mga sistema.

Handa ka nang mag-failover sa mga alternatibong solusyon para sa mission-critical system, kung hindi natutugunan ang mga pamantayan sa seguridad. Gumagamit ka ng mga mapagkukunan tulad ng Patnubay sa Supply Chain ng NCSC at mga framework gaya ng Supply Chain Levels para sa Software Artifacts (SLSA)[10] para sa pagsubaybay sa mga patotoo ng supply chain at ng mga software development life cycle.

Kilalanin, subaybayan at protektahan ang iyong mga asset

Nauunawaan mo ang halaga sa iyong organisasyon ng iyong mga asset na nauugnay sa AI, kabilang ang mga modelo, datos (kabilang ang puna ng gumamit), mga prompt, software, dokumentasyon, mga log at mga pagtatasa (kabilang ang impormasyon tungkol sa mga potensyal na hindi ligtas na kakayahan at mga mode ng pagkabigo), pagkilala kung saan kinakatawan ng mga ito ang makabuluhang pamumuhunan at kung saan ang pag-access sa mga ito ay nagbibigay-daan sa isang umaatake. Itinuturing mo ang mga log bilang sensitibong datos at nagpapatupad ka ng mga kontrol para protektahan ang pagiging kompidensiyal, integridad at availability ng mga ito.

Alam mo kung saan naninirahan ang iyong mga asset at nasuri at tinanggap mo ang anumang nauugnay na mga panganib. Mayroon kang mga proseso at tool upang subaybayan, patotohanan, kontrolin ang bersyon at gawing ligtas ang iyong mga asset, at maaaring ibalik sa isang kilalang mabuting kalagayan kung sakaling magkaroon ng kompromiso.

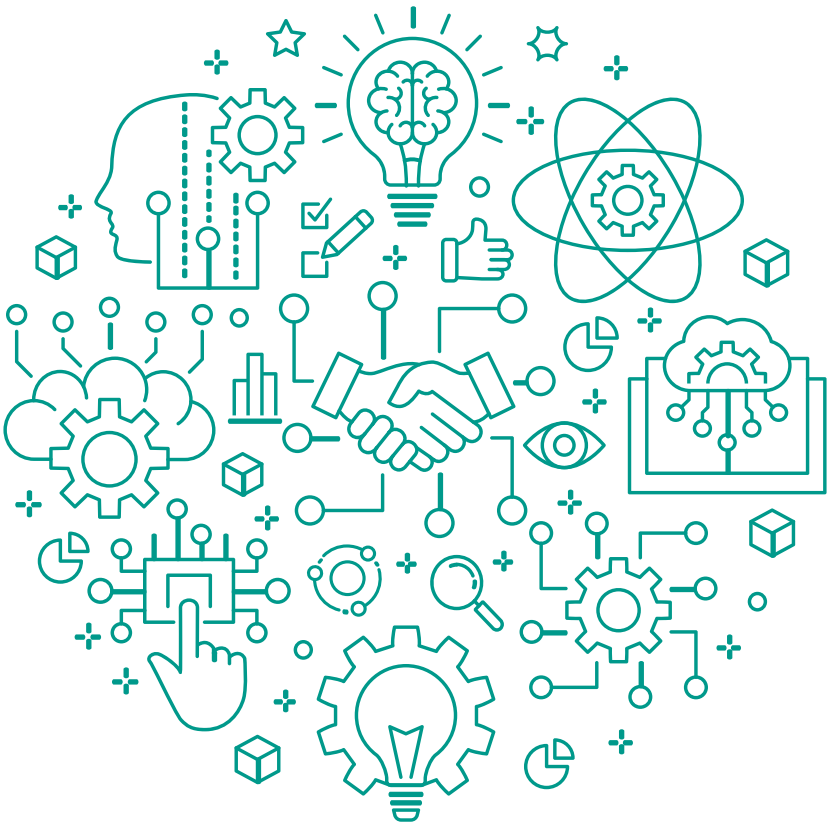
Mayroon kang mga proseso at kontrol sa lugar upang pamahalaan kung anong datos ang maaaring magamit ng mga sistema ng AI, at upang pamahalaan ang nilalamang nabuo ng AI ayon sa pagiging sensitibo nito (at ang pagiging sensitibo ng mga input na napunta sa pagbuo nito).

