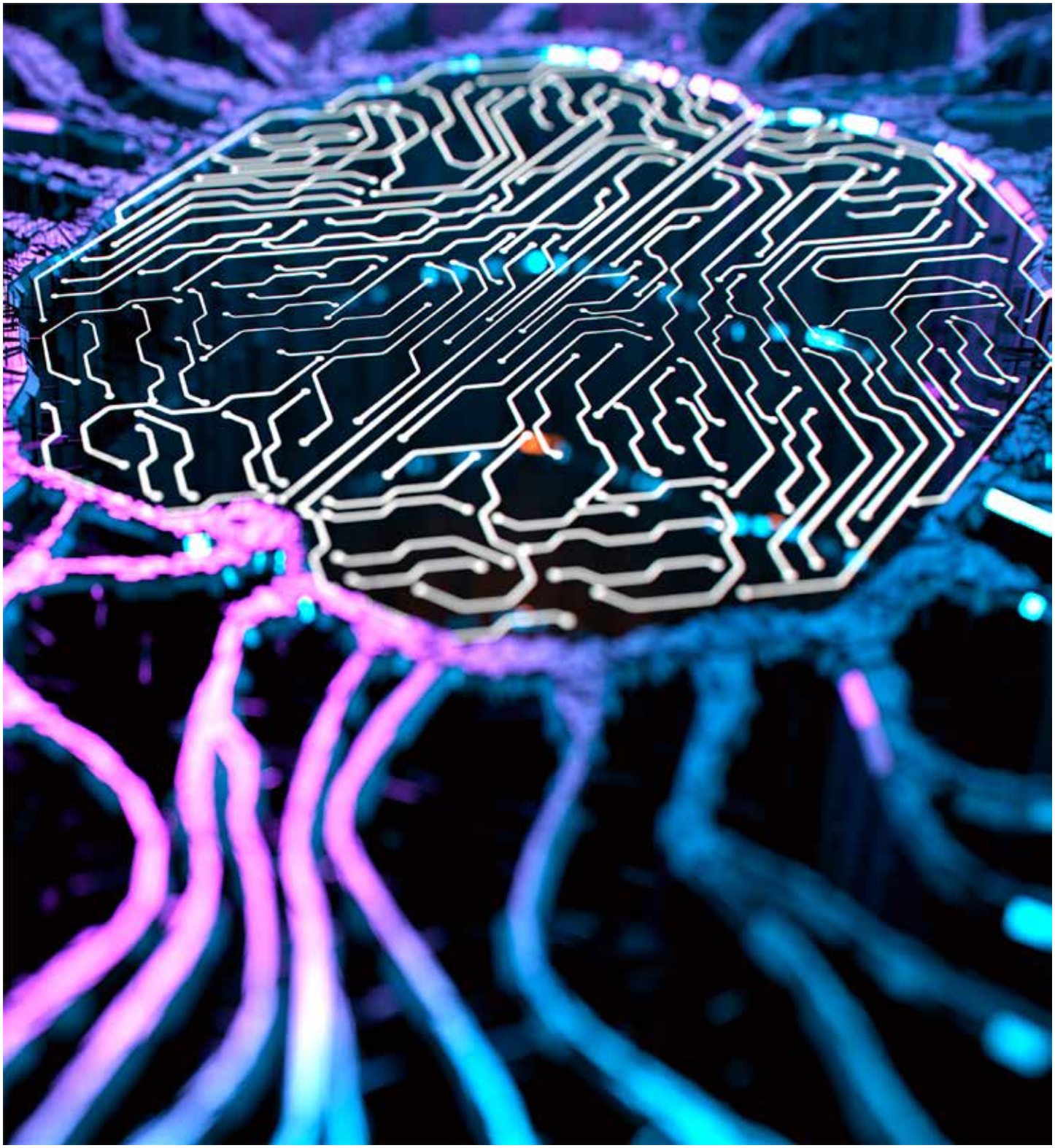
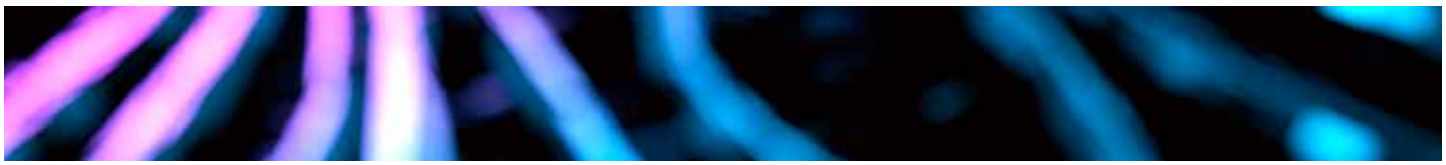
Panduan untuk pengembangan sistem AI yang aman





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

Tentang dokumen ini

Dokumen ini diterbitkan oleh National Cyber Security Centre (NCSC) Inggris, US Cybersecurity and Infrastructure Security Agency (CISA), dan mitra internasional berikut:

- 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)

Ucapan Terima Kasih

Organisasi-organisasi berikut berkontribusi terhadap pengembangan panduan ini:

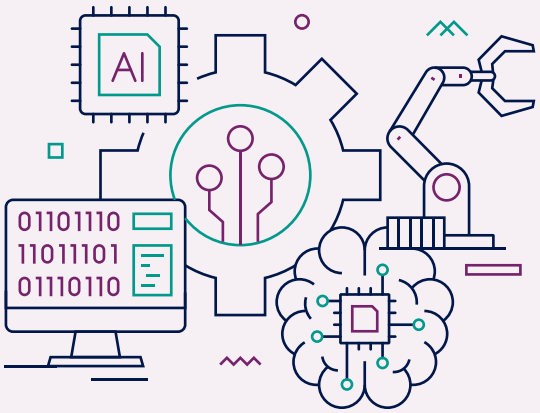
- Alan Turing Institute
- Anthropic
- Databricks
- Georgetown University's Center for Security and Emerging Technology
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute at Carnegie Mellon University
- Stanford Center for AI Safety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

Penafian

Informasi dalam dokumen ini disediakan "sebagaimana adanya" oleh NCSC dan organisasi pembuatnya yang tidak bertanggung jawab atas kehilangan, cedera, atau kerusakan apa pun yang disebabkan oleh penggunaannya kecuali diwajibkan oleh hukum. Informasi dalam dokumen ini bukan merupakan atau menyiratkan dukungan atau rekomendasi terhadap organisasi pihak ketiga, produk, atau layanan mana pun oleh NCSC dan lembaga pembuatnya. Tautan dan referensi ke situs web dan materi pihak ketiga disediakan hanya untuk informasi dan tidak mewakili dukungan atau rekomendasi sumber daya tersebut terhadap sumber lain.

Dokumen ini tersedia berdasarkan TLP:CLEAR (https://www.first.org/tlp/).

