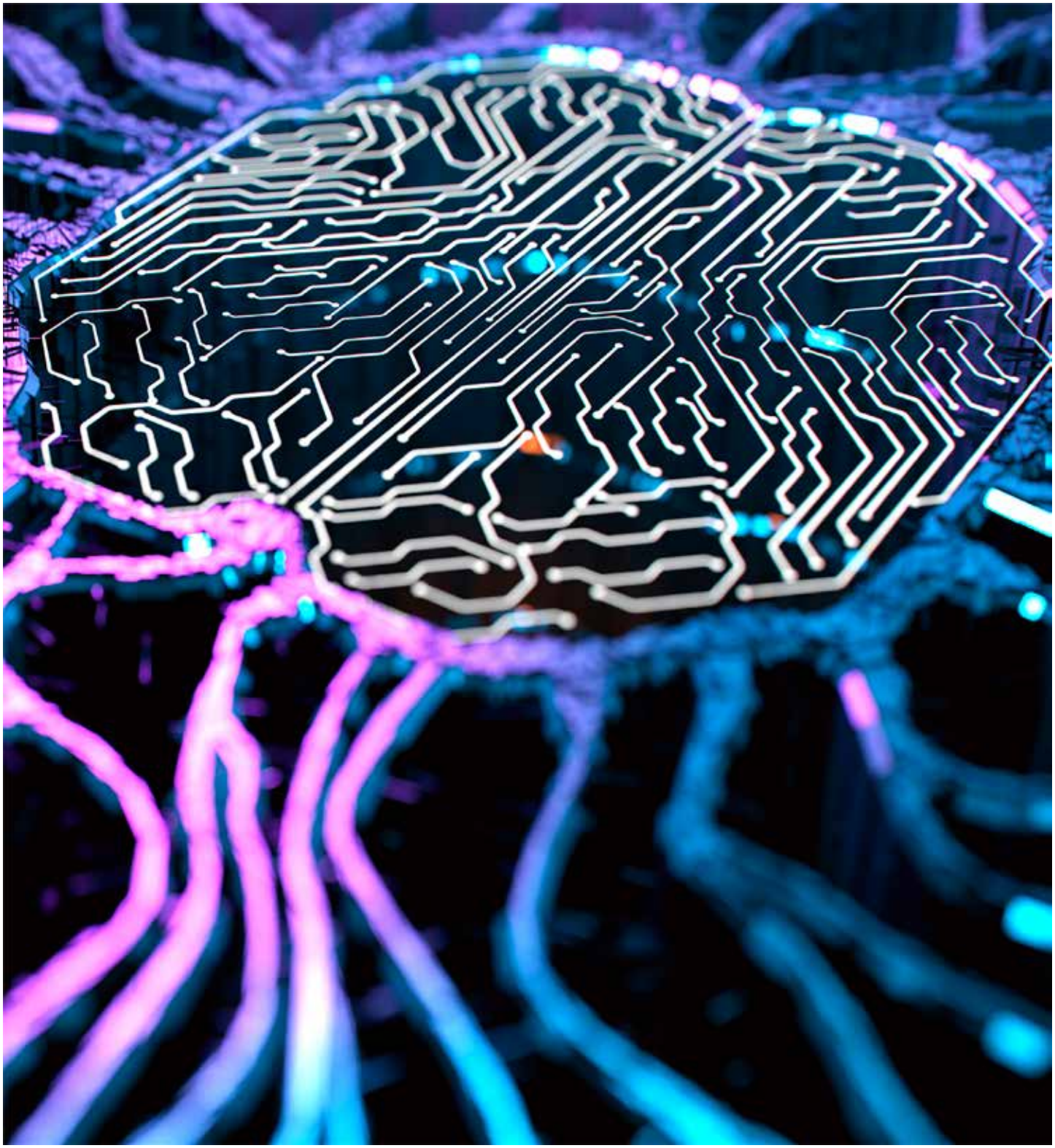
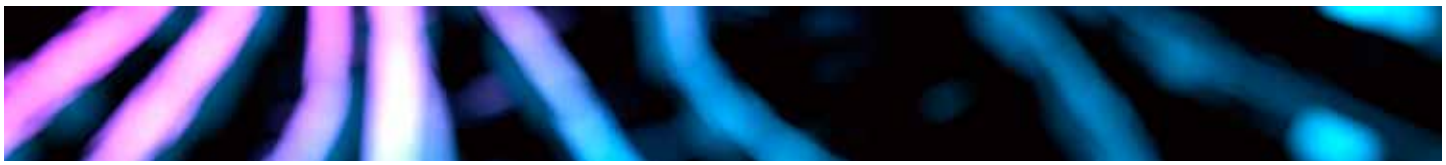
Guia ba dezenvolvimentu seguru sistema IA





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

Kona-ba dokumentu ida ne'e

Dokumentu ida ne'e publika ho Sentru Nasional Seguransa Sibernétika Reinu Unidu (NCSC), Ajênsia Seguransa Sibernétika no Infraestrutura EUA (CISA) no parseiru internasional sira tuir mai:

- Ajênsia Seguransa Nasional (NSA)
- Departamentu Federal Investigasaun (FBI)
- Sentru Australianu Seguransa Sibernétika (ACSC) Diretorate Sinal Australianu
- Sentru Kanadense Seguransa Sibernétika (CCCS)
- Sentru Nasional Seguransa Sibernétika Nova Zelândia (NCSC-NZ)
- CSIRT Governu Chile
- Ajênsia Nasional Seguransa Sibernétika no Informasaun Tcheka (NUKIB)
- Autoridade Sistema Informasaun Estónia (RIA) no Sentru Nasional Seguransa Sibernétika Estónia (NCSC-EE)
- Ajênsia Francesa Siberseguransa (ANSSI)
- Eskritóriu Federal Seguransa Informasaun Alemanha (BSI)
- Diresaun Nasional Sibernétika Israel (INCD)
- Ajênsia Nasional Italiana Siberseguransa (ACN)
- Sentru Nasional Preparasaun ba Insidente no Estratéjia Seguransa Sibernétika Japaun (NISC)
- Sekretaria Polítika Siênsia, Teknolojia no Inovasaun Japaun, Gabinete Governu
- Ajênsia Nasional Dezenvolvimentu Teknolojia Informasaun Nigéria (NITDA)
- Sentru Nasional Norueguês Seguransa Sibernétika (NCSC-NO)
- Ministériu Asuntu Dijital Polónia
- Institutu Nasional Peskiza NASK Polónia (NASK)
- Servisu Nasional Intelijênsia República Koreia (NIS)
- Ajênsia Seguransa Sibernétika Singapura (CSA)

Rekoñesimentu

Kontribuisaun organizasaun tuir mai ne'e ba dezenvolvimentu guia sira ne'e:

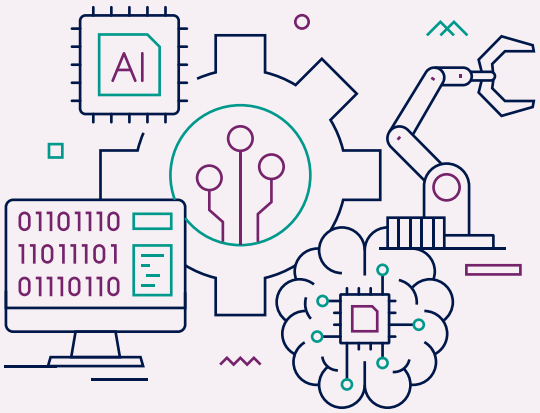
- Institutu Alan Turing
- Anthropic
- Databricks
- Sentru Seguransa no Teknolojia Emerjente Universidade Georgetown
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Institutu Enjeñaria Software iha Universidade Carnegie Mellon
- Sentru Stanford ba Seguransa IA
- Programa Stanford kona-ba Geopolítika, Teknolojia no Governansa