Idokumento ang iyong datos, mga modelo at mga senyas

Idodokumento mo ang paggawa, pagpapatakbo, at pamamahala sa life cycle ng anumang mga modelo, dataset at meta-o system-prompt. Kasama sa iyong dokumentasyon ang impormasyong nauugnay sa seguridad gaya ng mga pinagmumulan ng datos ng pagsasanay (kabilang ang fine-tuning na datos at feedback ng tao o iba pang operational na feedback), nilalayong saklaw at limitasyon, mga guardrail, cryptographic na mga hash o lagda, oras ng pagpapanatili, iminungkahing dalas ng pagsusuri at mga potensyal na mode ng pagkabigo. Kasama sa mga kapaki-pakinabang na istrukturang makakatulong sa paggawa nito ang mga model card, data card at software bill of materials (SBOMs). Ang paggawa ng komprehensibong dokumentasyon ay sumusuporta sa katapatan at pananagutan[11].

Pamahalaan ang iyong teknikal na utang

Tulad ng anumang sistema ng software, tinutukoy mo, sinusubaybayan at pinamamahalaan mo ang iyong 'teknikal na utang' sa buong lilfe cycle ng isang sistema ng AI (ang teknikal na utang ay kung saan ginawa ang mga desisyon sa engineering na kulang sa pinakamahuhusay na kagawian upang makamit ang mga panandaliang resulta, at sinasakripisyo ang matagalang benepisyo). Tulad ng utang sa pananalapi, ang teknikal na utang ay hindi likas na masama, ngunit dapat na pamahalaan mula sa pinakamaagang yugto ng pagbuo[12]. Kinikilala mo na ang paggawa nito ay maaaring maging mas mahirap sa isang konteksto ng AI kaysa para sa karaniwang software, at na ang iyong mga antas ng teknikal na utang ay malamang na mataas dahil sa mabilis na mga yugto ng pagbuo at kakulangan ng mahusay na itinatag na mga protocol at interface. Tinitiyak mo ang iyong mga plano sa life cycle (kabilang ang mga proseso sa pag-decommission ng mga sistema ng AI) na tinatasa, tinatanggap at pinapagaan ang mga panganib sa mga katulad na sistema sa hinaharap.



3. Ligtas na deployment

Naglalaman ang seksyong ito ng mga alituntunin na nalalapat sa yugto ng **pagtatalaga** ng development lifecycle ng sistema ng AI, kabilang ang pagprotekta sa imprastraktura at mga modelo mula sa kompromiso, pagbabanta o pagkawala, pagbuo ng mga proseso ng pamamahala ng insidente, at responsableng pagpapahayag.

Gawing ligtas ang iyong imprastraktura

Inilalapat mo ang mahusay na mga prinsipyo sa seguridad ng imprastraktura sa imprastraktura na ginagamit sa bawat bahagi ng life cycle ng iyong sistema. Maglalapat ka ng naaangkop na mga kontrol sa pag-access sa iyong mga API, modelo at datos, at sa kanilang pagsasanay at pagpoproseso ng mga pipeline, sa pananaliksik at pagpapaunlad pati na rin sa pag-deploy. Kabilang dito ang naaangkop na paghihiwalay ng mga kapaligiran na may sensitibong code o datos. Makakatulong din ito na mapagaan ang mga karaniwang pag-atake sa seguridad sa cyber na naglalayong magnakaw ng isang modelo o makapinsala sa pagganap nito.

Patuloy na protektahan ang iyong modelo

Maaaring muling buuin ng mga umaatake ang functionality ng isang modelo[13] o ang datos kung saan ito sinanay[14], sa pamamagitan ng direktang pag-access sa isang modelo (sa pamamagitan ng pagkuha ng mga timbang ng modelo) o hindi direkta (sa pamamagitan ng pagtatanong sa modelo sa pamamagitan ng isang aplikasyon o serbisyo). Ang mga umaatake ay maaari ring pakialaman ang mga modelo, datos o mga senyas sa panahon o pagkatapos ng pagsasanay, na ginagawang hindi mapagkakatiwalaan ang output.