Isi



Ringkasan eksekutif

Dokumen ini merekomendasikan panduan bagi penyedia sistem apa pun yang menggunakan kecerdasan buatan (AI), baik sistem tersebut dibuat dari awal atau dibangun di atas alat dan layanan yang disediakan oleh pihak lain. Penerapan panduan ini akan membantu penyedia layanan membangun sistem AI yang berfungsi sebagaimana mestinya, tersedia saat dibutuhkan, dan berfungsi tanpa mengungkapkan data sensitif kepada pihak yang tidak berwenang.

Dokumen ini ditujukan terutama bagi penyedia sistem AI yang menggunakan model yang dihosting oleh suatu organisasi, atau menggunakan antarmuka pemrograman aplikasi (API) eksternal. Kami mendesak **semua** pemangku kepentingan (termasuk ilmuwan data, pengembang, manajer, pengambil keputusan, dan pemilik risiko) untuk membaca panduan ini guna membantu mereka mengambil keputusan yang tepat mengenai **desain**, **pengembangan**, **penerapan** dan **operasi** sistem AI mereka.

Tentang panduan

Sistem AI berpotensi membawa banyak manfaat bagi masyarakat. Namun, agar peluang AI dapat terwujud sepenuhnya, AI harus dikembangkan, diterapkan, dan dioperasikan dengan cara yang aman dan bertanggung jawab.

Sistem AI memiliki kerentanan keamanan baru yang perlu dipertimbangkan bersamaan dengan ancaman keamanan siber standar. Ketika laju pengembangan tinggi – seperti halnya AI – keamanan seringkali menjadi pertimbangan kedua. Keamanan harus menjadi persyaratan inti, tidak hanya pada tahap pengembangan, tetapi sepanjang siklus hidup sistem.

Oleh karena itu, panduan ini dibagi menjadi empat bidang utama dalam siklus hidup pengembangan sistem AI: **desain yang aman**, **pengembangan yang aman**, **penerapan yang aman**, dan **pengoperasian dan pemeliharaan yang aman**. Untuk setiap bagian, kami menyarankan pertimbangan dan mitigasi yang akan membantu mengurangi risiko keseluruhan pada proses pengembangan sistem AI organisasi.

1. **Desain yang aman**
   Bagian ini berisi panduan yang berlaku pada tahap desain siklus hidup pengembangan sistem AI. Hal ini mencakup pemahaman risiko dan pemodelan ancaman, serta topik spesifik dan trade-off yang perlu dipertimbangkan dalam desain sistem dan model.

2. **Pengembangan yang aman**
   Bagian ini berisi panduan yang berlaku pada tahap pengembangan siklus hidup pengembangan sistem AI, termasuk keamanan rantai pasokan, dokumentasi, serta pengelolaan aset dan utang teknis.

3. **Penerapan yang aman**
   Bagian ini berisi panduan yang berlaku pada tahap penerapan siklus hidup pengembangan sistem AI, termasuk melindungi infrastruktur dan model dari gangguan, ancaman, atau kerugian, mengembangkan proses manajemen insiden, dan rilis yang bertanggung jawab.

4. **Pengoperasian dan pemeliharaan yang aman**
   Bagian ini berisi panduan yang berlaku untuk tahap pengoperasian dan pemeliharaan yang aman pada siklus hidup pengembangan sistem AI. Panduan ini memberikan panduan mengenai tindakan yang sangat relevan setelah sistem diterapkan, termasuk pencatatan dan pemantauan, pengelolaan pembaruan, dan berbagi informasi.

Panduan ini mengikuti pendekatan 'aman secara default', dan selaras dengan praktik yang ditetapkan dalam Panduan pengembangan dan penerapan yang aman NCSS, Kerangka Pengembangan Perangkat Lunak Aman NIST, dan 'aman berdasarkan prinsip desain' yang diterbitkan oleh CISA, NCSC, dan badan siber internasional. Mereka memprioritaskan:

- mengambil kepemilikan atas hasil keamanan bagi pelanggan
- merangkul transparansi dan akuntabilitas yang radikal
- membangun struktur organisasi dan kepemimpinan yang aman secara desain adalah prioritas bisnis utama

Pengantar

Sistem kecerdasan buatan (AI) berpotensi membawa banyak manfaat bagi masyarakat. Namun, agar peluang AI dapat terwujud sepenuhnya, AI harus dikembangkan, diterapkan, dan dioperasikan dengan cara yang aman dan bertanggung jawab. Keamanan siber adalah prasyarat penting untuk keselamatan, ketahanan, privasi, keadilan, kemanjuran dan keandalan sistem AI.

Namun, sistem AI mempunyai kerentanan keamanan baru yang perlu dipertimbangkan bersamaan dengan ancaman keamanan siber standar. Ketika laju pengembangan tinggi – seperti halnya AI – keamanan sering kali menjadi pertimbangan kedua. Keamanan harus menjadi persyaratan inti, tidak hanya dalam tahap pengembangan, namun sepanjang siklus hidup sistem.

Dokumen ini merekomendasikan panduan bagi penyedia[1] sistem apa pun yang menggunakan AI, baik sistem tersebut dibuat dari awal atau dibuat berdasarkan alat dan layanan yang disediakan oleh yang lain. Penerapan panduan ini akan membantu penyedia membangun sistem AI yang berfungsi sebagaimana mestinya, tersedia saat dibutuhkan, dan berfungsi tanpa mengungkapkan data sensitif kepada pihak yang tidak berwenang.

Panduan ini harus dipertimbangkan bersamaan dengan keamanan siber, manajemen risiko, dan praktik terbaik respons insiden. Secara khusus, kami mendesak penyedia layanan untuk mengikuti prinsip 'aman berdasarkan desain'[2] yang dikembangkan oleh Badan Keamanan Siber dan Infrastruktur AS - US Cybersecurity and Infrastructure Security Agency (CISA), Pusat Keamanan Siber Nasional Inggris – UK National Cyber Security Centre (NCSC), dan semua mitra internasional kami . Prinsip yang diprioritaskan adalah:

- mengambil kepemilikan atas hasil keamanan bagi pelanggan
- merangkul transparansi dan akuntabilitas yang radikal
- membangun struktur organisasi dan kepemimpinan yang aman secara desain adalah prioritas bisnis utama

Mengikuti prinsip 'aman berdasarkan desain' memerlukan sumber daya yang signifikan di seluruh siklus hidup sistem. Ini berarti pengembang harus berinvestasi dalam memprioritaskan **fitur**, **mekanisme**, dan **implementasi** alat yang melindungi pelanggan di setiap lapisan desain sistem, dan di seluruh tahapan siklus hidup perkembangan. Melakukan hal ini akan mencegah desain ulang yang mahal di kemudian hari, serta melindungi pelanggan dan data mereka dalam jangka pendek.

Mengapa keamanan AI berbeda?

