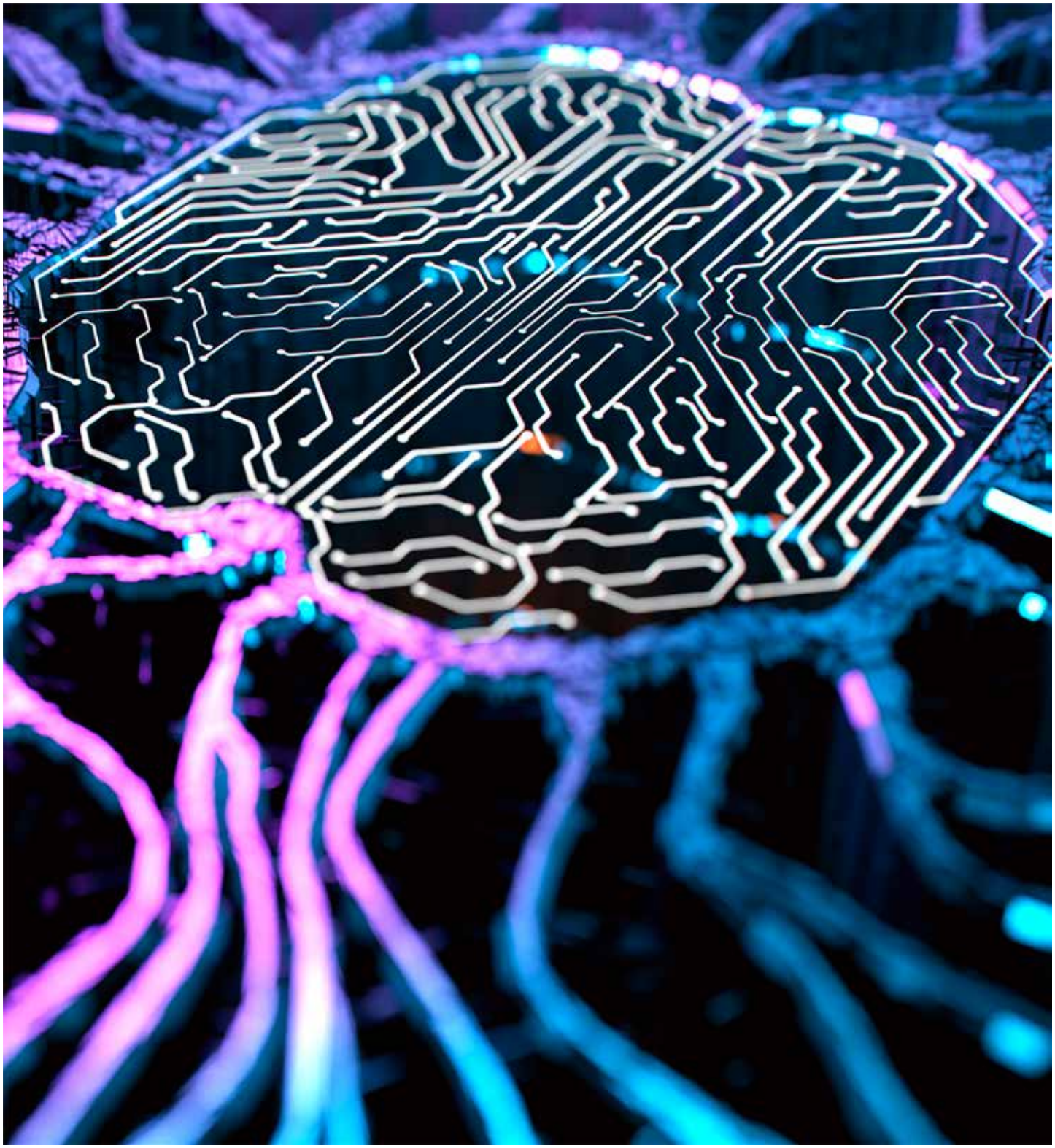
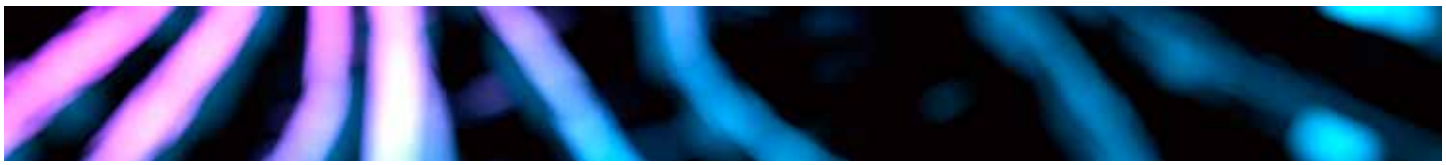
Garis panduan bagi pembangunan sistem AI teguh