Desaprovadór

Informasaun iha dokumentu ida ne'e fornese "tuir loloos nian" ho NCSC no organizasaun autora sira ne'ebé sei la responsável ba kualkér perda, danus ka defeitu husi tipu kualkér tipu husi ninia utiliza só deit obriga tanba lei. Informasaun iha dokumentu ida ne'e la konstitui ka implika endosu ka rekomendasaun kualkér organizasaun partidu terseiru, produtu ka servisu ho NCSC no ajênsia autora. Links no referênsia ba site no material partidu terseiru sira, fornese hela ba informasaun deit no la reprezenta endossa ka rekomendasaun husi rekursu sira seluk.

Dokumentu ida ne'e halo disponivel ona iha baze TLP:CLEAR (https://www.first.org/tlp/).

Konteúdu



Sumáriu ezekutivu

Dokumentu ida ne'e rekomenda guia ba fornesedór husi kualkér sistema ne'ebé uza intelijênsia artifisial (IA), karik sistema sira ne'e kria ona husi zero ka konstrui iha ekipamentu no servisu tutun sira mak fornese ho sira seluk. Implementa guia sira ne'e sei ajuda provedór sira konstrui sistema IA ne'ebé funsaun hanesan intende ona, disponivel hela kuandu presija, no traballu sein revela dadus sensitivu ba partidu naun autorizadu sira.

Dokumentu ida ne'e destina prinsipalmente ba fornesedór sistema IA sira ne'ebé uza hela modelu mak ospedadou ho organizasaun, ka uza interface programasaun aplikativu (APIs) esterna. Ami husu **sira hotu** parte interesadu (inklui sientista dadus, dezenvolvedór, jerente, tomadór dezisaun no proprietáriu risku sira) ba lee guia sira ne'e atu ajuda sira foti dezisaun kona-ba **dezeñu**, **dezenvolvimentu** , **implantasaun** no **operasaun** ba sira-nia sistema IA.

Kona-ba guia ne'e

Sistema IA iha potensial atu fó benefisiu barak ba komunidade. Entantu, oportunidade ba IA atu sai koñese tomak, nia tenke dezenvolve, implanta no opera iha medidas seguru no responsavel.

Sistema IA ne'e sujeitu vulnerabilidade foun ba seguransa ne'ebé presija atu konsidera hamutuk ho padraun ameasa seguransa sibernétika. Kuandu ritmu dezenvolvimentu elevadu – hanesan kazu ho IA – seguransa dalaruma bele sai nu'udar konsiderasaun sekundária. Seguransa tenke sai rezikitu fundamental, laos deit iha faze dezenvolvimentu laran, maibe sistema siklu vida tomak.

Tanba razaun ida ne'e, guia sira ne;e divida ba kuatru área prinsipal iha siklu vida dezenvolvimentu sistema IA: **dezeñu seguru**, **dezenvolvimentu seguru**, **implantasaun seguru**, no **operasaun no manutensaun seguru**. Ba kada seksaun, ami sujere konsiderasaun no mitigasaun sira ne'ebé sei ajuda redus risku jeral ba prosesu dezenvolvimentu sistema organizasional IA.

1. **Dezeñu seguru**
   Seksaun ida ne'e kontéin ho guia sira ne'ebé aplika ba etapa dezeñu husi siklu vida dezenvolvimentu sistema IA. Ne'e inklui komprensaun kona-ba risku no modelu ameasa, no mós topiku espesifiku no kompenzasaun atu konsidera iha sistema no dezeñu modelu.

2. **Dezenvolvimentu seguru**
   Seksaun ida ne'e kontéin ho guia sira ne'ebé aplika ba etapa dezenvolvimentu siklu vida dezenvolvimentu sistema IA, inklui seguransa kadeia suprimentu, dokumentasaun no jerensiamentu ativu no dívida téknika.

3. **Implantasaun seguru**
   Seksaun ida ne'e kontéin ho guia sira ne'ebé aplika ba etapa implantasaun siklu vida dezenvolvimentu sistema IA, inklui infraestrutura protesaun no modelu sira husi kompromisu, ameasa ka perda, prosesu jerensiamentu insidente dezenvolve, no publikasaun responsavel.

4. **Operasaun no manutensaun seguru**
   Seksaun ida ne'e kontéin ho guia sira ne'ebé aplika ba etapa operasaun no manutensaun siklu vida dezenvolvimentu sistema IA. Ida ne'e fornese guia sira ba asaun partikularmente relevante kuandu sistem ida implanta ona, inklui rejistru no monitoramentu, jerensiamentu atualizasaun no kompartillamentu informasaun.

Guia ne'e tuir abordajen 'seguru tanba padraun', no aliña rabat ba prátika sira mak define ona iha NCSC's Guia dezenvolvimentu no implantasaun seguru, NIST's Estrutura Dezenvolvimentu Software Seguru, no 'prinsipiu seguru tanba dezeñu sira' publika ho CISA, NCSC no ajênsia sibernétika internasional. Sira prioritiza:

- asumi propriedade rezultadu seguransa ba kliente sira
- adota transparênsia no responsabilidade radikal
- konstrui estrutura organizasional no lideransa para seguru tanba dezeñu sai nu'udar prioridade tutun negosiu nian

Introdusaun

Sistema intelijênsia artifisial (IA) iha potensial atu lori mai benefísiu barak ba sosiedade. Entantu, ba oportunidade IA para bele implementa tomak, nia tenke dezenvolve no opera iha forma seguru no responsável. Siberseguransa ne'e mak pre-kondisaun nesesária ba seguransa, reziliênsia, privasidade, justisa, efikásia no fiabilidade sistema IA.

Entantu, sistema IA ne'e sujeitu ba vulnerabilidade seguransa foun ne'ebé tenke konsidera hamutuk ho ameasa normal seguransa sibernétika. Kuandu ritmu dezenvolvimentu ne'e elevadu – hanesan iha kazu ho IA – seguransa sempre bele sai konsiderasaun sekundária. Seguransa tenke sai rekizitu fundamental ida, laos deit iha faze dezenvolvimentu, maibe iha siklu vida sistema tomak.

Dokumentu ida ne'e rekomenda guia ba provedór sira[1] ba kualkér sistema mak uza iA, karik sistema sira ne'eba kria ona husi zero ka konstruí iha ekipamentu tutun nian no servisu fornese ho sira seluk. Implementa guia sira ne'e sei ajuda provedór sira ba konstrui sistema IA ne'ebé funsiona nu'udar intende tiha ona, sira disponivel kuandu presija, no traballu sein revela dadus sensitivu sira ba partidu naun autorizadu.