Dalam dokumen ini kami menggunakan 'AI' untuk merujuk secara khusus pada aplikasi pembelajaran mesin (ML)[3]. Semua jenis ML ada dalam cakupannya. Kami mendefinisikan aplikasi ML sebagai aplikasi yang:

- melibatkan komponen perangkat lunak (model) yang memungkinkan komputer mengenali dan membawa konteks ke pola dalam data tanpa aturan yang harus diprogram secara eksplisit oleh manusia.
- menghasilkan prediksi, rekomendasi, atau keputusan berdasarkan alasan statistik

Selain ancaman keamanan siber yang ada, sistem AI juga mempunyai kerentanan baru. Istilah 'adversarial machine learning' (AML), digunakan untuk menggambarkan eksploitasi kerentanan mendasar dalam komponen ML, termasuk perangkat keras, perangkat lunak, alur kerja, dan rantai pasokan. AML memungkinkan penyerang menyebabkan perilaku yang tidak diinginkan dalam sistem ML yang dapat mencakup:

- memengaruhi klasifikasi model atau kinerja regresi
- memungkinkan pengguna untuk melakukan tindakan yang tidak sah
- menyadap informasi model yang sensitif

Ada banyak cara untuk mencapai efek ini, seperti serangan injeksi cepat di domain Large Language Model (LLM), atau dengan sengaja merusak data pelatihan atau umpan balik pengguna (dikenal sebagai 'keracunan data').

Siapa yang harus membaca dokumen ini?

Dokumen ini ditujukan terutama pada penyedia sistem AI, baik berdasarkan model yang dihosting oleh suatu organisasi atau menggunakan antarmuka pemrograman aplikasi (API) eksternal. Namun, kami mendesak **semua** pemangku kepentingan (termasuk ilmuwan data, pengembang, manajer, pengambil keputusan, dan pemilik risiko) untuk membaca panduan ini guna membantu mereka mengambil keputusan yang tepat mengenai **desain**, **penerapan** dan **pengoperasian** sistem AI pembelajaran mesin mereka.

Meskipun demikian, tidak semua panduan ini dapat diterapkan secara langsung pada semua organisasi. Tingkat kecanggihan dan metode serangan akan bervariasi tergantung pada penyerang yang menargetkan sistem AI, sehingga penggunaan panduan ini harus dipertimbangkan bersamaan dengan kasus dan profil ancaman terhadap organisasi Anda.

Siapa yang bertanggung jawab mengembangkan AI yang aman?

Seringkali terdapat banyak aktor dalam rantai pasokan AI modern. Pendekatan sederhana mengasumsikan dua entitas:

- ‘penyedia’ yang bertanggung jawab atas kurasi data, pengembangan algoritmik, desain, penerapan, dan pemeliharaan
- ‘pengguna’, yang memberikan masukan dan menerima keluaran.

Meskipun pendekatan penyedia-pengguna ini digunakan di banyak aplikasi, hal ini menjadi semakin jarang[4], karena penyedia mungkin berupaya menggabungkan perangkat lunak, data, model, dan/atau layanan jarak jauh yang disediakan oleh pihak ketiga ke dalam sistem mereka sendiri. Rantai pasokan yang rumit ini mempersulit pengguna akhir untuk memahami tanggung jawab atas AI yang aman.

Pengguna (baik ‘pengguna akhir’, atau penyedia yang menggunakan komponen AI eksternal[5]) biasanya tidak memiliki visibilitas dan/atau keahlian yang memadai untuk sepenuhnya memahami, mengevaluasi, atau mengatasi risiko yang terkait dengan sistem yang mereka gunakan. Oleh karena itu, sejalan dengan prinsip ‘aman berdasarkan desain’, **penyedia komponen AI harus bertanggung jawab atas hasil keamanan pengguna di bagian bawah rantai pasokan**.

Penyedia harus menerapkan kontrol dan mitigasi keamanan jika memungkinkan dalam model, saluran pipa dan/atau sistem mereka, dan jika pengaturan digunakan, terapkan opsi paling aman sebagai default. Jika risiko tidak dapat dimitigasi, penyedia layanan harus bertanggung jawab untuk:

- memberi tahu pengguna di bagian bawah rantai pasokan mengenai risiko yang mereka dan (jika berlaku) diterima oleh pengguna mereka sendiri
- menasihati mereka tentang cara menggunakan komponen dengan aman

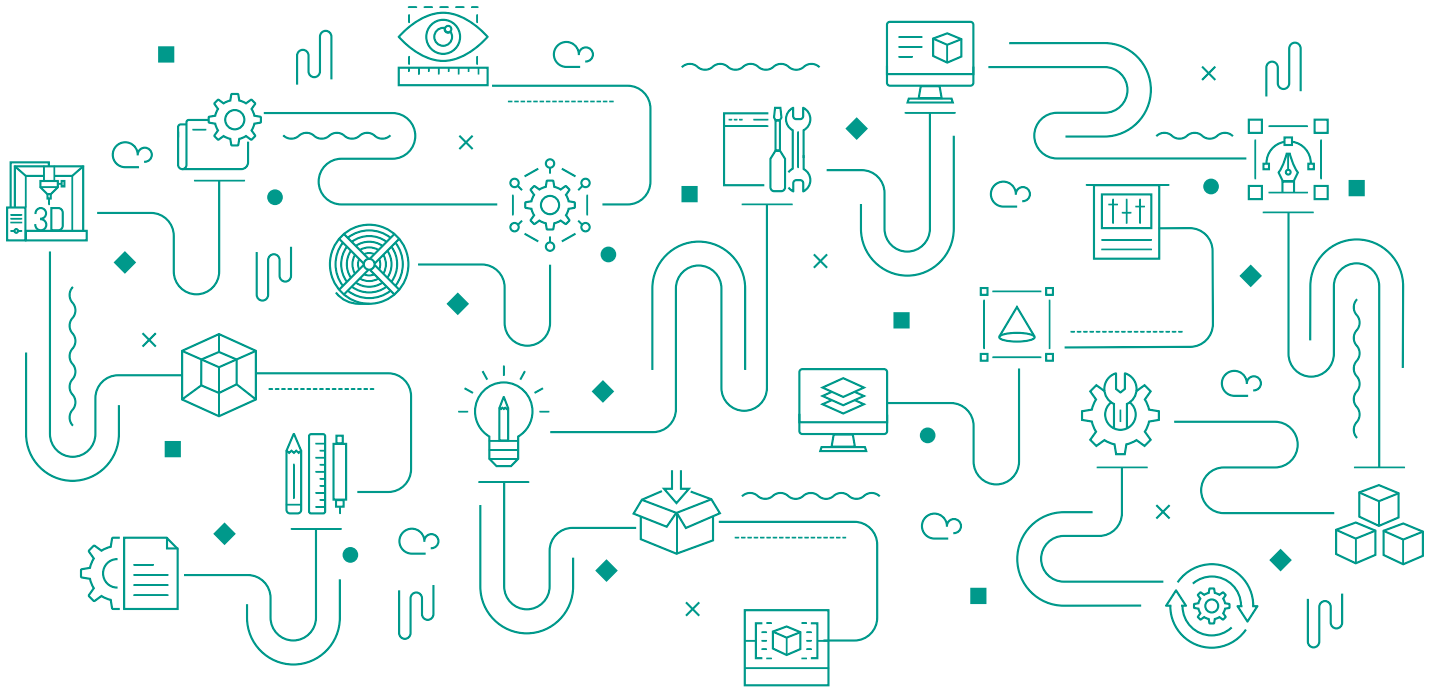
Jika penyusupan sistem dapat menyebabkan kerusakan fisik atau reputasi yang nyata atau meluas, kerugian besar pada operasi bisnis, kebocoran informasi sensitif atau rahasia dan/atau implikasi hukum, risiko keamanan siber AI harus dianggap **amat penting**.

