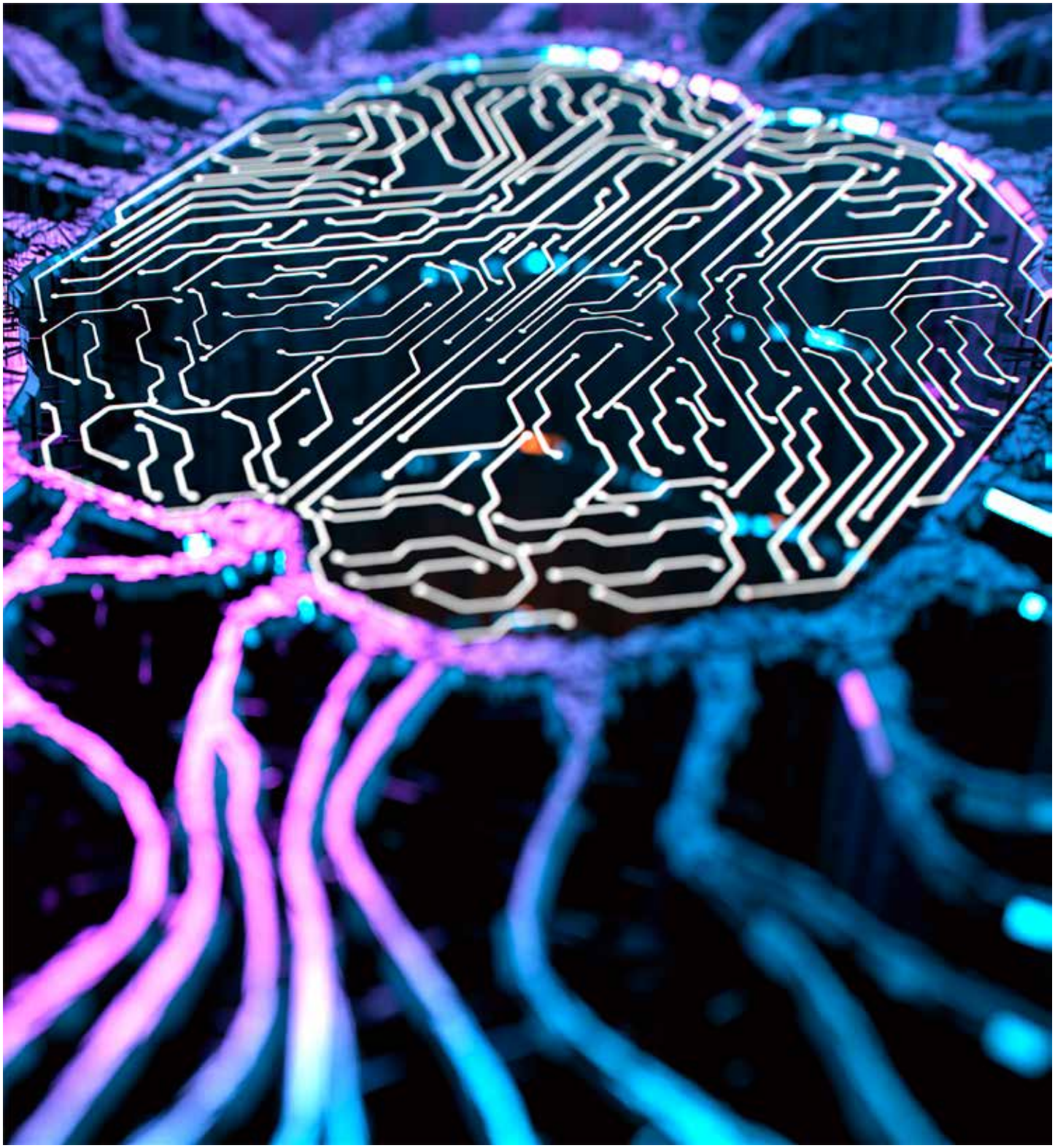
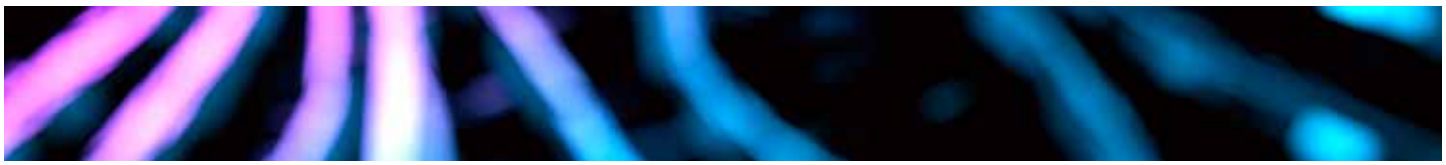
OI guideline bilong secure AI system development





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

Toksave long dispela document

Dispela document em UK National Cyber Security Centre (NCSC), US Cybersecurity and Infrastructure Security Agency (CISA), na ol dispela international partner pablisim:

- National Security Agency (NSA)
- Federal Bureau of Investigations (FBI)
- Australian Signals Directorate's Australian Cyber Security Centre (ACSC)
- Canadian Centre for Cyber Security (CCCS)
- New Zealand National Cyber Security Centre (NCSC-NZ)
- Chile's Government CSIRT
- Czechia's National Cyber and Information Security Agency (NUKIB)
- Information System Authority of Estonia (RIA) and National Cyber Security Centre of Estonia (NCSC-EE)
- French Cybersecurity Agency (ANSSI)
- Germany's Federal Office for Information Security (BSI)
- Israeli National Cyber Directorate (INCD)
- Italian National Cybersecurity Agency (ACN)
- Japan's National center of Incident readiness and Strategy for Cybersecurity (NISC)
- Japan's Secretariat of Science, Technology and Innovation Policy, Cabinet Office
- Nigeria's National Information Technology Development Agency (NITDA)
- Norwegian National Cyber Security Centre (NCSC-NO)
- Poland Ministry of Digital Affairs
- Poland's NASK National Research Institute (NASK)
- Republic of Korea National Intelligence Service (NIS)
- Cyber Security Agency of Singapore (CSA)

OI acknowledgement

OI dispela ogenaisesin halvim long development bilong ol dispela guideline:

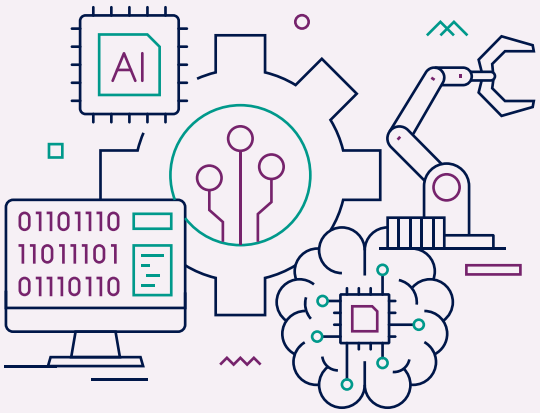
- Alan Turing Institute
- Anthropic
- Databricks
- Georgetown University's Center for Security and Emerging Technology
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute at Carnegie Mellon University
- Stanford Center for AI Safety
- Stanford Program on Geopolitics, Technology and Governance

OI disclaimer

OI infomesin stap insait long dispela document em kam "olsem tasol" long NCSC na ol authoring ogenaisesin na ol bai no inap kisim hevi sapos sampela kain bagarap kamap taim yu yusim ol dispela infomesin na dispela samting no stap aninit long wanpela lo. Infomesin insait long dispela document em no makim endorsement o recommendation bilong ol third party ogenaisesin, product, o service long lukluk bilong NCSC na ol arapela authoring agency. OI link na ol reference go long ol website na ol third party material em stap long givim infomesin tasol na em no makim endorsement o recommendation bilong ol dispela samting antap long ol arapela.

Dispela document em stap na bilong yusim aninit long TLP:CLEAR (https://www.first.org/tlp/).

OI content



Executive summary

Dispela document givim ol gutpela guideline long ol provider bilong olgeta kain system we em yusim artificial intelligence (AI), na dispela em ol system we ol mekim ol yet o ol mekim antap long ol tool o service bilong ol arapela provider. Taim ol provider yusim ol dispela guideline em bai halvim long mekim ol AI system wok olsem ol laikim, em bai mekim ol available o stap taim ol laik yusim, na bai wok gut na no inap lusim ol sensitive data go long ol unauthorised party.

Dispela document em bilong ol provider bilong ol AI system husait wok long yusim ol model em stap wantaim ol arapela ogenaisesin, o wok long yusim ol external application programming interfaces (APIs). Mipela tok strong long **olgeta** stakeholder (em olsem ol data scientist, developer, manager, decision-maker na risk owner) long ridim ol dispela guideline long halvim ol long mekim ol gutpela decision long **design**, **development**, **deployment** na **operation** bilong ol AI system.

Toksave long ol guideline

Ol AI system ken bringim planti gutpela samting long yumi olgeta. Tasol, sapos yumi laik kisim tru ol dispela gutpela samting em ken kam long AI, yumi mas mekim na yusim long wei em secure na responsible.

Ol AI system ken gat ol sampela kain security vulnerability bilong em yet we yumi mas tingim wantaim ol standard cyber security threat. Taim development em kamap hariap – kain olsem wantaim AI – security em ken kamap olsem liklik samting. Security mas kamap olsem bikpela samting, na no long development phase tasol, em mas olsem long olgeta hap bilong life cycle bilong system.

Wantaim dispela tingting, ol brukim ol guideline go long fopela key area insait long AI system development life cycle: **secure design**, **secure development**, **secure deployment**, na **secure operation na maintenance**. Long wanwan section mipela givim ol tingting na pasin we em bai halvim long daunim bikpela risk long AI system development process bilong ogenaisesin.

1. **Secure design**
   Dispela section gat ol guideline bilong design stage bilong AI system developent life cycle. Em karamapim save bilong ol risk na threat modelling, na tu ol topic na trade-off yu mas tingim taim yu wokim system na model design.

2. **Secure development**
   Dispela section em gat ol guideline bilong development stage bilong AI system development life cycle, wantaim supply chain security, documentation, na asset na technical debt management.

3. **Secure deployment**
   Dispela section em gat ol guideline bilong deployment stage bilong AI system development life cycle, wantaim wei bilong lukautim ol infrastructure na model long ol samting olsem compromise, threat o loss, incident management process, na responsible release.

4. **Secure operation and maintenance**
   Dispela section gat guideline bilong secure operation na maintenance stage bilong AI system development life cycle. Em givim guideline long ol action yu mas mekim taim system em wok stap, olsem logging na monitoring, update management na information sharing.

Ol guideline bihainim rot bilong 'secure by default', na tu bihainim ol pasin stap long Secure development and deployment guidance bilong NCSC, Secure Software Development Framework bilong NIST, na 'secure by design principles' em CISA, NCSC na ol international cyber agency pablisim. Ol lukluk moa long:

- kisim hevi bilong security wari bilong customer husait baim product
- holim strong pasin bilong tok tru na wokim samting klia
- strongim bun na ol lida bilong ogenaisesin long mekim secure by design kamap bikpela samting long business.

Introduction

Ol AI system ken bringim planti gutpela samting long yumi olgeta. Tasol, sapos yumi laik kisim tru ol dispela gutpela samting em ken kam long AI, yumi mas mekim na yusim long wei em secure na responsible. Cyber security em mas kamap pastaim long strongim safety, resilience, privacy, fairness, efficacy na reliablity bilong ol AI system.

Ol AI system ken gat ol sampela kain security vulnerability bilong em yet we yumi mas tingim wantaim ol standard cyber security threat. Taim development em kamap hariap – kain olsem wantaim AI – security em ken kamap olsem liklik samting. Security mas kamap olsem bikpela samting, na no long development phase tasol, em mas olsem long olgeta hap bilong life cycle bilong system.

Dispela document givim ol gutpela guideline long ol provider[1] bilong olgeta kain system em yusim AI, maski sapos ol systems em ol mekim ol yet o ol mekim antap long ol tool o service bilong ol arapela provider. Taim ol provider yusim ol dispela guideline em bai halvim long mekim ol AI system wok olsem ol laikim, em bai mekim ol available o stap taim ol laik yusim, na bai wok gut na no inap lusim ol sensitive data go long ol unauthorised party.