Guia sira ne'e tenke konsidera iha konjuntu ho estabelese seguransa sibernétika, jerensiamentu risku, no prátika melloria ba responde insidenti. Partikularmente, ami urjente provedór sira atu halo tuir prinsipiu 'seguru tanba dezeñu'[2] dezenvolve ho Seguransa sibernétika US no Ajênsia Seguransa Infraestrutura (CISA), Sentru Seguransa sibernétika Reinu Unidu (NCSC), no ami-nia parseira internasional hotu-hotu. Prinsipiu sira ne'e prioritiza:

- asumi propriedade rezultadu seguransa ba kliente sira
- adota transparênsia no responsabilidade radikal
- konstrui estrutura organizasional no lideransa para seguru tanba dezeñu sai nu'udar prioridade tutun negosiu nian

Segui prinsípiu "seguransa tanba dezeñu" presija rekursu signifikativu iha siklu vida sistema hotu-hotu. Ida ne'e signifika katak dezenvolvedór sira tenke investe ho prioritiza **rekursu**, **mekanizmu**, no **implementasaun** ekipamentu sira ne'ebé proteke kliente iha kada xamada husi dezeñu sistema, no iha etapa hotu-hotu siklu vida dezenvolvimentu. Halo ida ne'e sei prevene dezeñu fali ho folin karun iha futuru, no mós nu'udar guarda salva ba kliente sira no sira-nia dadus iha periodu badak.

Tanba sá seguransa IA ne'e diferente?

Iha dokumentu ida ne'e ami uza 'IA' espesifikamente refere ba aplikativu makina aprendizajen (ML)[3]. Tipu hotu-hotu husi ML ne'e mak iha eskopu. Ami define aplikativu ML sira nu'uda aplikativu ne'ebé:

- envolve komponente software sira (modellu sira) ne'ebé permite komputadór rekoñese no lori kontekstu ba padraun iha dadus sein regras tenke programa esplisitamente ho ema ida
- jeneraliza prediksaun sira, rekomendasaun sira ka dezisaun sira baze ba razaun estatistikal

No mós ameasa seguransa sibernétika mak ezistente, sistema IA sai sujeitu ba vulnerabilidade tipu foun nian. Termu "aprendizadu mákina adversáriu" (AML) ne'e uza atu deskreve eplorasaun vulnerabilidade fundamental sira iha komponente ML laran, inklui hardware, software, fluxu traballu no kadeia suprimentu sira. AML permite invasór sira kazua komportamentu la intensionadu iha sistema ML laran, ne'ebé inklui:

- Afeita klasifikasaun modelu nian ka realiza regresaun
- permite uzuáriu sira realiza asaun naun autorizadu
- estrai modelu informasaun sensitivu

Iha maneira barak atu obten efeitu sira ne'e, hanesan atake injesaun imediatu iha dominiu modelu linguajen grande (LLM) ka korupsaun delibera iha treinamentu dadus nian ka feedback uzuáriu (koñesidu nu'udar "envenenamento dadus").

Sé mak tenke lee dokumentu ida ne'e?

Dokumentu ida ne'e destina prinsipalmente ba fornesedór sistema IA sira, baze ba modelu ospedadu ho organizasaun ida ka uza interface programasaun aplikativu esterna sira (APIs). Entantu, ami urjente parte interesadu **hotu-hotu** (inklui sientista dadus, dezenvolvedór sira, jerente sira, tomadór dezisaun sira no proprietáriu risku sira) atu lee guia sira hodi ajuda sira halo dezisaun informadu kona-ba **implantasaun**, **dezeñu** no **operasaun** ho sira-nia makina aprendizajen sistema IA.

Maski nune'e, laos guia sira hotu-hotu sei aplikavél diretamente ba organizasaun hotu-hotu. Nível sofistikasaim no métodu atake sei varia depende ba adversáriu tarjetu hela ba sistema IA, para guia ne'e tenke konsidera konjuntu ho kazu uza no perfil informasaun ho Ita-nia organizasaun.

Sé mak responsavél ba dezenvolve IA seguru?

Sempre iha atór barak iha kadeia suprimentu modernu IA. Abordajen simples ida prezumu entidade rua:

- 'Provedór' nia mak responsável ba kurasaun dadus, dezenvolvimentu algorítmiku, dezeñu, implantasaun no manutensaun
- 'Uzuáriu', nia mak fornese entrada no simu rezultadu sira

Embora abordajen provedór-uzuáriu ida ne'e uza iha aplikativu barak, nia aumenta sai la komun[4], hanesan provedór bele tenta atu inkorpora software, dadus, modelu no/ka servisu remotu sira fornese ho partidu terseiru tama ba sira-nia sistema rasik. Kadeia suprimentu kompleksu sira ne'e halo araska liutan ba utilizadór final sira atu komprende iha ne'ebé responsabilidade seguru ba IA.

Uzuáriu (karik "uzuáriu final" ka provedór sira inkorpora komponente esternu IA[5]) normalmente la iha visibilidade no/ka esperiênsia sufisiente atu komprende tomak, avalia, ka rezolve risku asosiadu ho sistema mak sira uza. Tanba ne'e, liña ho prinsípiu "seguransa tanba dezeñu", **fornesedór komponente IA tenke asumi responsabilidade ba rezultadu seguransa utilizadór, tuun to'o kadeia suprimentu.**

Provedór tenke implementa kontrola no mitigasaun seguransa iha ne'ebé posível iha sira-nia modelu laran, pipelines no/ka sistema, no iha ne'ebé konfigurasaun uza ona, implementa opsaun seguru tebes nu'udar padraun. Iha ne'ebé risku labele mitiga, fornesedór tenke responsavel ba:

- informa utilizadór sira, tuun to'o kadeia suprimentu kona-ba risku mak sira infrenta no (karik aplikável) sira-nia utilizadór rasik aseita hela
- sujere ba sira kona-ba oinsá uza komponente ho seguru