Pinoprotektahan mo ang modelo at datos mula sa direkta at hindi direktang pag-access sa pamamagitan ng:

- pagpapatupad ng karaniwang mga pinakamahusay na kasanayan sa seguridad sa cyber
- pagpapatupad ng mga kontrol sa interface ng query upang makita at maiwasan ang mga pagtatangka na i-access, baguhin, at alisin ang kumpidensyal na impormasyon

Upang matiyak na makakapag-validate ng mga modelo ang mga consuming system, magkuwenta at magbahagi ng mga cryptographic na hash at/o mga lagda ng mga file ng modelo (halimbawa, mga timbang ng modelo) at mga dataset (kabilang ang mga checkpoint) sa sandaling nasanay ang modelo. Gaya ng nakasanayan sa cryptography, ang mahusay na pamamahala ng key ay mahalaga[15].

Ang iyong diskarte sa pagpapagaan ng confidentiality risk ay lubos na nakasalalay sa use case at modelo ng pagbabanta. Ang ilang mga application, halimbawa ang mga kinasasangkutan ng napakasensitibong data, ay maaaring mangailangan ng mga teoretikal na garantiya na maaaring mahirap o mahal na ilapat. Kung pwede, ang mga teknolohiyang nagpapahusay sa privacy (gaya ng differential privacy o homomorphic encryption) ay maaaring gamitin upang galugarin o tiyakin ang mga antas ng panganib na nauugnay sa mga konsumer, user at attacker na may access sa mga modelo at output.

Bumuo ng mga pamamaraan sa pamamahala ng insidente

Ang hindi maiiwasang mga insidente sa seguridad na nakakaapekto sa iyong mga sistema ng AI ay makikita sa iyong pagtugon sa insidente, mga plano sa paglaki at remedyo.. Ang iyong mga plano ay sumasalamin sa iba't ibang mga sitwasyon at regular na muling sinusuri habang nagbabago ang sistema at mas malawak na pananaliksik. Nag-iimbak ka ng mga kritikal na mapagkukunan ng digital ng kumpanya sa mga offline na backup. Ang mga tumugon ay sinanay upang tasahin at tugunan ang mga insidenteng nauugnay sa AI. Nagbibigay ka ng mataas na kalidad na mga log ng pag-audit at iba pang mga katangian ng seguridad o impormasyon sa mga kustomer at user nang walang dagdag na bayad, upang paganahin ang kanilang mga proseso ng pagtugon sa insidente.

Ilabas ang AI nang responsable

Maglalabas ka lang ng mga modelo, application o sistema pagkatapos na isailalim ang mga ito sa naaangkop at epektibong pagsusuri sa seguridad tulad ng benchmarking at red teaming (pati na rin ang iba pang mga pagsubok na wala sa saklaw para sa mga alituntuning ito, gaya ng kaligtasan o pagiging patas), at malinaw sa iyong mga user tungkol sa mga kilalang limitasyon o potensyal na mga mode ng pagkabigo. Ang mga detalye ng open-source na mga library ng pagsubok sa seguridad ay ibinibigay sa seksyon ng karagdagang pagbabasa sa dulo ng dokumentong ito.

Gawing madali para sa mga gumagamit na gawin ang mga tamang bagay

Kinikilala mo na ang bawat bagong setting o opsyon sa pagsasaayos ay susuriin kasabay ng benepisyo ng negosyo na nakukuha nito, at anumang mga panganib sa seguridad na ipinakilala nito. Mas mabuti kung ang pinakaligtas na setting ay isasama sa sistema bilang ang tanging opsyon. Kapag kailangan ang pagsasaayos, dapat na malawak na ligtas ang default na opsyon laban sa mga karaniwang banta (iyon ay, secure by default). Naglalapat ka ng mga kontrol upang maiwasan ang paggamit o pag-deploy ng iyong sistema sa mga nakakahamak na paraan.

Nagbibigay ka ng gabay sa mga user sa naaangkop na paggamit ng iyong modelo o sistema, na kinabibilangan ng pag-highlight ng mga limitasyon at mga potensyal na modo ng pagkabigo. Malinaw mong isinasaad sa mga user kung aling mga aspeto ng seguridad ang kanilang pananagutan, at malinaw kung saan (at paano) maaaring gamitin, i-access o iimbak ang kanilang datos (halimbawa, kung ginagamit ito para sa muling pagsasanay ng modelo, o sinusuri ng mga empleyado o kasosyo).