Yu mas lukim ol dispela guideline wantaim ol gutpela cyber security, risk management, na incident response best practice. Mipela tok strong long ol provider long bihainim ol 'secure by design'[2] pasin em US Cybersecurity and Infrastructure Security Agency (CISA), UK National Cyber Security Centre (NCSC), na olgeta international partner bilong mipela mekim. Ol dispela pasin lukluk moa long:

- kisim hevi bilong security wari bilong customer husait baim product
- holim strong pasin bilong tok tru na wokim samting klia
- strongim bun na ol lida bilong ogenaisesin long mekim secure by design kamap bikpela samting long business

Bihainim ol 'secure by design' pasin bai nidim planti resource long olgeta hap bilong life cycle bilong system. Em min olsem ol developer mas lukluk moa long ol **feature**, **mechanism**, na **implementation** bilong ol tool long lukautim ol customer long olgeta hap bilong system design, na long olgeta stage bilong development life cycle. Taim ol mekim olsem, dispela bai stopim ol bikpela senis bihain, na tu bai lukautim ol customer na data bilong ol nau yet.

Long wanem AI security em narakain?

Insait long dispela document yumi yusim 'AI' long makim ol machine learning (ML) application[3]. Olgeta kain ML tu yumi toktok long em. Yumi lukim ol ML application olsem ol application we:

- gat ol software component (model) em larim ol computer luksave na kisim save long ol pattern em stap long data na no nidim ol manmeri long givim program long wokim dispela samting
- em kamapim ol prediction, recommendation na decision wantaim statistical reasoning

Na wantaim ol cyber security threat em stap nau, ol AI system ken gat ol niupela kain birua o vulnerability. Dispela samting ol kolim 'adversarial machine learning' (AML), em yumi toktok long we ol birua ken painim vulnerability long ol ML component, kain olsem long ol hardware, software, workflow na supply chain. AML em givim ol attacker wei bilong kamapim birua insait long ol ML system kain olsem:

- bagarapim classification bilong model o regression performance
- larim ol user mekim ol unauthorised wok o action
- kisim ol sensitive model infomesin

Em gat planti wei long kamapim ol dispela birua, kain olsem ol prompt injection attack insait long large language model (LLM) domain, o bagarapim training data o user feedback (ol kolim 'data poisoning').

Husait mas ridim dispela document?

Dispela document em bilong ol provider bilong ol AI system husait wok long yusim ol model em stap wantaim ol arapela ogenaisesin, o wok long yusim ol external application programming interfaces (APIs). Mipela tok strong long **olgeta** stakeholder (em olsem ol data scientist, developer, manager, decision-maker na risk owner) long ridim ol dispela guideline long halvim ol long mekim ol gutpela decision long **design**, **deployment** na **operation** bilong ol machine learning AI system bilong ol.

Dispela em tru, tasol yumi bai no inap long yusim ol guideline wantaim olgeta ogenasesin. Ol kain na rot bilong attack em kamap long wei bilong husait lain laik bringim birua long AI system, olsem na ol ogenaisesin mas lukluk long ol guideline wantaim ol use case na threat profile bilong ol yet.

Husait em mas karim hevi bilong mekim secure AI?

Em gat planti ol actor insait long ol modern AI supply chain. Gutpela wei long lukim dispela em long tingting long tupela samting:

- 'provider' em husait karim hevi bilong data curation, algorithmic development, design, deployment na maintenance
- 'user' em husait givim ol input na save kisim ol output

Dispela provider-user wei bilong tingting em yumi save yusim wantaim planti ol application, tasol nau yumi no save lukim dispela samting tumas[4], taim ol provider wok long bringim ol software, data, model na/o remote service bilong ol third party kam insait long ol system bilong ol. Dispela ol bikpela supply chain em mekim hat long ol end user long save gut long husait karim hevi bilong strongim security bilong AI.

Ol user (em ol 'end user', o provider husait yusim wanpela external AI component[5]) no gat wei long lukim gut na/o gat gutpela save long painim na stretim gut ol risk em ken kamap wantaim ol system ol yusim. Olsem na, wantaim ol 'secure by design' principle o pasin, **ol provider bilong AI component mas karim hevi bilong ol security outcome bilong ol user em stap tamblo long supply chain.**

Ol provider mas putim ol security control na mitigation insait long ol model, pipeline na/o system bilong ol, na tu we ol yusim ol setting, ol mas putim ol secure option olsem default. Long we ol no inap long daunim ol risk, ol provider mas karim hevi bilong:

- toksave long ol user stap tamblo long supply chain long risk em ol na (sapos em tru) ol user bilong ol bai karim
- toksave long wei bilong yusim ol component wantaim gutpela security

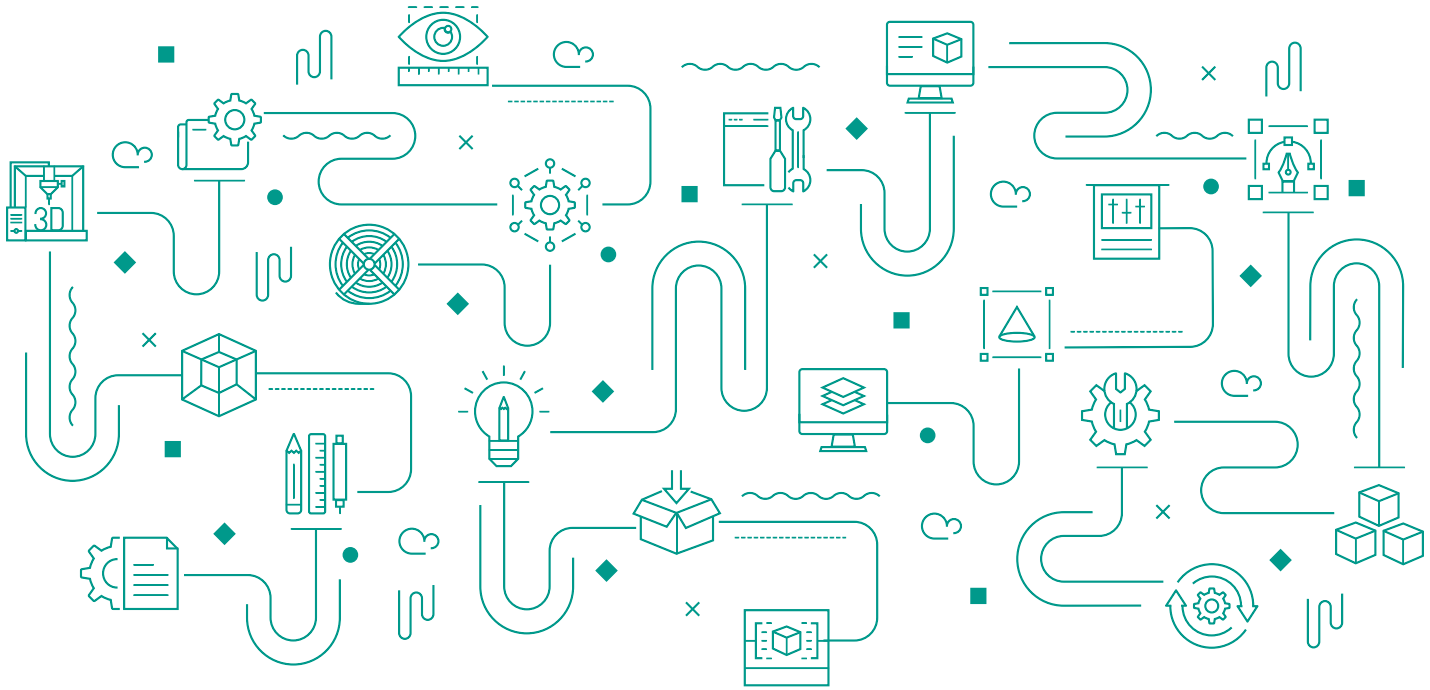
We ol birua long system ken kamapim bikpela bagarap, na lusim business operation, larim ol sensitive o confidential infomesin go long ol arapela, na/o samting we lo mas kam insait, ol mas soim ol AI cyber security risk olsem **critical.**