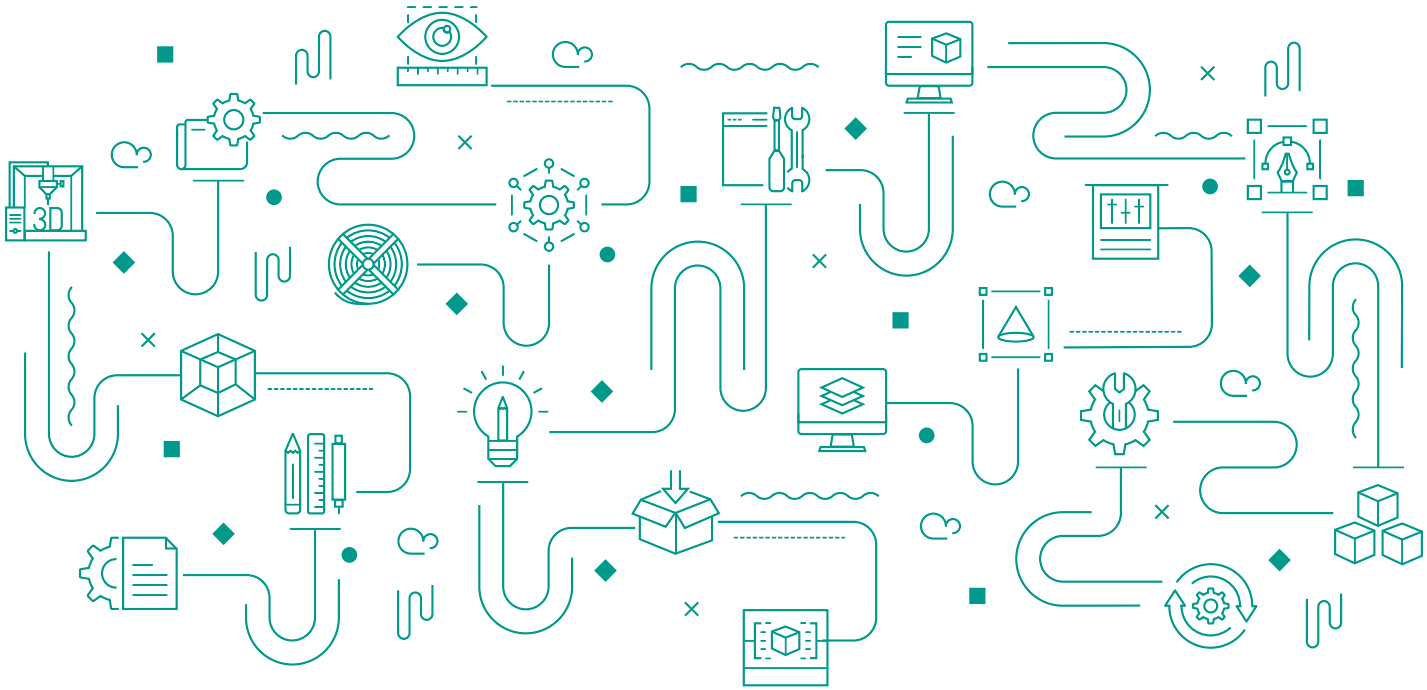
Iha ne'ebé kompromet imentu sistema bele leva ba danus fíziku ka reputasaun tangível ka jeneralizadu, perda signifikativu husi operasaun komersial, vazamentu informasaun sensível ka konfidensial no/ka implikasaun legal, risku seguransa sibernétika IA tenke trata nu'udar **krítiku.**

Guia ba seguru dezenvolvimentu sistema IA

Guia ne'e dividida tama ba kuatru área prinsipal iha sistema siklu vida dezenvolvimentu sistema IA: **dezeñu seguru**, **dezenvolvimentu seguru**, **implantasaun seguru** no **operasaun seguru no manutensaun**. Ba kada área, ami sujere konsiderasaun no mitigasaun ne'ebé sei ajuda redus risku jeral ba prosesu dezenvolvimentu sistema organiziaonal IA.

Guia mak konfigura iha dokumentu ida ne'e aliña rabat ba prátika siklu vida dezenvolvimentu software, define iha:

- Guia seguru dezenvolvimentu no implantasaun NCSC
- Institutu Nasional Padraun no Teknolojia (NIST) Estrutura dezenvolvimentu software seguru (SSDF)[6]



1. Dezeñu seguru

Seksaun ida ne'e kontéin guia sira ne'ebé aplika ba etapa **dezeñu** siklu vida dezenvolvimentu sistema IA nian. Nia kobre komprensaun ba risku no modelu ameasa, no mós topiku espesifiku no kompenzasaun atu konsidera iha sistema no dezeñu modelu.

Levanta konsiente funsinariu ba ameasa no risku sira

Proprietáriu sistema no líderes senior komprende ameasa sira ba seguransa IA no sira-nia mitigasaun. Ita-nia sientista dadus no dezenvolvedór sira mantein koñesimentu ameasa seguransa no modu falla relevante sira no ajuda proprietáriu risku sira atu toma dezisaun informadu. Ita fornese uzuáriu sira ho guia iha risku seguransa esklusivu enfrenta sistema IA (por ezemplu, nu'udar parte treinamentu padraun InfoSec) no treina dezenvolvedór iha téknika kodifikasaun segura no prátika IA segura no responsável.

Modelu ameasa sira ba Ita-nia sistema

Nu'udar parte husi Ita-nia prosesu jerensiamentu risku, Ita aplika prosesu olistiku atu avalia ameasa ba Ita-nia sistema, ne'ebé inklui komprensaun impaktu potensial ba sistema, uzuariu sira, organizasaun sira no sosiedade luan tan karik komponente IA kompromisu ka halo komportamentu la espera[7]. Prosesu ida ne'e envolve avaliasaun impaktu ba ameasa espesifiku IA[8] no dokumenta Ita-nia prosesu toma dezisaun.

Ita rekoñese katak sensitivu no tipu dadus mak uza iha Ita-nia sistema laran bele influensa ninia valór nu'udar tarjetu ba atakante ida. Ita-nia avaliasaun tenke konsidera katak ameasa balun bele dezenvolve tanba AI sistema aumenta haree nu'udar tarjetu ho valór altu, no hanesan IA rasik permite vektor, atake automatiku foun.

Dezeñu Ita-nia sistema ba seguransa no mós funsionalidade no dezempeñu

Ita konfidente katak tarefa mak iha di'ak liu realiza ho uza IA. Depois determina ida ne'e, Ita avalia adekuasaun husi Ita nia opsaun dezeñu espesifiku IA. Ita konsidera Ita-nia modelu ameasa no mitigasaun seguransa sira asosiadu juntamente funsionalidade, esperiênsia uzuáriu, ambiente implantasaun, dezempeñu, garantia, supervizaun, rekizitu étiku no legal, entre konsiderasaun sira seluk. Por ezemplu:

- Ita konsidera seguransa kadeia suprimentu kuandu hili karik atu dezenvolve iha uma laran ka uza komponente esterna, por ezemplu:
  - Ita-nia opsaun atu treinu modelu ida foun, uza modelu ezistente ida (ho ka sein finu di'ak) ka asesu modelu ida liuhusi API esternu ida ne'e apropriadu ba Ita-nia rekizitu.
  - Ita hili atu servisu ho provedór modelu esterna inklui avaliasaun due diligence husi própria postura seguransa provedór
  - karik uza biblioteka esterna ida, Ita kompleta avaliasaun due diligence (por ezemplu, atu garante biblioteka iha kontrola ne'ebé prevene sistema karga modelu la konfiansa sein espozisaun imediatamente ezekusaun arbitrária kódigu[9])
  - Ita implementa varedura no izolamentu/sandbox kuandu importa modelu terseiru ka serializadu todan, ne'ebé tenke trata nu'udar kódigu partidu terseiru naun konfiável no bele permite ezekuzaun remota kódigu

    - Karik uza API esterna, using an external APIs, Ita implementa kontrolu apropriadu ba dadus ne'ebé bele manda ba servisu esterna husi Ita-nia organizasaun kontrola, hanesan presija uzuáriu ba login no konfirma molok manda potensialidade informasaun sensitivu
    - Ita aplika verifikasaun no ijienizasaun apropriadu ba dadus no insumu; inklui ida ne'e kuandu inkorpora feedback uzuáriu ka dadus aprendizajen kontinua tama ba Ita-nia modelu, rekoñese katak dadus treinamentu define komportamentu sistema

- Ita integra dezenvolvimentu sistemas software IA ba prátika ezistente de dezenvolvimentu no operasaun segura melloria nian; elementu sistema IA hotu-hotu eskrita iha ambiente apropriadu uza prátika kodifikasaun no linguajen ne'ebé redus ka elimina klasse koñesidu vulnerabilidade iha ne'ebé plausível

- karik komponente IA tenke asiona asaun, por ezemplu, altera arkuivu ka diresiona rezultadu ba sistema esternu, Ita aplika restrisaun apropriadu ba asaun posível (ida ne'e inklui sistema seguransa kontra falla IA esternu no naun-IA, karik nesessáriu)