4. Ligtas na operasyon at pagmementena

Ang seksyong ito ay naglalaman ng mga alituntunin na nalalapat sa **ligtas na operasyon at pagmementena** na yugto ng lifecycle ng pagbuo ng sistema ng AI. Nagbibigay ito ng mga alituntunin sa mga aksyon na partikular na nauugnay kapag ang isang sistema ay naitalaga, kabilang ang paglista at pagsubaybay, pamamahala ng update at pagbabahagi ng impormasyon.

Subaybayan ang gawi ng iyong sistema

Sinusukat mo ang mga output at performance ng iyong modelo at sistema upang maobserbahan mo ang biglaan at unti-unting pagbabago sa pag-uugaling nakakaapekto sa seguridad. Maaari mong isaalang-alang at tukuyin ang mga potensyal na panghihimasok at kompromiso, pati na rin ang natural na daloy ng datos .

Subaybayan ang mga input ng iyong sistema

Alinsunod sa mga kinakailangan sa pagka-pribado at proteksyon ng datos, sinusubaybayan at inila-log mo ang mga input sa iyong sistema (tulad ng mga kahilingan sa inference, query o senyas) upang paganahin ang mga obligasyon sa pagsunod, pag-audit, pagsisiyasat at remediation sa kaso ng kompromiso o maling paggamit. Maaaring kabilang dito ang tahasang pagtuklas ng wala-sa-pamamahagi at/o mga adversarial input, kabilang ang mga naglalayong pagsamantalahan ang mga hakbang sa paghahanda ng datos (gaya ng pagpuputol (cropping) at pagbabago ng laki para sa mga larawan).

Sundin ang isang ligtas na diskarte sa disenyo sa mga update

Isasama mo ang mga awtomatikong pag-update bilang default sa bawat produkto at gumamit ng ligtas, modular na mga pamamaraan sa pagsapanahon upang ipamahagi ang mga ito. Ang iyong mga proseso sa pag-update (kabilang ang mga pagsubok at pagsusuri) ay sumasalamin sa katotohanan na ang mga pagbabago sa datos, mga modelo o mga senyas ay maaaring humantong sa mga pagbabago sa gawi ng sistema (halimbawa, tinatrato mo ang mga pangunahing update tulad ng mga bagong bersyon). Sinusuportahan mo ang mga user na suriin at tugonin ang mga pagbabago sa modelo (halimbawa sa pamamagitan ng pagbibigay ng paggamit sa preview at mga naka-bersyon na API).

Kolektahin at ibahagi ang mga natutunan

Lumalahok ka sa mga komunidad na nagbabahagi ng impormasyon, nakikipagtulungan sa buong pandaigdigang ecosystem ng industriya, akademya at pamahalaan upang ibahagi ang pinakamahusay na kasanayan kung naaangkop. Pinapanatili mo ang mga bukas na linya ng komunikasyon para sa puna tungkol sa seguridad ng sistema, sa parehong panloob at panlabas sa iyong organisasyon, kabilang ang pagbibigay ng pahintulot sa mga mananaliksik ng seguridad na magsaliksik at mag-ulat ng mga kahinaan. Kapag kinakailangan, ipaparating mo ang mga isyu sa mas malawak na komunidad, halimbawa ang pag-publish ng mga bulletin na tumutugon sa mga pagsisiwalat ng kahinaan, kabilang ang detalyado at kumpletong enumeration ng karaniwang kahinaan. Kumilos ka upang bawasan at ayusin ang mga isyu nang mabilis at naaangkop.

Karagdagang pagbabasa

Pag-unlad ng AI

Mga Prinsipyo para sa seguridad ng machine learning
Ang detalyadong patnubay ng NCSC sa pagbuo, pag-deploy o pagpapatakbo ng sistema na may bahagi na ML.