OI guideline bilong secure AI system development

OI ken brukim ol guideline go daun long fopela area insait long AI system development life cycle: **secure design**, **secure development**, **secure deployment**, na **secure operation na maintenance**. Long wanwan area, mipela givim ol consideration na mitigation long halvim daunim bikpela risk long AI system development process bilong ogenaisesin.

OI guideline stap long dispela document em klostu long ol software development life cycle practice em stap long:

- Secure development and deployment guidance bilong NCSC
- National Institute of Standards and Technology (NIST) Secure Software Development Framework (SSDF)[6]



1. Secure design

Dispela section em gat ol guideline em toktok long **design** stage bilong AI system development life cycle. Dispela em givim save long ol risk na threat modelling, na tu ol topic na trade-off long tingim taim yu lukluk long system na model design.

Strongim save bilong ol wok manmeri long ol threat na risk

Ol system owner na ol senior leader save gut long ol threat long secure AI na ol mitigation bilong ol. Ol data scientist na developer mas oltaim gat gutpela save long ol security threat na ol failture model na halvim ol risk owner long mekim ol gutpela decision. Yu givim ol user halvim long save long ol kain security risk em kam wantaim ol AI system (kain olsem, hap bilong standard InfoSec training) na lainim ol developer long yusim ol secure coding technique na secure na responsible AI pasin o practice.

Traim makim ol birua em ken kamap long system

Long risk management process bilong yu, yu mas putim wanpela bikpela o holistic process long luksave long ol threat long system bilong yu, na dispela em min olsem yu mas gat gutpela save long ol impact long system, ol user, ol ogenaisesin, na long olgeta manmeri sapos wanpela AI component em painim birua o no wok stret[7]. Dispela process em nidim taim long luksave long ol hevi em ken kamap wantaim ol AI threat[8] na yu mas raitim ol dispela decision go daun.

Yu luksave olsem ol sensitivity na kain data yu yusim long system bilong yu ken senisim value bilong em olsem wanpela target bilong ol attacker. Assessment bilong yu mas tingim olsem sampela threat bai kamap bikpela taim ol AI system kamap olsem ol bikpela target, na ol AI yet kamapim ol niupela, automated attack vector.

Taim yu mekim system bilong yu, tingim security tu wantaim functionality na performance

Yu mas gat strongpela tingting olsem wok yu laik wokim em bai mas yusim AI long wokim gut. Wantaim dispela tingting, yu mas lukluk gut long ol design choice yu mekim wantaim AI. Yu mas lukluk gut long threat model bilong yu na ol security mitigation em stap tu wantaim functionality, user experience, deployment environment, performance, assurance, oversight, ol ethical na legal requirement, na ol arapela kain samting olsem tu. Kain olsem:

- yu tingting gut long security bilong supply chain taim yu wok long mekim decision long mekim samting yu yet o yusim ol external component, kain olsem:
    - yu mekim training bilong wanpela niupela model, wantaim wanpela model em stap pinis (wantaim fine-tuning o nogat) o kisim wanpela model wantaim external API we em ken wok olsem yu laikim
    - sapos yu wok wantaim wanpela external model provider yu mas wokim due diligence evaluation long soim strong bilong security bilong dispela provider
    - sapos yu yusim wanpela external library, yu mekim due diligence evaluation (kain olsem, long luksave sapos dispela library em gat ol control long stopim system long yusim ol untrusted model na banisim ol yet long ol birua olsem arbitrary code execution[9])
    - yu yusim ol scanning na isolation/sandboxing taim yu laik kisim ol third-party model o ol serialised weight, na dispela em bai mas makim olsem untrusted third-party code na em ken yusim remote code execution

  - sapos yu yusim ol external API, yu mas yusim ol gutpela control long ol data bilong yu we em ken go long ol service autsait long control bilong ogenaisesin bilong yu, kain olsem ol user mas log in na givim tok orait bipo long ol salim sensitive information go aut
  - yu mas yusim ol gutpela wei bilong sekim na klinim ol data na input; dispela em olsem taim yu kisim user feedback o ol continuous learning data long model bilong yu, na luksave olsem training data bai kamapim wei system bai wok

- yu putim AI software system development go insait long ol gutpela secure development na operation pasin; olgeta hap bilong AI system em ol raitim insait long ol gutpela environment wantaim ol coding practice na language we em daunim o rausim ol kain vulnerability we em luksave long em na ken mekim

- sapos ol AI component mas mekim ol sampela kain action, kain olsem senisim ol file o putim ol output go long ol external system, yu mas yusim ol gutpela banis long ol action bai kamap (dispela em kain olsem ol external AI na non-AI fail-safe sapos dispela mas kamap)

- ol decision bilong user interaction em bai kam long ol risk em stap long AI, kain olsem:

  - system bilong yu mas givim ol user ol gutpela output tasol mas noken givim ol infomesin we ol birua ken yusim
  - sapos yu mas mekim, system bilong yu mas givim ol gutpela banis long lukautim ol model output
  - sapos yu laik givim API long ol external customer o collaborator, yu mas putim ol gutpela control we em bai daunim ol attack o birua em ken kamap long AI system wantaim dispela API
  - yu mas yusim ol strongpela security setting long system pastaim tru
  - yu mas yusim least privilege principle long givim liklik access long functionality bilong system
  - yu mas toksave long ol capability em ken kamapim hevi na askim ol user long tok orait bipo long ol yusim ol dispela samting; yu mas tok klia long ol prohibited use case, na, sapos yu inap, toksave long ol user long ol arapela wei bilong wokim dispela samting

Tingting gut long ol security benefit na trade-off taim yu wok long painim wanpela AI model

Wanem AI model yu yusim bai nidim gutpela tingting long ol kainkain requirement. Dispela em kain samting olsem wanem kain model architecture, configuration, training data, training algorithm na ol hyperparameter. Decision bilong yu bai mas kisim tingting long threat model bilong yu, na yu mas sekim gen wantaim ol niupela AI security research na save em kam taim dispela ol threat senis.

Taim yu laik yusim wanpela AI model, yu mas tingting ol dispela kain samting:

- sais bilong model yu yusim, em olsem architecture na hamas parameters em bai gat; dispela ol samting em bai givim tingting long hamas training data bai yu nidim na strong bilong em taim ol samting senis long input data taim yu wok long yusim

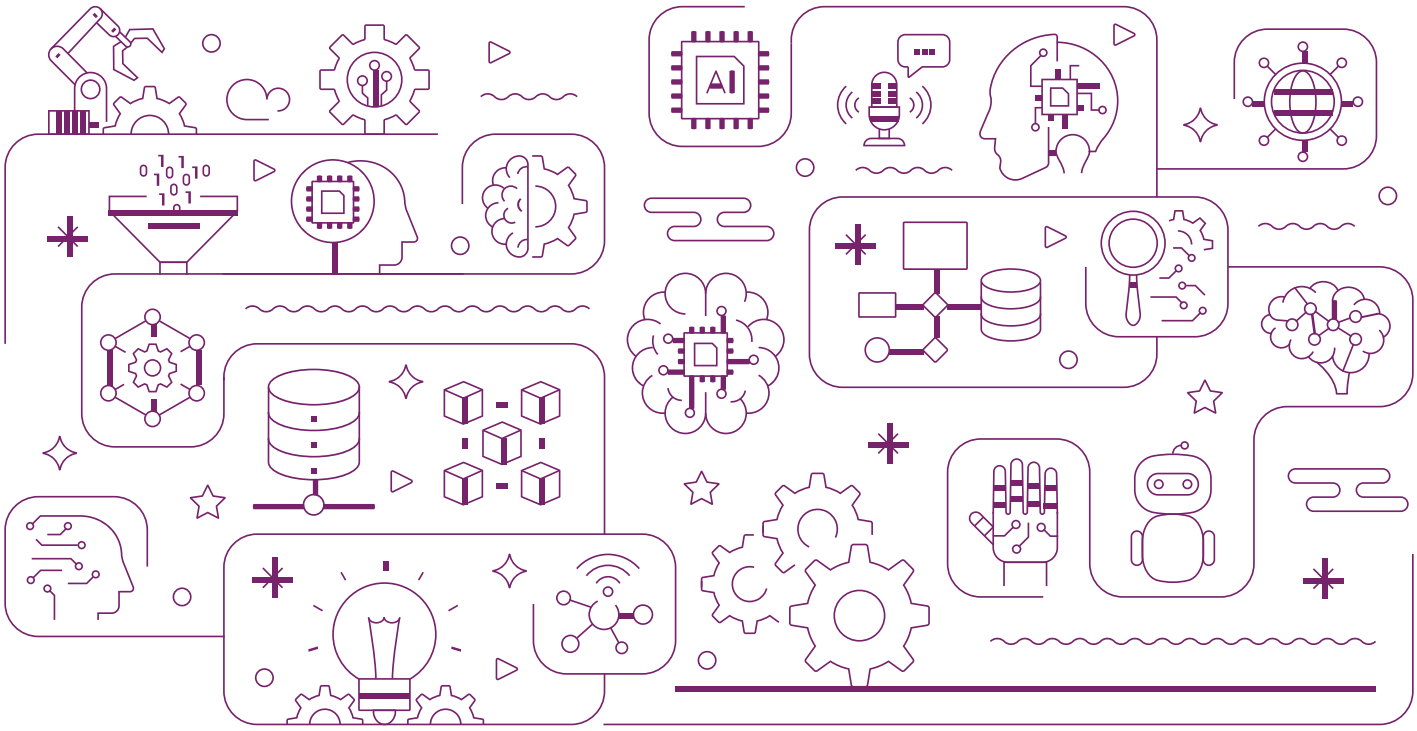
- sapos model yu yusim em bai fit wantaim ol use case bilong yu na/o em bai isi long senisim long wok wantaim samting yu laik mekim (kain olsem ol fine-tuning)

- save bilong yu long senisim, luksave na toksave long ol output bilong model bilong yu (kain olsem long debugging, audit o long bihainim ol regulation o lo); em gutpela tu sapos yu yusim ol model em isi na klia na no yusim ol bikpela model we em ken hat long wok wantaim

- ol kain samting long training dataset, olsem size, integrity, quality, sensitivity, age, relevance na diversity

- value bilong yusim ol model hardening (kain olsem adversarial training), regularisation na/o ol privacy-enhancing technique
- ol rot na supply chain bilong ol component wantaim ol model o foundation model, training data na ol tool

Long painim sampela moa infomesin long hamaspela bilong ol dispela samting ken senisim ol security outcome, go long 'Principles for the Security of Machine Learning' bilong NCSC, na lukluk gut long Design for security (model architecture).



2. Secure development

Dispela section em gat ol guideline em toktok long **development** stage bilong AI system development life cycle, na tu em toktok long supply chain security, documentation, na asset na technical debt management.