- dezisaun kona-ba interasaun uzuáriu ne'e informa ho risku espisifiku IA, por ezemplu:

    - Ita-nia sistema fornese uzuáriu ho rezultadu utilizável sein revela nível la nesesáriu husi detalle ba invasor potensial
    - karik nesesáriu, Ita-nia sistema fornese protesaun efikas besik rezultadu modelu nian
    - karik oferese API ida ba kliente esterna ka kolaboradór sira, Ita aplika kontrola apropriadu ne'ebé mitiga atake sira iha sistema IA liuhusi API
    - Ita integra konfigurasaun seguru tebtebes tama ba sistema ho padraun
    - Ita aplika prinsípiu priviléjiu mínimu atu limita asesu ba funsionalidade sistema
    - Ita esplika rekursu risku liuta ba uzuáriu no presija uzuáriu sira atu optein hodi uza sira; Ita komunika bandu uza kazu sira no iha ne'ebé posivelm informa uzuáriu sira kona-ba solusaun alternativua

Konsidera benefisiu seguransa no kompenzasaun kuandu hili modelu Ita-nia IA

Ita-nia opsaun ba modelu IA sei envolve ekuilíbriu série ida husi rekizitu sira. Ida ne'e inklui opsaun arkuitetura modelu, konfigurasaun, dadus treinamentu, algoritmu treinamentu no hiperparâmetru. Ita-nia dezisaun informa ho Ita-nia modelu ameasa, no regularmente avalia nu'udar peskiza avansadu seguransa IA no komprende ameasa mak evolui.

Kuandu hili modelu IA ida, Ita-nia konsiderasaun sei posivel inklui, maibe la limita ba:

- kompleksidade modelu Ita uza hela, ne'ebé arkuitetura mak hili ona no kuantidade parâmetru; Modelu Ita-nia arkuitetura mak hili on númeru parâmetru sira, entre fator sira seluk, afeita hanusa dadus treinamentu mak nia presija no hanusa robustu nia ba mudansa iha dadus entrada kuandau uza hela

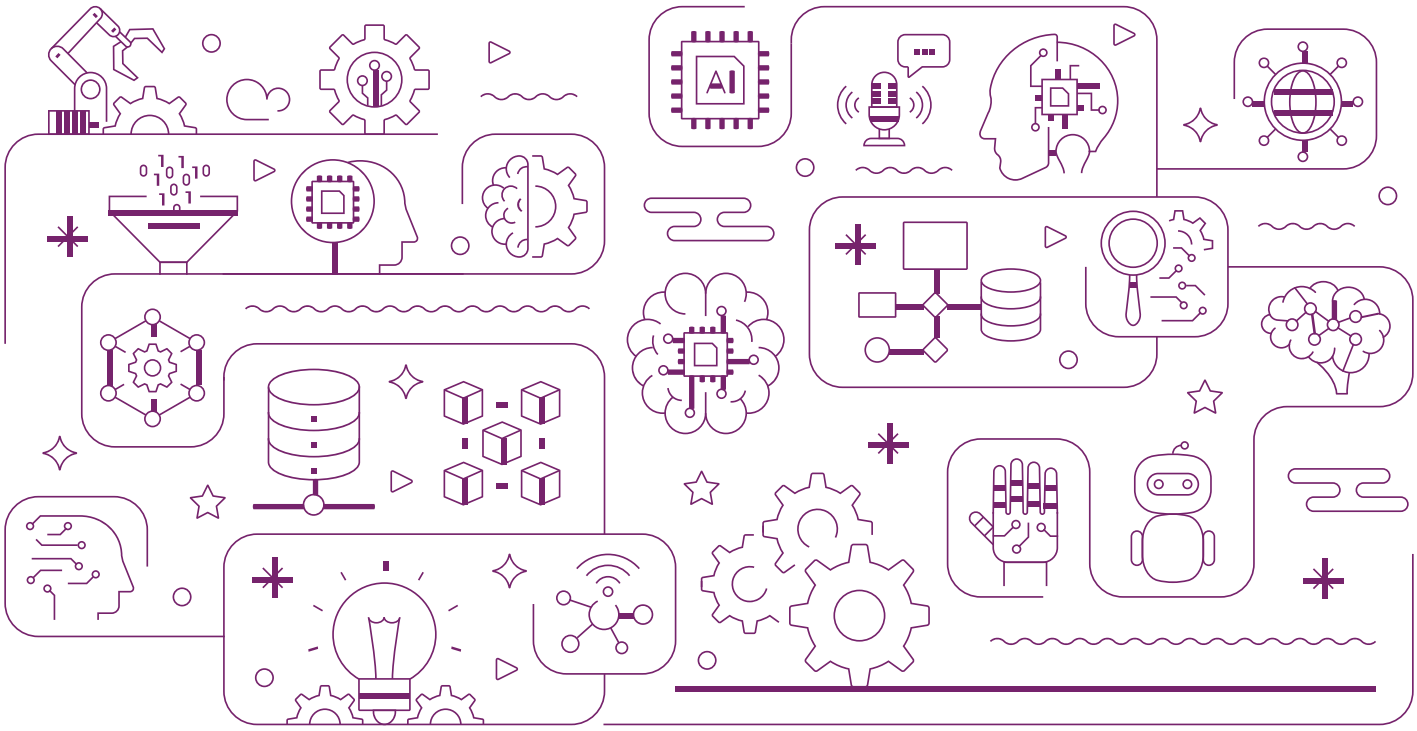
- Adekuasaun modelu ba Ita-nia kazu uza nian no/ka viabilidade adapta nia ba Ita-nia nesesisade espesifika (por ezemplu, liuhusi ajusta finu)

- kapasidade atu aliña, interpre no esplika Ita-nia modelu rezultadu (por ezamplu ba debugging, auditoria ka konformidade regulatória); iha benefísiu atu uza modelu simples liutan, transparente liutan duke modelu boot no komplikadu liutan ne'ebé araska liutan atu interpreta

- karakterístika husi konjuntu dadus treinamentu, inklui tamañu, integridade, kualidade, sensibilidade, idade, relelevânsia no diversidade

- valor ba uza fortalesimentu modelu (hanesan treinamentu adversáriu), regularizasaun no/ka téknika aprimoramento privasidade
- proveniênsia no kadeia suprimentu komponente sira, inklui modelu ka modelu báziku, dadus treinamentu no ekipamentu asosiadu

Ba informasaun liutan kona-ba hira husi fatór sira ne'e impaktu rezultadu seguransa, refere ba 'Prinsipiu ba Seguransa Aprendizadu Makina' NCSC, partikularmente Dezeñu ba seguransa (arkuitetura modelu).



2. Dezenvolvimentu seguru

Seksaun ne'e kontéin kontéin guia sira ne'ebé aplika ba etapa **dezenvolvimentu** siklu vida dezenvolvimentu sistema IA inklui seguransa kadeia suprimentu, dokumentasaun no propriedade no jerensiamentu deve tekniku.

Seguru Ita-nia kadeia suprimentu

Ita avalia no monitoriza seguransa Ita-nia kadeia suprimentu IA iha siklu vida sistema tomak, no presija fornesedór halo tuir padraun hanesan ho Ita-nia organizasaun rasik aplika ba software seluk. Karik fornesedór la bele halo tuir Ita-nia padraun organizasaun, Ita atu akordu ho Ita-nia politika jerensiamentu risku mak ezistente.