Panduan untuk pengembangan sistem AI yang aman

Panduan ini dibagi menjadi empat bidang utama dalam siklus hidup pengembangan sistem AI: **desain yang aman**, **pengembangan yang aman**, **penerapan yang aman**, dan **operasi dan pemeliharaan yang aman**. Untuk setiap area, kami menyarankan pertimbangan dan mitigasi yang akan membantu mengurangi risiko keseluruhan terhadap proses pengembangan sistem AI organisasi.

Panduan yang ditetapkan dalam dokumen ini selaras dengan praktik siklus hidup pengembangan perangkat lunak yang didefinisikan dalam:

- NCSC Panduan pengembangan dan penerapan yang aman
- National Institute of Standards and Technology (NIST) Software Development Framework (SSDF)[6]



1. Desain yang aman

Bagian ini berisi panduan yang berlaku pada tahap **desain** siklus hidup pengembangan sistem AI. Ini mencakup pemahaman risiko dan pemodelan ancaman, serta topik spesifik dan trade-off yang perlu dipertimbangkan pada desain sistem dan model.

Meningkatkan kesadaran staf akan ancaman dan risiko

Pemilik sistem dan pemimpin senior memahami ancaman terhadap keamanan AI dan mitigasinya. Ilmuwan dan pengembang data Anda menjaga kesadaran akan ancaman keamanan yang relevan dan mode kegagalan serta membantu pemilik risiko untuk membuat keputusan yang tepat. Anda memberikan panduan kepada pengguna tentang risiko keamanan unik yang dihadapi sistem AI (misalnya, sebagai bagian dari pelatihan InfoSec standar) dan melatih pengembang dalam teknik pengkodean yang aman serta praktik AI yang aman dan bertanggung jawab.

Modelkan ancaman terhadap sistem Anda

Sebagai bagian dari proses manajemen risiko, Anda menerapkan proses holistik untuk menilai ancaman terhadap sistem Anda, yang mencakup pemahaman potensi dampak terhadap sistem, pengguna, organisasi, dan masyarakat luas jika komponen AI disusupi atau berperilaku tidak terduga[7]. Proses ini melibatkan penilaian dampak ancaman khusus AI[8] dan mendokumentasikan pengambilan keputusan Anda.

Anda menyadari bahwa sensitivitas dan jenis data yang digunakan dalam sistem Anda dapat memengaruhi nilainya sebagai target penyerang. Penilaian Anda harus mempertimbangkan bahwa beberapa ancaman mungkin muncul ketika sistem AI semakin dipandang sebagai target bernilai tinggi, dan karena AI itu sendiri memungkinkan vektor serangan otomatis yang baru.

Rancang sistem Anda untuk keamanan serta fungsionalitas dan kinerja

Anda yakin bahwa tugas yang ada dapat diselesaikan dengan paling tepat menggunakan AI. Setelah menentukan hal ini, Anda menilai kesesuaian pilihan desain khusus AI Anda. Anda mempertimbangkan model ancaman dan mitigasi keamanan terkait serta fungsionalitas, pengalaman pengguna, lingkungan penerapan, kinerja, jaminan, pengawasan, persyaratan etika dan hukum, serta pertimbangan lainnya. Misalnya:

- Anda mempertimbangkan keamanan rantai pasokan ketika memilih apakah akan mengembangkan sendiri atau menggunakan komponen eksternal, misalnya:
    - pilihan Anda untuk melatih model baru, menggunakan model yang sudah ada (dengan atau tanpa penyesuaian) atau mengakses model melalui API eksternal sesuai dengan kebutuhan Anda
    - pilihan Anda untuk bekerja sama dengan penyedia model eksternal mencakup evaluasi uji tuntas terhadap postur keamanan penyedia tersebut
    - jika menggunakan pustaka eksternal, Anda menyelesaikan evaluasi uji tuntas (misalnya, untuk memastikan pustaka memiliki kontrol yang mencegah sistem memuat model yang tidak tepercaya tanpa langsung mengekspos dirinya ke eksekusi kode arbitrer[9])
    - Anda menerapkan pemindaian dan isolasi/sandboxing saat mengimpor model pihak ketiga atau bobot serial, yang harus diperlakukan sebagai kode pihak ketiga yang tidak tepercaya dan dapat mengaktifkan eksekusi kode jarak jauh

  - jika menggunakan API eksternal, Anda menerapkan kontrol yang sesuai pada data yang dapat dikirim ke layanan di luar kendali organisasi Anda, seperti mengharuskan pengguna untuk masuk dan mengonfirmasi sebelum mengirim informasi yang berpotensi sensitif
  - Anda menerapkan pemeriksaan dan sanitasi data dan masukan yang sesuai; hal ini termasuk saat menggabungkan umpan balik dari pengguna atau data pembelajaran berkelanjutan ke dalam model Anda, dengan menyadari bahwa data pelatihan menentukan perilaku sistem

- Anda mengintegrasikan pengembangan sistem perangkat lunak AI ke dalam praktik terbaik pengembangan dan operasi aman yang ada; semua elemen sistem AI ditulis dalam lingkungan yang sesuai dengan menggunakan praktik pengkodean dan bahasa yang mengurangi atau menghilangkan kelompok kerentanan yang diketahui jika masuk akal

- jika komponen AI perlu memicu tindakan, misalnya mengubah file atau mengarahkan keluaran ke sistem eksternal, Anda menerapkan pembatasan yang sesuai pada tindakan yang mungkin dilakukan (ini mencakup pengamanan kegagalan AI dan non-AI eksternal jika diperlukan)

- keputusan seputar interaksi pengguna dipengaruhi oleh risiko spesifik AI, misalnya:

  - sistem Anda memberikan keluaran yang dapat digunakan kepada pengguna tanpa mengungkapkan tingkat detail yang tidak perlu kepada calon penyerang
  - jika perlu, sistem Anda menyediakan pagar pembatas yang efektif di sekitar keluaran model
  - jika menawarkan API kepada pelanggan atau kolaborator eksternal, Anda menerapkan kontrol yang sesuai untuk memitigasi serangan pada sistem AI melalui API
  - Anda mengintegrasikan pengaturan paling aman ke dalam sistem secara default
  - Anda menerapkan prinsip hak istimewa terkecil untuk membatasi akses ke fungsionalitas sistem
  - Anda menjelaskan kemampuan yang lebih berisiko kepada pengguna dan mengharuskan pengguna untuk ikut serta menggunakannya; Anda mengomunikasikan kasus penggunaan yang dilarang, dan, jika memungkinkan, memberi tahu pengguna tentang solusi alternatif

Pertimbangkan manfaat dan keuntungan keamanan saat memilih model AI Anda

Pilihan model AI Anda akan melibatkan penyeimbangan berbagai persyaratan. Ini termasuk pilihan arsitektur model, konfigurasi, data pelatihan, algoritma pelatihan, dan hyperparameter. Keputusan Anda didasarkan pada model ancaman Anda, dan secara berkala dinilai ulang seiring kemajuan penelitian keamanan AI dan pemahaman tentang ancaman tersebut.