Lukautim na strongim supply chain bilong yu

Yu mas lukluk gut na bihainim security bilong ol AI supply chain long olgeta hap long life cycle bilong system bilong yu, na askim ol supplier long bihainim ol wankain standard olsem ogenaisesin bilong yu save yusim long ol arapela software. Sapos ol supplier no inap long bihainim ol standard bilong ogenaisesin bilong yu, yu mas bihainim rot em stap insait long ol risk management policy bilong ogenaisesin bilong yu.

Taim yu no mekim samting yu yet, yu mas kisim na lukautim ol gutpela na strongpela hardware na software component (kain olsem, ol model, data, software library, module, middleware, framework, na external API) kam long ol gutpela commercial, open source, na ol arapela third-party developer long strongim security long ol system bilong yu.

Yu mas redi long kalap go long ol arapela solution wantaim ol mission-critical o bikpela system bilong yu, sapos ol security wari kamap. Yu yusim ol resource olsem dispela bilong NCSC Supply Chain Guidance na ol framework olsem Supply Chain Levels for Software Artifacts (SLSA)[10] long bihainim ol wok em kamap long ol supply chain na ol software development life cycle.

Makim, bihainim na lukautim ol asset bilong yu

Yu save gut long ol gutpela samting ol AI asset bringim long ogenaisesin bilong yu, kain olsem ol model, data (wantaim ol user feedback), prompt, software, documentation, log na assessment (wantaim infomesin long ol sampela unsafe capability na failure mode), na luksave long wanem hap ol dispela samting em wokim bikpela wok na wanem hap ol ken bringim birua sapos ol attacker go insait. Yu mekim ol log olsem sensitive data na putim ol control long lukautim confidentiality, integrity na availability bilong ol.

Yu save wanem hap ol asset bilong yu stap long em na luksave gut long ol risk em stap wantaim dispela ol samting. Yu gat ol process na tool long bihainim, sekim gut, wokim version control na lukautim ol asset bilong yu, na ken kisim go bek long wanpela gutpela state sapos birua o bagarap kamap.

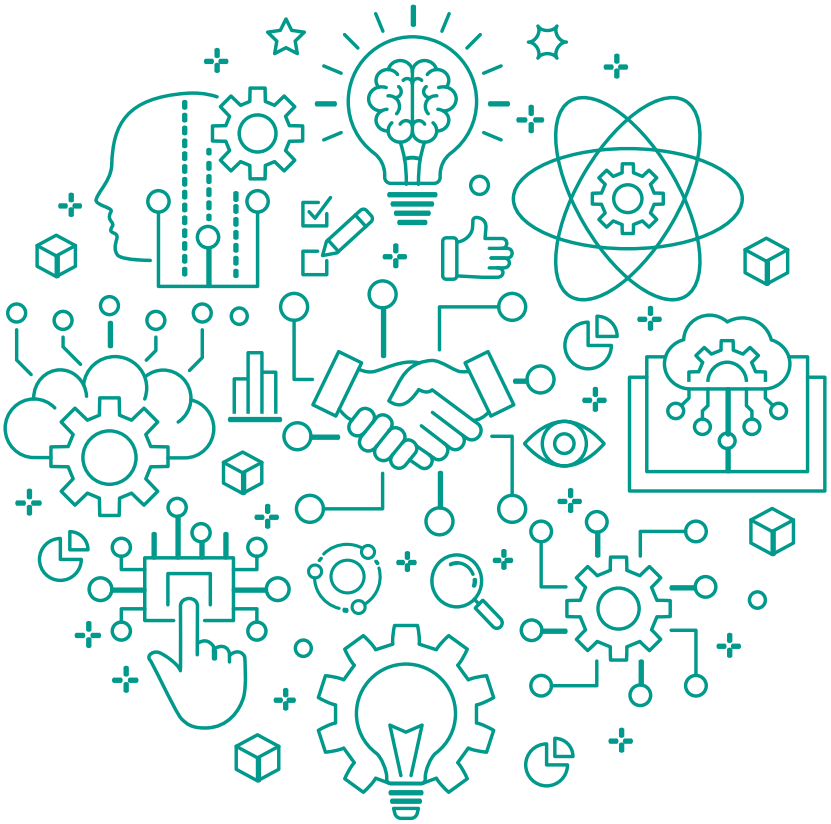
Yu gat ol process na control stap long lukautim wanem data ol AI system ken go insait, na long lukautim ol content AI em mekim wantaim wanem kain sensitivity em mas gat (na wanem kain sensitivity em stap long ol input em yusim long mekim).

Raitim go daun olgeta data, model na prompt

Yu raitim go daun ol wok bilong mekim, ronim na lukautim life cycle bilong olgeta kain model, dataset na meta-o system-prompt. Yu raitim go daun ol security infomesin olsem ol source bilong training data (wantaim ol fine-tuning data na human o arapela operational feedback), intended scope na ol limitation, ol banis o guardrail, ol cryptographic hash o signature, retention time, suggested review frequency na ol potential failure mode. OI gutpela samting long halvim wantaim dispela em ol model card, data card na software bills of materials (SBOMs). Taim yu kamapim ol gutpela documentation dispela bai sapotim transparency na accountability[11].

Lukaut gut long technical debt

Kain olsem ol arapela software system, yu luksave, bihainim na lukautim ol 'technical debt' bilong yu long life cycle bilong AI system (technical debt em we ol engineering decision lukluk long kamapim gutpela short-term result tasol na lusim ol longer-term benefit). Olsem financial debt, technical debt em no samting nogut, tasol yu mas lukautim gut long taim yu statim ol development wok[12]. Yu luksave olsem dispela em bai hat moa long AI olsem long ol standard software, na tu ol technical debt em bai antap moa taim ol wok hariap long ol development cycle na nogat ol gutpela protocol na interface em stap yet. Ol life cycle plan bilong yu (wantaim ol process bilong pinisim ol AI system tu) mas luksave gut long olgeta risk em ken kamap long wankain system em ken kamap bihain taim.



3. Secure deployment

Dispela section em gat ol guideline em toktok long **deployment** stage bilong AI system development life cycle, na tu toktok long lukautim ol infrastructure na model long ol birua o bagarap, mekim incident management process, na wei bilong salim samting go aut gut.