Secure by Design - Pagbabago sa Balanse ng Cybersecurity Risk: Mga Prinsipyo at Pamamaraan para sa Secure by Design Software
Nakatuwang sa pagkaka-akda ng CISA, NCSC at iba pang ahensya, inilalarawan ng gabay na ito kung paano dapat gumawa ng mga hakbang ang mga tagagawa ng sistema ng software, kabilang ang AI, upang isali ang seguridad sa disenyo yugto ng pagbuo ng produkto, at pagpapadala ng mga produktong ligtas na out of the box.

Buod ng mga Alalahanin sa Seguridad ng AI
Ginawa ng Pampederal na Opisina ng Seguridad sa Impormasyon ng Alemanya(BSI), ang dokumentong ito ay nagbibigay ng panimula sa mga posibleng pag-atake sa machine learning system at mga potensyal na depensa laban sa mga pag-atakeng iyon.

Hiroshima Process International Gabay na mga Prinsipyo para sa Mga Organisasyong Bumubuo ng Mga Advanced na Sistema ng AI at Hiroshima Process International Code of Conduct para sa Mga Organisasyong Bumubuo ng Advanced na Sistema ng AI
Ang mga dokumentong ito, na ginawa bilang bahagi ng G7 Hiroshima AI Process, ay nagbibigay ng gabay para sa mga organisasyong bumubuo ng mga pinaka-advanced na sistema ng AI, kabilang ang mga pinakabagong foundation model at generative AI system na may layuning i-sulong ang ligtas, secure, at mapagkakatiwalaang AI sa buong mundo.

Ang AI Verify
AI Governance Testing Framework at Software toolkit ng Singapore na nagpapatunay sa performance ng sistema ng AI laban sa isang hanay ng mga internasyonal na kinikilalang prinsipyo sa pamamagitan ng mga standardized na pagsubok.

Multilayer Framework para sa Mabuting Pagsasanay sa Seguridad sa Cyber para sa AI — ENISA (europa.eu)
Isang framework para gabayan ang National Competent Authority at AI stakeholders sa mga hakbang na kailangan nilang sundin para gawing ligtas ang kanilang sistema ng AI, mga operasyon at proseso.

ISO 5338: Mga proseso ng life cycle ng sistema ng AI (Sinusuri)
Isang set ng mga proseso at nauugnay na konsepto para sa paglalarawan sa life cycle ng mga sistema ng AI batay sa machine learning at heuristic system.

AI Cloud Service Compliance Criteria Catalog (AIC4)
Ang AI Cloud Service Compliance Criteria Catalog ng BSI ay nagbibigay ng pamantayang tukoy sa AI, na nagbibigay-daan sa pagsusuri ng seguridad ng isang serbisyo ng AI sa buong lifecycle nito.

NIST IR 8269 (Draft) Isang Taxonomy at Terminology ng Adversarial Machine Learning
Isang hanay ng mga proseso at nauugnay na konsepto para sa paglalarawan ng life cycle ng sistema ng AI batay sa machine learning at heuristic system.

MITRE ATLAS
Isang base ng kaalaman sa mga taktika, diskarte, at case study ng kalaban para sa mga machine learning (ML) na sistema, na namodelo at naka-link sa balangkas ng MITER ATT&CK.

Isang Pangkalahatang-ideya ng Mga Malalang Panganib sa AI (2023)
Ginawa ng Sentro para sa Seguridad ng AI, ang dokumentong ito ay nagtatakda ng mga lugar ng panganib na dulot ng AI.

Mga Malalaking Modelo ng Wika: Mga Oportunidad at Mga Panganib para sa Industriya at Mga Awtoridad
Dokumentong ginawa ng BSI para sa mga kumpanya, awtoridad at developer na gustong matuto nang higit pa tungkol sa mga pagkakataon at panganib ng pagbuo, pagtalaga at/o paggamit ng mga LLM.

Ang mga open-source na proyekto upang matulungan ang mga user sa pagsubok sa seguridad ng mga modelo ng AI ay kinabibilangan ng:

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

Seguridad sa Cyber

Mga Layunin sa Pagganap ng Seguridad sa Cyber ng CISA
Isang karaniwang hanay ng mga proteksyon na dapat ipatupad ng lahat ng kritikal na entity ng imprastraktura upang makabuluhang bawasan ang posibilidad at epekto ng mga kilalang panganib at diskarte ng kalaban.