Saat memilih model AI, pertimbangan Anda kemungkinan besar mencakup, namun tidak terbatas pada:

- kompleksitas model yang Anda gunakan, yaitu arsitektur yang dipilih dan jumlah parameter; arsitektur model yang Anda pilih dan jumlah parameternya, antara lain, akan memengaruhi jumlah data pelatihan yang diperlukan dan kekuatan model tersebut terhadap perubahan data masukan saat digunakan

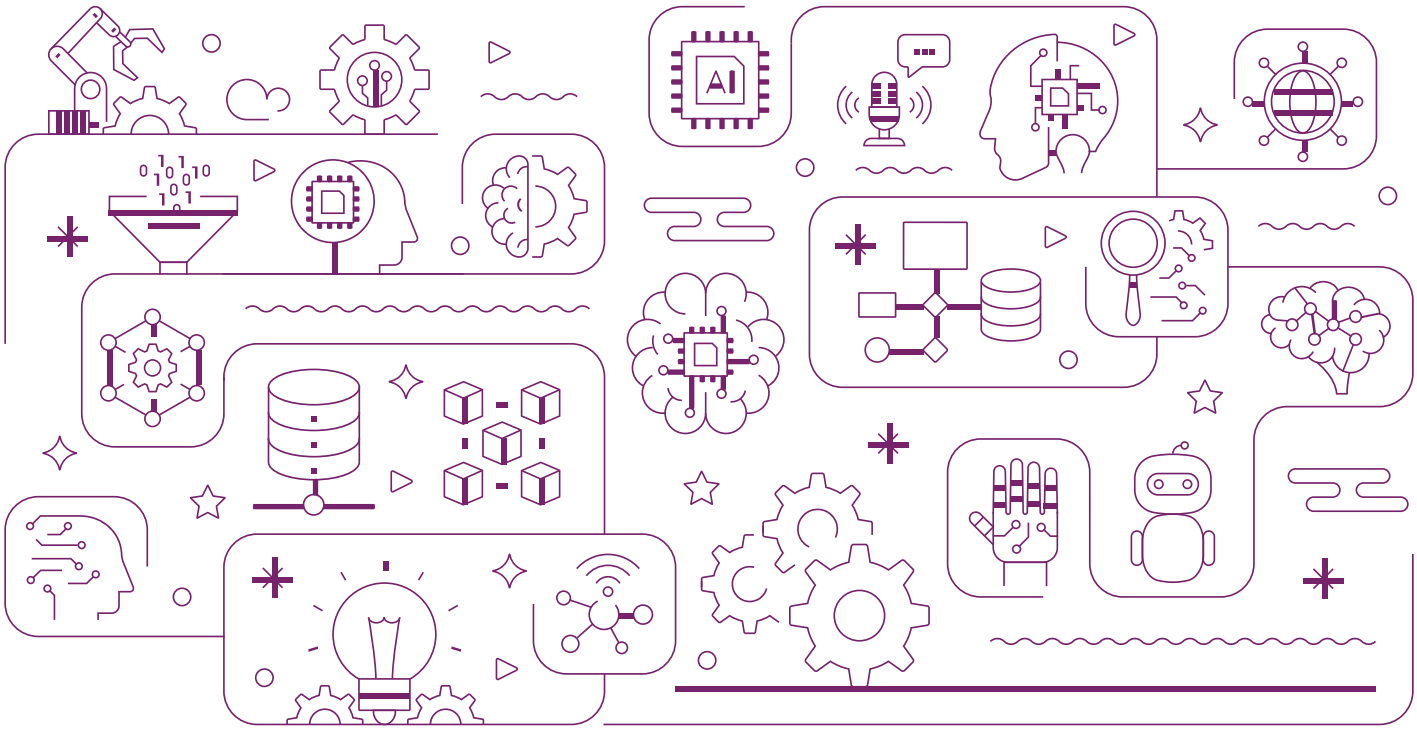
- kesesuaian model untuk kasus penggunaan Anda dan/atau kelayakan untuk mengadaptasinya sesuai kebutuhan spesifik Anda (misalnya dengan melakukan penyesuaian)

- kemampuan untuk menyelaraskan, menafsirkan, dan menjelaskan keluaran model Anda (misalnya untuk proses debug, audit atau kepatuhan terhadap peraturan); mungkin ada keuntungan menggunakan model yang lebih sederhana dan transparan dibandingkan model yang besar dan kompleks yang lebih sulit diinterpretasikan

- karakteristik kumpulan data pelatihan, termasuk ukuran, integritas, kualitas, sensitivitas, usia, relevansi, dan keragaman

- nilai penggunaan model pengerasan (seperti pelatihan permusuhan), regularisasi dan/atau teknik peningkatan privasi
- asal dan rantai pasokan komponen termasuk model atau model dasar, data pelatihan, dan alat terkait

Untuk informasi selengkapnya tentang seberapa banyak faktor ini memengaruhi hasil keamanan, lihat 'Prinsip Keamanan Pembelajaran Mesin' NCSC, khususnya Desain untuk keamanan (arsitektur model).



2. Pembangunan yang aman

Bagian ini berisi panduan yang berlaku pada tahap **pengembangan** siklus hidup pengembangan sistem AI, termasuk keamanan rantai pasokan, dokumentasi, serta pengelolaan utang aset dan teknis.

Amankan rantai pasokan Anda

Anda menilai dan memantau keamanan rantai pasokan AI Anda di seluruh siklus hidup sistem, dan mengharuskan pemasok untuk mematuhi standar yang sama yang diterapkan organisasi Anda pada perangkat lunak lain. Jika pemasok tidak dapat mematuhi standar organisasi Anda, Anda bertindak sesuai dengan kebijakan manajemen risiko yang ada.

Jika tidak diproduksi sendiri, Anda memperoleh dan memelihara komponen perangkat keras dan perangkat lunak yang aman dan terdokumentasi dengan baik (misalnya, model, data, pustaka perangkat lunak, modul, middleware, kerangka kerja, dan API eksternal) dari komersial terverifikasi, sumber terbuka, dan pengembang pihak ketiga lainnya untuk memastikan keamanan yang kuat di sistem Anda.

Anda siap melakukan failover (kegagalan) ke solusi alternatif untuk sistem yang sangat penting, jika kriteria keamanan tidak terpenuhi. Anda menggunakan sumber daya seperti Panduan Rantai Pasokan NCSC dan kerangka kerja seperti Tingkat Rantai Pasokan untuk Artefak Perangkat Lunak (SLSA)[10] untuk melacak pengesahan rantai pasokan dan perangkat lunak perkembangan siklus hidup.