Strongim infrastructure bilong yu

Yu mas yusim ol gutpela infrastructure principle long infrastructure em stap long olgeta hap bilong life cycle bilong system bilong yu. Yu mas yusim ol gutpela access control long ol API, model na data, na long ol training na processing pipeline, long research na development na deployment bilong ol tu. Dispela em ol samting olsem gutpela segregation o banis long ol environment em holim ol sensitive code o data. Dispela tu bai halvim long daunim ol standard cyber security attack we ol laik traim long stilim model bilong yu o bagarapim performance bilong em.

Lukautim model bilong yu olgeta taim

Ol attacker bai traim long bihainim pasin bilong wanpela model[13] o data ol yusim long mekim training bilong em[14], na long wokim ol bai traim go insait long model stret (wantaim ol model weight) o traim arapela wei (wantaim ol arapela application o service). Ol attacker bai traim long senisim ol model, data o ol prompt long taim bilong training o bihain, long traim bagarapim output bilong model.

Yu mas lukautim gut model na data bilong yu long ol birua husait bai traim go insait stret o long ol arapela wei, wantaim ol dispela samting:

- yusim ol standard cyber security practice o pasin
- yusim ol control long ol query interface long luksave na pasim ol birua husait laik traim long go insait na senisim o kisim ol confidential infomesin

Long strongim ol consuming system long luksave gut long ol model, yu mas kamapim na givim o sherim ol cryptographic hash na/o ol signature bilong ol model file (kain olsem ol model weight) na dataset (wantaim ol checkpoint) long wankain taim yu pinisim training long model bilong yu. Wantaim olgeta cryptography, yu mas gat gutpela key management[15].

Rot yu bihainim long lukautim confidentiality risk em bai kam long ol use case na threat model bilong yu. Sampela application, kain olsem ol dispela em gat ol sensitive data, em bai nidim sampela ol theoretical guarantee na dispela em bai hat o bai nidim bikpela moni long kisim. Sapos yu nidim, sampela ol privacy technology (olsem differential privacy o homomorphic encryption) em stap we yu ken yusim long lukluk gut long ol risk em stap wantaim ol consumer, user na ol attacker sapos ol gat access long ol model o output.

Kamapim ol incident management procedure

Yu tok klia long ol security incident em ken kamap long ol AI system bilong yu long incident response, escalation na remediation plan bilong yu. Ol plan bilong yu tok klia long ol kain samting ken kamap na yu lukluk gen long ol dispela samting taim system o research em senis. Yu holim ol critical digital resource bilong company long ol offline backup. Ol responder bin kisim training long luksave na stretim ol AI incident. Yu givim ol gutpela audit log na ol arapela security feature o infomesin long ol customer na user na no askim ol long baim moa, long halvim incident response process bilong ol.

Strongim AI gut bipo long yu salim go aut 

Yu salim go aut ol model, application o system bihain long yu sekim gut wantaim ol gutpela security evaluation olsem benchmarking na red teaming (na tu ol arapela test em autsait long dispela ol guideline olsem safety o fairness), na yu tok klia long ol user long ol hap we em sot o bai no inap wok stret. Infomesin long ol open-source security testing library em ol givim long sampela moa samting long ridim section stap baksait long dispela document.

Mekim isi long ol user long wokim samting stret 

Yu luksave olsem yu mas lukluk gut long olgeta niupela setting o configuration option em kamap na luksave long wanem gutpela samting em bai bringim long business, na wanem kain ol risk bai kam wantaim em. Gutpela samting tru, em we yu yusim strongpela security setting tru insait long system olsem wanpela option o rot tasol. Sapos configuration mas kamap, wanem option em stap mas gat strong long stopim olgeta common threat o birua (em olsem, secure by default). Yu putim ol control long stopim ol birua long yusim system bilong yu long bringim hevi o bagarap.

Yu givim guidance na sapot long ol user bilong yu long yusim model o system bilong yu stret, na tu soim ol hap we ol em sot liklik o bai no wok gut. Yu tok klia long ol user long ol hap bilong security em ol mas lukautim, na mas toksave gut long wanem hap (na wei) ol ken yusim, kisim o holim data bilong ol (kain olsem, sapos ol yusim long model training, o taim ol wok manmeri o partner bai sekim).

4. Secure operation na maintenance

Dispela section em gat ol guideline em toktok long **secure operation na maintenance** stage bilong AI system development life cycle. Dispela givim ol guideline long ol action mas kamap taim system em go aut, kain olsem logging na monitoring, update management na information sharing.

Lukluk gut long behaviour bilong system bilong yu

Yu ken lukim output na performance bilong model na system bilong yu we yu ken lukim ol senis wok long kamap long behaviour bilong system na wok long senisim security bilong em. Yu ken makim na lukim gut ol birua woklong kamap, na tu ol senis em woklong kamap em yet long data.

Lukluk gut long ol input bilong system

Wantaim ol privacy na data protection requirement, yu lukluk gut na putim ol input long system bilong yu (kain olsem ol inference request, ol query o prompt) long halvim wantaim ol compliance obligation, audit, investigation na remediation sapos yu painim birua or bagarap. Dispela em ken kamap wantaim ol explicit detection bilong ol out-of-distribution na/o ol adversarial input, kain olsem ol dispela em laik bringim birua long ol data preparation step (kain olsem cropping na resizing long ol image o piksa).

Bihainim rot bilong secure by design wantaim ol update

Yu mas putim ol automated update olsem default insait long olgeta product na yusim ol secure, modular update procedure long salim ol go aut. Ol update process bilong yu (wantaim ol testing na evaluation) mas soim olsem ol senis long data, model o ol prompt ken senisim ron bilong system (kain olsem, ol bikpela update em olsem niupela version bilong system). Yu halvim ol user long luksave na wok wantaim ol senis long model (kain olsem yu givim ol preview access na ol versioned API).

Bungim na tokaut long ol samting yu lainim

Yu wok wantaim ol information-sharing community, na wok wantaim ol arapela lain long industry, academia na government long sherim ol gutpela rot na pasin. Yu toktok oltaim long kisim feedback long system security, insait na autsait long ogenaisesin bilong yu, na tu givim tok orait long ol security researcher long lukluk na toksave long ol birua o vulnerability. Sapos samting kamap, yu toksave long community, kain olsem yu pablisim ol bulletin long toksave long ol vulnerability, wantaim gutpela infomesin long halvim ol long save gut long ol birua em stap. Yu wokim wok long stretim gut ol wari o birua hariap tasol.