Iha ne'ebé la produs internamente, Ita obten no mantein komponente hardware no software protejidu no dokumentadu ho di'ak (por ezemplu, modelu, dadus, biblioteka software, módulu, middleware, estrutura no API esterna) husi komersial verifikadu, kódigu abertu, no dezenvolvedór terseiru seluk atu garante seguransa robusta iha Ita-nia sistema.

Ita prontu atu failover ba solusaun alternativu ba sistema misaun krítika, karik kritériu seguransa la atende. Ita uza rekursu hanesan Guia Kadeia Suprimentu NCSC no estrutura hanesan nivel Kadeia Suprimentu ba Artefatu Software (SLSA)[10] ba rastreia atestasaun husi kadeia suprimentu no siklu vida dezenvolvimentu software.

Identifika, rastreia no proteje Ita-nia ativu sira

Ita komprende valór ba Ita-nia organizasaun husi Ita-nia ativu relasiona ho IA, inklui modelu dadus (inklui feedback uzuáriu), prompts, software, dokumentasaun, rejistru no avaliasaun (inklu informasaun kona-ba kapasidade potensialmente insegura no modu falla), rekoñese iha ne'ebé sira reprezenta investimentu signifikativu no iha ne'ebé asesu ba sira permite invasor ida. Ita trata logs nu'udar dadus sensitivu no implementa kontrola atu proteje sira-nia konfidensialidade, integridade no disponibilidade.

Ita hatene iha ne'ebé Ita-nia ativu sira no avalia ona no aseita kualkér risku asosiadu sira. Ita prosesu ona no ekipamentu atu rastreia, autentika, kontrola versaun no seguru Ita-nia ativu sira no bele restaura iha estadu koñesidu di'ak iha kazu kompromet imentu.

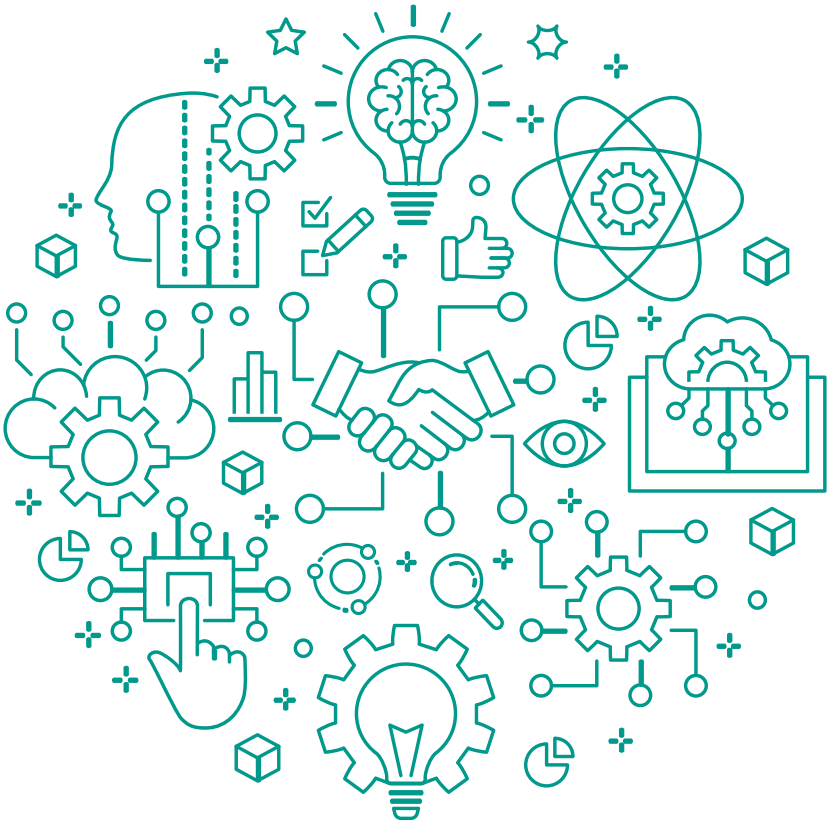
Ita prosesu ona no kontrola iha lokasaun atu jere sistema IA dadus sa'ida mak bele asesu, no atu jere konteudu jeneraliza ho IA akordu ba ninia sensitividade (no sensitividade husi entrada ne'ebé tama ba jeneraliza nia).

Dokumenta Ita-nia dadus, modelu sira no prompts

Ita dokumenta kriasaun, operasaun, no jerensiamentu siklu vida husi kualkér modelu, konjuntu dadus no meta-ka prompts sistema. Ita-nia dokumentasaun inklui informasaun relevante seguransa, hanesan rekursu dadus treinamentu (inklui dadus ajusta finu no feedback humanu ka feedback operasional seluk), eskopu no limitasaun pretendidu, protesaun, hashes ka asinatura kriptográfika, tempu retensaun, frekuênsia revisaun sujere ona no posível modu falla. Estrutura útel atu ajuda halo ida ne'e inklui kartaun modelu, kartaun dadus no lista material software (SBOMs). Produsaun dokumentasaun abranjente apoiu transparênsia no responsabilizasaun[11].

Jerensia Ita-nia dívida téknika

Hanesan ho kualkér sistema software, Ita identifika, rastreia no jere Ita-nia 'dívida téknika' iha siklu vida sistema IA hotu-hotu (dívida téknika iha ne'ebé dezisaun enjeñaria la kompletu prátika melloria atu alkanse rezultadu prazu badak, ho sakrifísiu prazu naruk-benefísiu ba prazu naruk liutan). Hanesan dívida finanseira, dívida téknika la inerentemente aat, maibe tenke jere desde etapa sedu liu ba dezenvolvimentu[12]. Ita rekoñese katak halo hanesan ne'e bele sai dezafiu liutan iha kontekstu IA duke ba software padraun, no katak Ita-nia nivel dívida téknika posivel sai altu tanba siklu dezenvolvimentu rapidu no menus estabelese ho dí'ak ba protokulu no interfaces. Ita garante Ita-nia planu siklu vida (inklui prosesu atu desativasaun sistema IA) avalia, rekoñesementu no mitiga risku sira ba futuru sistema hanesan.



3. Implantasaun seguru

Seksaun ne'e kontéin kontéin guia sira ne'ebé aplika ba etapa **implantasaun** siklu vida dezenvolvimentu sistema IA , inklui infraestrutura no modelu protesaun husi kompromete, ameasa ka perda, prosesu jerensiamentu insidente mak dezenvolve, no liberasaun responsável.

Seguru Ita-nia infraestrutura

Ita aplika prinsipiu seguransa infraestrutura di'ak ba infaestrutura mak uza iha parte hotu-hotu husi Ita-nia siklu vida sistema. Ita aplika kontrola asesu apropriadu ba Ita-nia API, modelu no dadus, no ba sira treinamentu no pipeline prosesamentu, iha peskiza no dezenvolvimentu no mós implantasaun. Ida ne'e inklui segregasaun adekuadu husi ambiente rai kódigu ka dadus konfidensial. Ida ne'e mós ajuda mitiga atake padraun seguransa sibernétika sira ne'ebé intende atu nauk modelu ida ka prejudika ninia dezempeñu.