Identifikasi, lacak, dan lindungi aset Anda

Anda memahami nilai aset terkait AI bagi organisasi Anda, termasuk model, data (termasuk umpan balik pengguna), perintah, perangkat lunak, dokumentasi, log, dan penilaian (termasuk informasi tentang kemampuan yang berpotensi tidak aman dan mode kegagalan), dengan mengenali di mana aset tersebut mewakili signifikan investasi dan di mana mengakses mereka memungkinkan penyerang. Anda memperlakukan log sebagai data sensitif dan menerapkan kontrol untuk melindungi kerahasiaan, integritas, dan ketersediaannya.

Anda mengetahui di mana aset Anda berada dan telah menilai serta menerima segala risiko terkait. Anda memiliki proses dan alat untuk melacak, mengautentikasi, mengontrol versi, dan mengamankan aset Anda, serta dapat memulihkan ke kondisi baik yang diketahui jika terjadi kompromi.

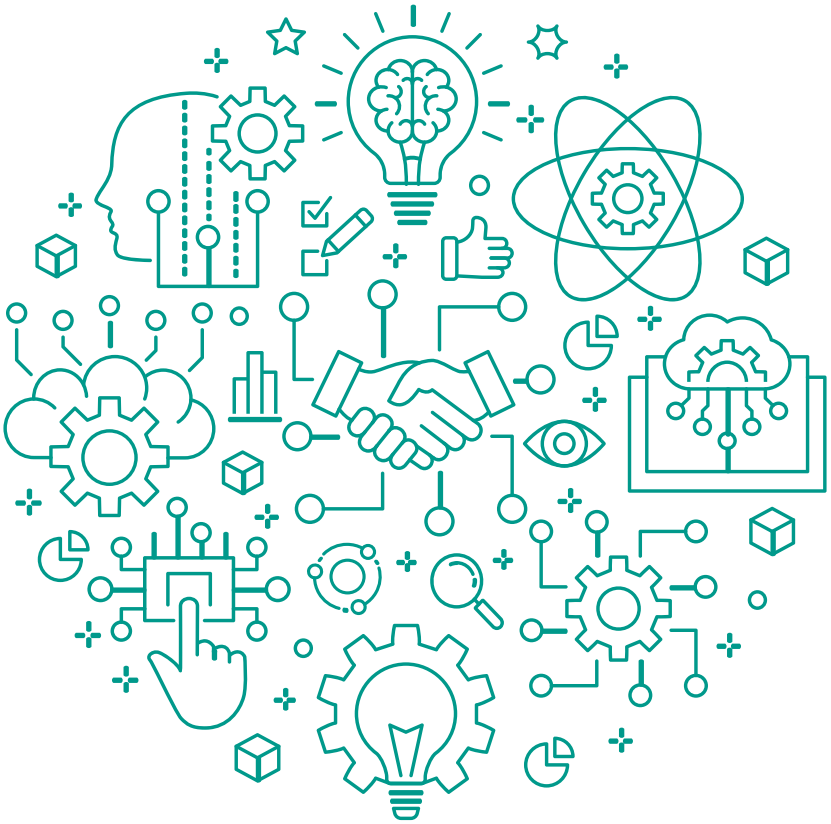
Anda memiliki proses dan kontrol untuk mengelola data apa yang dapat diakses oleh sistem AI, dan untuk mengelola konten yang dihasilkan oleh AI sesuai dengan sensitivitasnya (dan sensitivitas masukan yang digunakan untuk menghasilkannya).

Dokumentasikan data, model, dan petunjuk Anda

Anda mendokumentasikan pembuatan, operasi, dan manajemen siklus hidup model, kumpulan data, dan perintah meta atau sistem apa pun. Dokumentasi Anda mencakup informasi terkait keamanan seperti sumber data pelatihan (termasuk data penyesuaian dan umpan balik manusia atau operasional lainnya), cakupan dan batasan yang dimaksudkan, pagar pembatas, hash atau tanda tangan kriptografi, waktu retensi, frekuensi peninjauan yang disarankan, dan potensi mode kegagalan. Struktur yang berguna untuk membantu melakukan hal ini termasuk kartu model, kartu data, dan bill of material perangkat lunak (SBOM). Produksi dokumentasi yang komprehensif mendukung transparansi dan akuntabilitas[11].

Kelola utang teknis Anda

Seperti halnya sistem perangkat lunak apa pun, Anda mengidentifikasi, melacak, dan mengelola 'utang teknis' Anda di sepanjang siklus hidup sistem AI (utang teknis adalah tempat pengambilan keputusan teknis yang tidak memenuhi praktik terbaik untuk mencapai hasil jangka pendek, dengan mengorbankan jangka waktu yang lebih lama). Seperti halnya utang finansial, utang teknis pada dasarnya tidak buruk, namun harus dikelola sejak tahap awal pengembangannya[12]. Anda menyadari bahwa melakukan hal ini bisa menjadi lebih menantang dalam konteks AI dibandingkan dengan perangkat lunak standar, dan bahwa tingkat utang teknis Anda kemungkinan besar akan tinggi karena siklus pengembangan yang cepat serta kurangnya protokol dan antarmuka yang mapan. Anda memastikan rencana siklus hidup Anda (termasuk proses untuk menonaktifkan sistem AI) menilai, mengetahui, dan memitigasi risiko terhadap sistem serupa di masa depan.



3. Penerapan yang aman

Bagian ini berisi panduan yang berlaku pada tahap **penerapan** siklus hidup pengembangan sistem AI, termasuk melindungi infrastruktur dan model dari gangguan, ancaman, atau kerugian, mengembangkan proses manajemen insiden, dan pelepasan yang bertanggung jawab.

Amankan infrastruktur Anda

Anda menerapkan prinsip keamanan infrastruktur yang baik pada infrastruktur yang digunakan di setiap bagian siklus hidup sistem Anda.. Anda menerapkan kontrol akses yang sesuai pada API, model, dan data Anda, serta pada jalur pelatihan dan pemrosesannya, dalam penelitian dan pengembangan serta penerapan. Hal ini mencakup pemisahan yang tepat terhadap lingkungan yang menyimpan kode atau data sensitif. Hal ini juga akan membantu memitigasi serangan keamanan siber standar yang bertujuan untuk mencuri model atau merusak kinerjanya.

Lindungi model Anda secara terus menerus

Penyerang mungkin dapat merekonstruksi fungsionalitas model[13] atau data tempat model tersebut dilatih[14], dengan mengakses model secara langsung (dengan memperoleh bobot model) atau secara tidak langsung (dengan menanyakan model melalui aplikasi atau layanan). Penyerang juga dapat merusak model, data, atau perintah selama atau setelah pelatihan, sehingga menghasilkan keluaran yang tidak dapat dipercaya.

Anda melindungi model dan data dari akses langsung dan tidak langsung, dengan:

- menerapkan praktik terbaik keamanan siber standar
- menerapkan kontrol pada antarmuka kueri untuk mendeteksi dan mencegah upaya mengakses, mengubah, dan menyaring informasi rahasia

Untuk memastikan bahwa sistem yang menggunakan dapat memvalidasi model, Anda menghitung dan membagikan hash kriptografi dan/atau tanda tangan file model (misalnya, bobot model) dan kumpulan data (termasuk pos pemeriksaan) segera setelah model dilatih. Seperti halnya kriptografi, pengelolaan kunci yang baik sangatlah penting[15].