Sampela moa samting bilong ridim

AI development

Ol pasin bilong bihainim long strongim security bilong machine learning
Ol toksave bilong NCSC taim yu mekim, salim na yusim ol system wantaim ML component.

Secure by Design – Senisim Wei Bilong Cybersecurity Birua: Pasin Na Rot Bilong Secure by Design Software
CISA, NCSC na ol arapela agency kamapim dispela toksave long soim ol manufacturer bilong software system, wantaim AI tu, rot bilong bihainim long putim gutpela security insait long design stage bilong product development, na salim ol product go aut wantaim gutpela security.

Olgeta AI Security Wari Stap Wantaim
German Federal Office for Information Security (BSI) mekim dispela toksave, na em givim tingting long sampela attack long ol machine learning system na wei bilong pasim ol dispela kain attack.

Hiroshima Process International Rot Bilong Bihainim bilong ol Ogenaisesin wok long Mekim ol Advanced AI Systems na Hiroshima Process International Pasin Bilong Bihainim bilong ol Ogenaisesin wok long Mekim ol Advanced AI Systems
Dispela ol document, ol mekim wantaim G7 Hiroshima AI Process, givim toksave long ol ogenaisesin wok long mekim ol advanced AI system, na tu ol advanced foundation model na generative AI system wantaim bikpela tingting long kamapim ol safe, secure na trustworthy AI system long olgeta hap long world.

Sekim AI
AI Governance Testing Framework na Software toolkit bilong Singapore em ol save yusim long sekim performance bilong ol AI system wantaim ol internationally recognised principle we ol yusim ol standardised test.

Multilayer Framework bilong ol Gutpela Cybersecurity Pasin bilong AI — ENISA (europa.eu)
Wanpela framework long halvim ol National Competent Authority na ol AI stakeholder long rot bilong bihainim long strongim ol AI system, operation na process bilong ol.

ISO 5338: Ol AI system life cycle process (Sekim yet)
Ol process na tingting bilong soim life cycle bilong ol AI system kam long machine learning na ol heuristic system.

AI Cloud Service Compliance Criteria Catalogue (AIC4)
AI Cloud Service Compliance Criteria Catalogue bilong BSI em givim ol rot bilong sekim security bilong ol AI service long lifecycle bilong ol.

NIST IR 8269 (Draft) Ol Tingting na Toktok bilong Adversarial Machine Learning
Ol process na tingting bilong toktok long life cycle bilong ol AI systems kam long machine learning na ol heuristic system.

MITRE ATLAS
Em knowledge base o peles we ol bungim ol rot ol birua save bihainim, na ol case study bilong machine learning (ML) system em bihainim na go wantaim MITRE ATT&CK framework.

Lukluk long of Bikpela AI Risk (2023)
Center for AI Safety em mekim dispela document long toksave long ol risk o birua em ken kam wantaim AI.

Ol Large Language Model: Ol Gutpela Samting na Birua bilong Industry na ol Authority
Dispela document em BSI mekim bilong ol company, authority na ol developer husait laik lainim sampela moa samting long ol gutpela samting na ol birua em stap long mekim, salim go aut na/o yusim ol LLM.

OI open-source project long halvim ol user long traim security bilong ol AI model em:

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

Cyber security

OI Cybersecurity Performance Goal bilong CISA
OI rot olgeta infrastruture entity o ogenaisesin mas bihainim long strongim ol yet na daunim sans bilong painim birua long ol risk na rot ol adversary save bihainim.

NCSC CAF Framework
Dispela Cyber Assessment Framework (CAF) em givim halvim long ol ogenaisesin husait save lukautim ol bikpela service na activity.

Supply Chain Security Framework bilong MITRE
Em framework bilong sekim ol supplier na service provider insait long supply chain.

Risk management

NIST AI Risk Management Framework (AI RMF)
Dispela AI RMF em soim ol rot bilong halvim ol manmeri na ol ogenaisesin wantaim ol risk o birua em ken kam wantaim AI.

ISO 27001: Information security, cybersecurity na privacy protection
Dispela standard em givim ol ogenaisesin halvim long sanapim, mekim na lukautim ol information security management system.

ISO 31000: Risk management
Dispela international standard em givim ol ogenaisesin ol guideline na pasin bilong risk management insait long ol ogenaisesin.

NCSC Risk Management Guidance
Dispela guidance em halvim ol cyber security risk wok manmeri long save gut na lukautim ol cyber security risk em stap long ogenaisesin bilong ol.

OI note

1. Long hia yumi tok olsem ol manmeri, public authority, agency o arapela body em bai mekim wanpela AI system (o em mekim pinis wanpela AI system) na putim dispela system long market o salim aninit long nem o trademark bilong em
2. Long painim sampela moa infomesin long secure by design, go long CISA na lukim Secure by Design web page na guidance Senisim Wei Bilong Cybersecurity Birua: Rot na Pasin bilong Secure by Design Software
3. Na narakain long ol non-ML AI rot kain olsem ol rule-based system
4. CEPS em toktok long sevenpela kain AI development rot long publication bilong ol 'Reconciling the AI Value Chain with the EU's Artificial Intelligence Act'
5. ISO/IEC 22989:2022(en) em toktok long dispela olsem 'wanpela samting em kamapim ol AI system'
6. NIST em gat bikpela wok bilong mekim ol guideline (na bihainim ol arapela rot) long strongim go het ol safe, secure, na trustworthy rot bilong mekim na yusim Artificial Intelligence (AI). Lukim NIST's Responsibilities Under the October 30, 2023 Executive Order
7. Sampela moa infomesin long threat modelling em stap long OWASP Foundation
8. Lukim MITRE ATLAS Adversarial Machine Learning 101
9. GitHub: RCE PoC for Tensorflow using a malicious Lambda layer
10. SLSA: 'Safeguarding artifact integrity across any software supply chain'
11. METI (Japanese Ministry of Economy, Trade and Industry, 2023), 'Guide of Introduction of Software Bill of Materials (SBOM) for Software Management'
12. Google research: Machine Learning: The High Interest Credit Card of Technical Debt
13. Tramèr et al 2016, Stealing Machine Learning Models via Prediction APIs
14. Boenisch, 2020, Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework
15. National Cyber Security Centre, 2020, Design and build a privately hosted Public Key Infrastructure