Proteje Ita-nia modelu beibeik

Invasor sira bele konstrui funsionalidade modelu ida[13] ka dadus mak treinu ona[14], ho asesu modelu diretamente (ho obten modelu-nia toda) ka indiretamente (ho obten modelu liuhusi aplikativu ka servisu ida). Invasor sira mós bele estraga modelu sira, dadus ka prompts durante ka depois treinamentu, hodi hetan rezultadu konfiável.

Ita proteje modelu no dadus husi asesu direta no indireta, respetivamente, ho:

- implementa prátika melloria ba padraun seguransa sibernétika
- Implementa kontrola ba interface konsulta atu detekta no prevene tentativa ba asesu, modifika no esfiltra informasaun konfidensial

Atu garante katak sistema konsume hela bele valida modelu sira, Ita komputa no partilla hashes kriptográfiku no/ka asinatura arkivu modelu (por ezemplu, modelu nia todan) no konjuntu dadus (inklui pontu verifikadu) lalais depois modelu treinu ona. Hanesan mós ho kriptografia, jerensiamentu xave di'ak ne'e mak esensial [15].

Ita-nia abordajen atu mitigasun risku konfidensialidade sei depende konsideravelmente ba kazu mak uza no modelu ameasa. Aplikasaun balun, por ezemplu sira ne'eba envolve dadus sensitivu, sei presija garantia teórika ne'ebé bele sai araska ka karun atu aplika. Karik apropriadu, teknolojia ida melloria privasidade (hanesan privasidade diferensial ka kriptografia homomórfika) bele uza atu esplora ka garante nível risku asosiadu ho konsumidór, uzuáriu no invasor ba modelu no rezultadu.

Dezenvolve prosidementu jerensiamentu insidente

Inevitabilidade insidente seguransa qafeta Ita-nia sistema IA ne'e reflete iha Ita-nia planu resposta, eskalamentu no remediasaun insisdente. Ita-nia planu reflete diferente senáriu no reavaliadu regularmente hanesan sistema no investigasaun luan tan evolui. Ita armazen rekursu dijital krítiku empreza iha backups offline. Respondente treinu ona atu avalias no rezolve insidente relasiona ho IA. Ita fornese rejistru auditoria alta kualidade no rekursu seluk kao informasaun seguransa ba kliente no uzuáriu sein kustu adisional, atu permite sira-nia prosesu resposta insidente.

Libera IA responsabilidade

Ita lansa modelu, aplikativu ka sistema somente depois submetê ba avaliasaun seguransa apropriada no efikaz, hanesan benchmarking no red teaming (no mós teste sira seluk ne'ebé fora eskopu guia sira ne'e, hanesan seguransa ka imparsialidade), no Ita klaru ba Ita-nia uzuáriu kona-ba limitasaun koñesidu ka posivel modu falla. Detalle kona-ba biblioteka teste seguransa kódigu aberta fornese iha seksaun leitura liutan iha parte fin husi dokumentu fin ida ne'e.

Halo fasil ba uzuariu sira atu halo buat sira mak loos

Ita rekoñese katak kada konfigurasaun foun ka konfigurasaun foun atu avalia konjuntu ho benefísiu komersial nia deriva, no kualkér risku seguransa nia introdus. Idealmente, konfigurasaun seguransa tebtebes sei integra tama ba sistema nu'udar opsaun ida deit Kuandu konfigurasaun nesesariu, opsaun padraun tenke seguru luan tan kontra ameasa komun sira (ne'e mak, seguru tanba padraun). Ita aplika kontrola atu prevene uza ka implanta Ita-nia sistema iha maneira malisiozu.

Ita fornese uzuariu sira ho guia kona-ba uza apropriadu Ita-nia modelu ka sistema, ne'ebé inklui enfaze limitasaun no modu falla potensial. Ita deklara klaramente ba uzuariu sira kona-ba aspeitu seguransa ida ne'ebé sira tenke responsavel, no transparente kona-ba iha ne'ebé (no oinsa) sira-nia dadus bele uza, asesu ka armajen (por ezemplu, karik ida ne'e uza ba re-treinametu modelu ka revizaun ho funsinariu ka parseira sira).

4. Operasaun no manutensaun seguru

Seksaun ne'e kontéin guia sira ne'ebé aplika ba etapa **operasaun no manutensaun seguru** husi siklu vida dezenvolvimentu sistema IA Nia fornese guia ba asaun partikularmente relevante kuandu sistema ida implanta ona, inklui rejistru no monitoramentu, jerensiamentu atualizasaun no kompartilla informasaun.

Monitoriza komportamentu Ita-nia sistema

Ita mede rezultadu no dezempeñu Ita-nia modelu no sistema hanesan ne'eba ne'ebé Ita bele observa mudansa derepenti no gradual iha komportamentu mak afeita seguransa. Ita bele kontabiliza no identifika posível invasaun no komprometimentu sira, no mós desviu natural dadus.

Monitoriza entrada Ita-nia sistema

Akordu ho rekizitu privasidade no protesaun dadus, Ita monitora no rejistru tama ba Ita-nia sistema (hanesan solisitasaun inferênsia, konsulta ka prompts) atu permite obrigasaun konformidade, auditoria, investigasaun no koresaun iha kazu komprometimentu ka uza indevidu. Ida ne'e bele inklui deteksaun esplisitu husi entrada duke distribuisaun no/ka adversária, inklui sira ne'ebé tarjetu atu etapa esplora preparasaun dadus (hanesan korta no redimensionamentu imajen sira).

Segui abordajen seguru tanba dezeñu atu atualiza

Ita inklui atualizasaun automatiku ho padraun iha produtu hotu-hotu no uza prosidementu atualizasaun modular atu distribui sira. Ita atualiza prosesu (inklui rejime teste no avaliasaun) reflete faktu katak mudansa ba dadus, modelu ka prompts bele lidera ba mudansa iha komportametu sistema (por ezemplu, Ita trate atualizasaun maioria hanesan versaun foun). Ita apoiu uzuáriu sira atu avalia no responde ba mudansa modelu sira (por ezemplu ho fornese asesu previzaun no versaun API nian).

Kolleta no distribui lisaun aprende ona

Ita partisipa komunidade kompartillamentu informasaun, kolabora ho ekosistema global indústria tomak, akademia no governu sira atu kompartilla prátika melloria ida hanesan apropriadu. Ita mantein liña aberta husi komunikasaun ba feedback akordu seguransa sistema, internalmente no esternamente ba Ita-nia organizasaun, inklui fornese konkordasaun ba peskizadór seguransa sira atu halo peskiza no relata vulnerabilidade. Kuandu nesesáriu, Ita eskala problema ba komunidade luan tan, por ezemplu, publika boletin resposta divulgasaun vulnerabilidade, inklui enumerasaun detalla no kompleta vulnerabilidade komun. Ita foti asaun atu mitiga no koriji problema sira ho lalais no adekuadu.

Leitura adisional

Dezenvolvimentu IA

Prinsípiu ba seguransa mákina
Guia detalla NCSC nian kona-ba dezenvolvimentu, implantasaun ka operasaun sistema ho komponente ML ida.

Seguru tanba dezeñu - Muda Ekuilíbriu Risku Seguransa: Prinsípiu no abordajen ba Seguru tanba Software Dezeñu
Ko-autoria ho CISA, NCSC no ajênsia sira seluk, guia ida ne'e deskreve oinsa fabrikante sistema software sira, inklui IA, tenke foti etapa ba fator seguransa tama ba estájiu dezeñu dezenvolvimentu produtu no enviu produtu ne'ebé seguru atu uza.