Pendekatan Anda terhadap mitigasi risiko kerahasiaan akan sangat bergantung pada kasus penggunaan dan model ancaman. Beberapa aplikasi, misalnya aplikasi yang melibatkan data yang sangat sensitif, mungkin memerlukan jaminan teoretis yang mungkin sulit atau mahal untuk diterapkan. Jika memungkinkan, teknologi yang meningkatkan privasi (seperti privasi diferensial atau enkripsi homomorfik) dapat digunakan untuk mengeksplorasi atau memastikan tingkat risiko yang terkait dengan konsumen, pengguna, dan penyerang yang memiliki akses ke model dan keluaran.

Mengembangkan prosedur manajemen insiden

Insiden keamanan yang tidak dapat dihindari yang memengaruhi sistem AI Anda tercermin dalam rencana respons, eskalasi, dan remediasi insiden Anda. Rencana Anda mencerminkan skenario yang berbeda-beda dan secara berkala dinilai ulang seiring dengan berkembangnya sistem dan penelitian yang lebih luas. Anda menyimpan sumber daya digital penting perusahaan dalam cadangan offline. Responden telah dilatih untuk menilai dan menangani insiden terkait AI. Anda memberikan log audit berkualitas tinggi dan fitur atau informasi keamanan lainnya kepada pelanggan dan pengguna tanpa biaya tambahan, untuk mengaktifkan proses respons insiden mereka.

Rilis AI secara bertanggung jawab

Anda merilis model, aplikasi, atau sistem hanya setelah melakukan evaluasi keamanan yang sesuai dan efektif seperti benchmarking dan tim merah (serta pengujian lain yang berada di luar cakupan panduan ini, seperti keselamatan atau keadilan), dan Anda diizinkan untuk melakukannya kepada. pengguna Anda tentang batasan yang diketahui atau potensi mode kegagalan. Detail pustaka pengujian keamanan sumber terbuka diberikan di bagian bacaan lebih lanjut di akhir dokumen ini.

Permudah pengguna untuk melakukan hal yang benar

Anda menyadari bahwa setiap opsi pengaturan atau konfigurasi baru harus dinilai sehubungan dengan manfaat bisnis yang diperolehnya, dan risiko keamanan apa pun yang ditimbulkannya. Idealnya, pengaturan yang paling aman akan diintegrasikan ke dalam sistem sebagai satu-satunya pilihan. Ketika diperlukan konfigurasi, opsi default harus aman secara luas terhadap ancaman umum (yaitu, aman secara default). Anda menerapkan kontrol untuk mencegah penggunaan atau penerapan sistem Anda dengan cara yang berbahaya.

Anda memberikan panduan kepada pengguna tentang penggunaan model atau sistem Anda dengan tepat, termasuk menyoroti batasan dan potensi mode kegagalan. Anda menyatakan dengan jelas kepada pengguna aspek keamanan apa yang menjadi tanggung jawab mereka, dan transparan tentang di mana (dan bagaimana) data mereka dapat digunakan, diakses, atau disimpan (misalnya, jika data tersebut digunakan untuk pelatihan ulang model, atau ditinjau oleh karyawan atau mitra).

4. Pengoperasian dan pemeliharaan yang aman

Bagian ini berisi panduan yang berlaku pada tahap **operasi dan pemeliharaan yang aman** dalam siklus hidup pengembangan sistem AI. Bagian ini memberikan panduan mengenai tindakan yang sangat relevan setelah sistem diterapkan, termasuk pencatatan dan pemantauan, pengelolaan pembaruan, dan pembagian informasi.

Pantau perilaku sistem Anda

Anda mengukur keluaran dan kinerja model dan sistem Anda sedemikian rupa sehingga Anda dapat mengamati perubahan perilaku yang memengaruhi keamanan secara tiba-tiba dan bertahap. Anda dapat memperhitungkan dan mengidentifikasi potensi intrusi dan kompromi, serta penyimpangan data alami.

Pantau masukan sistem Anda

Sejalan dengan persyaratan privasi dan perlindungan data, Anda memantau dan mencatat masukan ke sistem Anda (seperti permintaan inferensi, pertanyaan, atau perintah) untuk mengaktifkan kewajiban kepatuhan, audit, investigasi, dan remediasi jika terjadi pelanggaran atau penyalahgunaan. Hal ini dapat mencakup deteksi eksplisit atas masukan yang tidak didistribusikan dan/atau masukan yang berlawanan, termasuk masukan yang bertujuan untuk mengeksploitasi langkah-langkah persiapan data (seperti memotong dan mengubah ukuran gambar).

Ikuti pendekatan desain yang aman terhadap pembaruan

Anda menyertakan pembaruan otomatis secara default di setiap produk dan menggunakan prosedur pembaruan modular yang aman untuk mendistribusikannya. Proses pembaruan Anda (termasuk sistem pengujian dan evaluasi) mencerminkan fakta bahwa perubahan pada data, model, atau perintah dapat menyebabkan perubahan dalam perilaku sistem (misalnya, Anda memperlakukan pembaruan besar seperti versi baru). Anda mendukung pengguna untuk mengevaluasi dan merespons perubahan model (misalnya dengan menyediakan akses pratinjau dan API berversi).

Kumpulkan dan bagikan pelajaran yang didapat

Anda berpartisipasi dalam komunitas berbagi informasi, berkolaborasi dengan seluruh ekosistem global industri, akademisi, dan pemerintah untuk berbagi praktik terbaik jika diperlukan. Anda menjaga jalur komunikasi terbuka untuk mendapatkan umpan balik mengenai keamanan sistem, baik secara internal maupun eksternal organisasi Anda, termasuk memberikan persetujuan kepada peneliti keamanan untuk meneliti dan melaporkan kerentanan. Bila diperlukan, Anda mengeskalasi permasalahan ke komunitas yang lebih luas, misalnya menerbitkan buletin yang menanggapi pengungkapan kerentanan, termasuk enumerasi kerentanan umum yang terperinci dan lengkap. Anda mengambil tindakan untuk memitigasi dan memulihkan masalah dengan cepat dan tepat.

Bacaan lebih lanjut

Perkembangan AI

Prinsip keamanan pembelajaran mesin
Panduan terperinci NCSC tentang pengembangan, penerapan, atau pengoperasian sistem dengan komponen ML.

Secure by Design - Menggeser Keseimbangan Risiko Keamanan Siber: Prinsip dan Pendekatan untuk Perangkat Lunak Secure by Design
Ditulis bersama oleh CISA, NCSC, dan lembaga lainnya, panduan ini menjelaskan bagaimana produsen sistem perangkat lunak, termasuk AI, harus mengambil langkah-langkah untuk memasukkan faktor keamanan pada tahap desain dalam pengembangan produk, dan mengirimkan produk yang sudah aman saat dikeluarkan dari kemasannya.