NCSC CAF Framework
Ang Cyber Assessment Framework (CAF) ay nagbibigay ng gabay para sa mga organisasyong responsable para sa napakahalagang mga serbisyo at aktibidad.

Ang Supply Chain Security Framework ng MITRE
Isang framework para sa pagsusuri ng mga tagapagtustos at nagbibigay serbisyo sa loob ng ugnayan sa pagtustos.

Pamamahala sa panganib

NIST AI Risk Management Framework (AI RMF)
Ang AI RMF ay nagbabalangkas kung paano pamahalaan ang mga socio-technical na panganib sa mga indibidwal, organisasyon, at lipunan na natatanging nauugnay sa AI.

ISO 27001: Seguridad ng impormasyon, cybersecurity at proteksyon sa pagka-pribado
Ang pamantayang ito ay nagbibigay sa mga organisasyon ng gabay sa pagtatatag, pagpapatupad at pagpapanatili ng isang sistema ng pamamahala ng seguridad ng impormasyon.

ISO 31000: Pamamahala sa peligro
Isang internasyonal na pamantayan na nagbibigay sa mga organisasyon ng mga patnubay at prinsipyo para sa pamamahala sa peligro sa loob ng mga organisasyon.

NCSC Risk Management Guidance
Ang gabay na ito ay tumutulong sa mga cyber security risk practitioner na mas maunawaan at mapamahalaan ang mga panganib sa seguridad sa cyber na nakakaapekto sa kanilang mga organisasyon.

Mga Tala

1. Tinutukoy dito bilang isang tao, pampublikong awtoridad, ahensya o iba pang katawan na bumuo ng sistema ng AI (o may binuo na sistema ng AI) at inilalagay ang sistema na iyon sa merkado o inilalagay ito sa serbisyo sa ilalim ng sarili nitong pangalan o trademark
2. Para sa higit pang impormasyon sa secure by design, tingnan ang Secure by Design na web page ng CISA at gabay Paglipat ng Balanse ng Cybersecurity Risk: Mga Prinsipyo at Pamamaraan para sa Secure by Design Software
3. Kabaligtaran sa mga non-ML AI approach gaya ng mga sistemang rule-based
4. Inilalarawan ng CEPS ang pitong uri ng pakikipag-ugnayan sa pagbuo ng AI sa kanilang publikasyon 'Pagkakasundo sa AI Value Chain sa Artificial Intelligence Act ng EU'
5. ISO/IEC 22989:2022(en) ay tumutukoy dito bilang 'isang functional na elemento na bumubuo ng sistema ng AI'
6. Ang NIST ay may tungkulin sa paggawa ng mga patnubay (at paggawa ng iba pang mga aksyon) upang isulong ang ligtas, secure, at mapagkakatiwalaang pag-unlad at paggamit ng Artificial Intelligence (AI). Tingnan ang Mga Responsibilidad ng NIST sa ilalim ng Oktubre 30, 2023 Executive Order
7. Higit pang impormasyon sa pagmomodelo ng pagbabanta ay makukuha mula sa OWASP Foundation
8. Tingnan ang MITER ATLAS Adversarial Machine Learning 101
9. GitHub: RCE PoC para sa Tensorflow gamit ang isang nakakahamak na layer ng Lambda
10. SLSA: 'Pag-iingat sa integridad ng artifact sa anumang software supply chain'
11. METI (Ministro ng Ekonomiya, Pangangalakal at Industriya ng Japan, 2023), 'Gabay sa Pagpapakilala ng Software Bill of Materials (SBOM) para sa Software Management'
12. Pananaliksik sa Google: Pag-aaral ng Machine: Ang Mataas na Interes na Credit Card ng Teknikal na Utang
13. Tramèr et al 2016, Pagnanakaw ng Mga Modelo ng Machine Learning sa pamamagitan ng Mga Prediction API
14. Boenisch, 2020, Mga Pag-atake laban sa Privacy ng Machine Learning (Bahagi 1): Mga Model Inversion Attack gamit ang IBM-ART Framework
15. Pambansang Sentro ng Seguridad sa Cyber, 2020, Magdisenyo at bumuo ng isang pribadong naka-host na Public Key Infrastructure