Preokupasaun Seguransa iha Rezumu ida
Produs ho Eskritóriu Federal Alemaun Seguransa Informasaun (BSI), dokumentu ida ne'e fornese introdusaun ida ba posivel atake sira iha sistema makina aprendizajen no defensa potensial kontra atake sira ne'eba.

Prinsípiu Orientadór Internasional Prosesu Hiroshima ba Organizasaun Dezenvolve Sistema Avansadu no Kódigu Konduta Internasional Prosesu Hiroshima ba Organizasaun Dezenvolve Sistema Avansadu IA
Dokumentu sira ne'e produs nu'udar parte Prosesu IA Hiroshima G7, fornese guia ba organizasaun dezenvolve sistema IA avansadu liu, inklui modelu báziku avansadu no sistema IA jenerativu ho objetivu atu promove sistema IA seguru no konfiável iha mundu tomak.

Verifika IA
Estrutura no Ekipamentu Software IA Singapura ne'ebé valida dezempeñu husi sistema IA relasaun ho konjuntu prinsípiu rekoñesidu internasionalmente liuhusi teste padronizadu sira.

Kuadru multixamada ba Prátika Siberseguransa ba IA — ENISA (europa.eu)
Kuadru ba orienta Autoridade Nasional Kompetente sira no parte interesadu IA sira ba etapa ne'ebé sira tenke halo tuir atu proteje sira-nia sistema IA, operasaun no prosesu.

ISO 5338 Prosesu siklu vida sistema A (Iha Revizaun Hela)
Konjuntu prosesu no konseitu sira atu deskreve sira-nia siklu vida husi sistema IA baze ba aprendizadu mákina no sistema eurístiku.

Katálogu kritériu konformidade servisu Cloud IA (AIC4)
Katálogu kritériu konformidade servisu Cloud IA BSI nian fornese kritériu espesífiku IA, permite avaliasaun seguransa servisu IA iha ninia siklu vida hotu-hotu.

NIST IR 8269 (Draft) Taxonomia no terminolojia aprendizadu mákina adversáriu
Konjuntu prosesu no asosiadu konseitu sira ba deskreve siklu vida sistema IA baze ba aprendizadu mákina no sistema eurístiku.

MITRE ATLAS
Baze koñesimentu ba tátika, téknika sira no estudu kazu adversáriu ba sistema aprendizadu mákina (ML), modeladu depois liga ona ba estrutura MITRE ATT&CK.

Vizaun jeral katastrófiku ba Risku IA (2023)
Produs ho Sentru ba Seguransa IA, dokumentu ida ne'e konfigura areia sira husi risku hatudu ho IA.

Modelu Lingua Grande: Oportunidade no risku ba Indústria no Autoridade sira
Dokumentu produs ho BSI ba emprezaria, autoridade no dezenvolvedór sira ne'ebé hakarak ba aprende liutan kona-ba oportunidade no risku husi dezenvolvimentu, implantasaun no/ka uza LLM.

Projetu kódigu abertu sira atu ajuda uzuáriu sira teste seguransa modelu IA, inklui:

- Kaixa Ekipamentu Robustez Adversárial (IBM)
- CleverHans (Universidade Toronto)
- TextAttack (Universidade Virginia)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft)
- AI Verify (Infocomm Media Development Authority, Singapore)

Síber seguransa

Tarjetu Dezempeñu Seguransa Sibernétika CISA
Konjuntu komun protesaun ne'ebé entidade infraestrutura krítika hotu-hotu tenke implementa ba redus signifikativamente probilidade no impaktu risku koñesidu no téknika adversáriu.

Estrutura NCSC CAF
Estrutura Avaliasaun Sibernétika (CAF) fornese orientasaun ba organizasaun responsável ba servisu no atividade importânsia vital.

Estrutura kadeia suprimentu MITRE
Estrutura ba avaliasaun fornesedór no provedór servisu sira iha kadeia suprimentu laran.

Jerensiamentu risku

NIST AI Estrutura Jerensiamentu Risku IA NIST (IA RMF)
IA RMF esplika oinsa atu jere risku sosiotékniku sira ba individual, organizasaun no sosiedade esklusivu asosia ho IA.

ISO 27001: Seguransa informasaun, seguransa sibernétika no protesaun privasidade
Padraun ida ne'e fornese organizasaun ho guia kona-ba estabelesementu, implementasaun no manutensaun informasaun sistema jerensiamentu seguransa.

ISO 31000: Jerensiamentu risku
Padraun internasional ida ne'ebé fornese organizasaun ho guia no prinsipiu sira ba jerensiamentu risku iha organizasaun laran.

Guia Jerensiamentu Risku NCSC
Guia ida ne'e ajuda profisional risku seguransa sibernétika sira atu komprende di'ak liutan no jere risku seguransa sibernétika afeita sira-nia organizasaun.

Notas

1. Iha ne'e definidu nu'udar ema ida, autoridade publiku, ajênsia ka orgaun seluk ne'ebé dezenvolve sistema IA (ka iha sistema IA ne'ebé implanta ona) no fatin sira ne'ebé sistema iha merkadu ka implementa iha servisu laran baze ninia naran ka marka rasik
2. Ba informasaun liutan kona-ba seguru tanba dezeñu, haree CISA Seguru tanba Dezeñu pajina web no guia Muda Balansu Risku Siber seguransa: Prinsípiu no Abordajen ba Seguru tanba Software Dezeñu
3. Opozisaun ba abordajen IA naun relasiona ho ML, sistema bazea regras
4. CEPS deskreve tipu diferente sete ba interasaun dezenvolvimentu IA iha sira-nia publikasaun 'Rekonsilia Kadeia Valór IA ho Lei Intelejênsia Artifisial UE'
5. ISO/IEC 22989:2022(en) define ida ne'e nu'udar 'elementu funsional ida ne'ebé konstrui sistema IA'
6. NIST iha tarefa ho produs guia sira (no foti asaun sira seluk) atu promove seguransan, proteje no konfiável ba dezenvolvimentu no utitliza Intelijênsia Artifisial (IA). Haree Responsabilidade NIST baze iha Orden Ezekutivu, Outubru 30, 2023
7. Ba informasaun liutan kona-ba modelu ameasa ne'e disponivel husi Fundasaun OWASP
8. Haree MITRE ATLAS Adversarial Machine Learning 101
9. GitHub: RCE PoC ba Tensorflow uza kamada Lambda malisiozu
10. SLSA: 'Salvaguarda integridade artefatu ba kualkér kadeia suprimentu software hotu-hotu'
11. METI (Ministériu Ekonomia, Komérsiu no Indústria Japaun, 2023), 'Guia Introdusaun Lista Material Software (SBOM) ba Jerensiamentu Software'
12. Google research: Aprendizadu Mákina: Kartaun kréditu Juru Altu Dívida Téknika
13. Tramèr et al 2016, Nauk Modelu Aprendizadu Mákina liuhusi API Previsaun
14. Boenisch, 2020, Atake kontra Privasidade Aprendizadu Mákina (Parte 1): Model Atake Inversaun ho Estrutura IBM-ART
15. Sentru Nasional Seguransa Sibernétika, 2020, Projeta no konstrui Infraestrutura Xave Publika ospedade ho privada