Sekilas Masalah Keamanan AI
Diproduksi oleh Kantor Federal Jerman untuk Keamanan Informasi (BSI), dokumen ini memberikan pengenalan tentang kemungkinan serangan terhadap sistem pembelajaran mesin dan potensi pertahanan terhadap serangan tersebut.

Prinsip-Prinsip Panduan Internasional Proses Hiroshima untuk Organisasi yang Mengembangkan Sistem AI Tingkat Lanjut dan Kode Etik Internasional Proses Hiroshima untuk Organisasi yang Mengembangkan Sistem AI Tingkat Lanjut
Dokumen-dokumen ini, yang diproduksi sebagai bagian dari Proses AI Hiroshima G7, memberikan panduan bagi organisasi yang mengembangkan sistem AI tercanggih, termasuk model dasar tercanggih dan sistem AI generatif dengan tujuan mempromosikan AI yang aman, terjamin, dan tepercaya di seluruh dunia.

AI Verify
Kerangka Pengujian Tata Kelola AI dan perangkat Perangkat Lunak Singapura yang memvalidasi kinerja sistem AI berdasarkan serangkaian prinsip yang diakui secara internasional melalui pengujian standar.

Kerangka Kerja Multilapis untuk Praktik Keamanan Siber yang Baik untuk AI — ENISA (europa.eu)
Kerangka kerja untuk memandu Otoritas Kompeten Nasional dan pemangku kepentingan AI mengenai langkah-langkah yang harus diikuti untuk mengamankan sistem, operasi, dan proses AI mereka.

ISO 5338: Proses siklus hidup sistem AI (Dalam peninjauan)
Seperangkat proses dan konsep terkait untuk menggambarkan siklus hidup sistem AI berdasarkan pembelajaran mesin dan sistem heuristik.

Katalog Kriteria Kepatuhan Layanan AI Cloud (AIC4)
Katalog Kriteria Kepatuhan Layanan AI Cloud BSI menyediakan kriteria khusus AI, yang memungkinkan evaluasi keamanan layanan AI di seluruh siklus hidupnya.

NIST IR 8269 (Draf) Taksonomi dan Terminologi Pembelajaran Mesin Adversarial
Seperangkat proses dan konsep terkait untuk menggambarkan siklus hidup sistem AI berdasarkan pembelajaran mesin dan sistem heuristik.

MITRE ATLAS
Basis pengetahuan tentang taktik, teknik, dan studi kasus melawan musuh untuk sistem pembelajaran mesin (ML), yang dimodelkan dan ditautkan ke kerangka kerja MITRE ATT&CK.

Ikhtisar Risiko Bencana AI (2023)
Diproduksi oleh Center for AI Safety, dokumen ini menguraikan area risiko yang ditimbulkan oleh AI.

Model Bahasa Besar: Peluang dan Risiko bagi Industri dan Otoritas
Dokumen yang dibuat oleh BSI untuk perusahaan, otoritas, dan pengembang yang ingin mempelajari lebih lanjut tentang peluang dan risiko pengembangan, penerapan, dan/atau penggunaan LLM.

Proyek sumber terbuka untuk membantu pengguna menguji keamanan model AI meliputi:

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

Keamanan siber

Sasaran Kinerja Keamanan Siber CISA
Seperangkat perlindungan umum yang harus diterapkan oleh semua entitas infrastruktur penting untuk mengurangi kemungkinan dan dampak risiko yang diketahui serta teknik musuh secara signifikan.

NCSC CAF Framework
Cyber Assessment Framework (CAF) memberikan panduan bagi organisasi yang bertanggung jawab atas layanan dan aktivitas yang sangat penting.

Kerangka Kerja Keamanan Rantai Pasokan MITRE
Kerangka kerja untuk mengevaluasi pemasok dan penyedia layanan dalam rantai pasokan.

Manajemen risiko

Kerangka Kerja Manajemen Risiko AI NIST (AI RMF)
AI RMF menguraikan cara mengelola risiko sosio-teknis terhadap individu, organisasi, dan masyarakat yang secara unik terkait dengan AI.

ISO 27001: Keamanan informasi, keamanan siber, dan perlindungan privasi
Standar ini memberikan panduan kepada organisasi mengenai pembuatan, penerapan, dan pemeliharaan sistem manajemen keamanan informasi.

ISO 31000: Manajemen risiko
Standar internasional yang memberikan panduan dan prinsip manajemen risiko dalam organisasi kepada organisasi.

Panduan Manajemen Risiko NCSC
Panduan ini membantu praktisi risiko keamanan siber untuk lebih memahami dan mengelola risiko keamanan siber yang memengaruhi organisasi mereka.

Catatan

1. Di sini didefinisikan sebagai seseorang, otoritas publik, agen atau badan lain yang mengembangkan sistem AI (atau yang mengembangkan sistem AI) dan menjual sistem tersebut di pasar atau menggunakannya di bawah nama atau merek dagangnya sendiri
2. Untuk informasi lebih lanjut tentang keamanan berdasarkan desain, lihat halaman web CISA Secure by Design halaman web dan panduan Mengubah Keseimbangan Risiko Keamanan Siber: Prinsip dan Pendekatan untuk Perangkat Lunak Secure by Design
3. Berlawanan dengan pendekatan AI non-ML seperti sistem berbasis aturan
4. CEPS menjelaskan tujuh jenis interaksi pengembangan AI dalam publikasinya 'Merekonsiliasi Rantai Nilai AI dengan Undang-Undang Kecerdasan Buatan EU'
5. ISO/IEC 22989:2022(en) mendefinisikan ini sebagai 'elemen fungsional yang membangun sistem AI'
6. NIST bertugas membuat panduan (dan mengambil tindakan lain) untuk memajukan pengembangan dan penggunaan Kecerdasan Buatan (AI) yang aman, terjamin, dan dapat dipercaya. Lihat Tanggung Jawab NIST Berdasarkan Perintah Eksekutif 30 Oktober 2023
7. Informasi selengkapnya tentang pemodelan ancaman tersedia dari OWASP Foundation
8. Lihat MITRE ATLAS Adversarial Machine Learning 101
9. GitHub: RCE PoC untuk Tensorflow menggunakan lapisan Lambda yang berbahaya
10. SLSA: 'Menjaga integritas artefak di semua rantai pasokan perangkat lunak'
11. METI (Kementerian Ekonomi, Perdagangan dan Industri Jepang, 2023), 'Panduan Pengenalan Bill of Materials Perangkat Lunak (SBOM) untuk Manajemen Perangkat Lunak'
12. Penelitian Google: Pembelajaran Mesin: Kartu Kredit Berbunga Tinggi untuk Utang Teknis
13. Tramèr dkk 2016, Mencuri Model Pembelajaran Mesin melalui API Prediksi
14. Boenisch, 2020, Serangan terhadap Machine Learning Privacy (Bagian 1): Model Serangan Inversi dengan IBM-ART Framework
15. National Cyber Security Centre, 2020, Rancang dan bangun Infrastruktur Kunci Publik yang dihosting secara pribadi