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

Mengenai dokumen ini

Dokumen ini diterbitkan oleh Pusat Keselamatan Siber UK (UK National Cyber Security Centre) (NCSC), Agensi Keselamatan Siber dan Keselamatan Infrastruktur AS (US Cybersecurity and Infrastructure Security Agency) (CISA), dan rakan-rakan kongsi antarabangsa berikut:

- Agensi Keselamatan Kebangsaan (National Security Agency) (NSA)
- Biro Penyiasatan Persekutuan (Federal Bureau of Investigations) (FBI)
- Pusat Keselamatan Siber Direktorat Semboyan Australia (Australian Signals Directorate's Australian Cyber Security Centre) (ACSC)
- Pusat Keselamatan Siber Kanada (Canadian Centre for Cyber Security) (CCCS)
- Pusat Keselamatan Siber Kebangsaan New Zealand (New Zealand National Cyber Security Centre) (NCSC-NZ)
- CSIRT Kerajaan Chile
- Agensi Keselamatan Maklumat dan Siber Kebangsaan Czechia (Czechia's National Cyber and Information Security Agency) (NUKIB)
- Pihak Berkuasa Sistem Maklumat Estonia (Information System Authority of Estonia) (RIA) dan Pusat Keselamatan Siber Kebangsaan Estonia (National Cyber Security Centre of Estonia) (NCSC-EE)
- Agensi Keselamatan Siber Perancis (French Cybersecurity Agency) (ANSSI)
- Pejabat Persekutuan Jerman untuk Keselamatan Maklumat (Germany's Federal Office for Information Security) (BSI)
- Direktorat Siber Kebangsaan Israel (Israeli National Cyber Directorate) (INCD)
- Agensi Keselamatan Siber Kebangsaan Itali (Italian National Cybersecurity Agency) (ACN)
- Pusat Kebangsaan Kesiapsiagaan dan Strategi Insiden untuk Keselamatan Siber Jepun (Japan's National center of Incident readiness and Strategy for Cybersecurity) (NISC)
- Urusetia Dasar Sains, Teknologi dan Inovasi, Pejabat Jemaah Kabinet Jepun (Japan's Secretariat of Science, Technology and Innovation Policy, Cabinet Office)
- Agensi Pembangunan Teknologi Maklumat Kebangsaan Nigeria (Nigeria's National Information Technology Development Agency) (NITDA)
- Pusat Keselamatan Siber Kebangsaan Norway (Norwegian National Cyber Security Centre) (NCSC-NO)
- Kementerian Hal-Ehwal Digital Poland (Poland Ministry of Digital Affairs)
- Institut Penyelidikan Kebangsaan NASK Poland (Poland's NASK National Research Institute) (NASK)
- Perkhidmatan Perisikan Kebangsaan Republik Korea (Republic of Korea National Intelligence Service) (NIS)
- Agensi Keselamatan Siber Singapura (Cyber Security Agency of Singapore) (CSA)

Sekalung Budi

Organisasi-organisasi berikut telah menyumbang kepada pembangunan garis panduan-garis panduan ini:

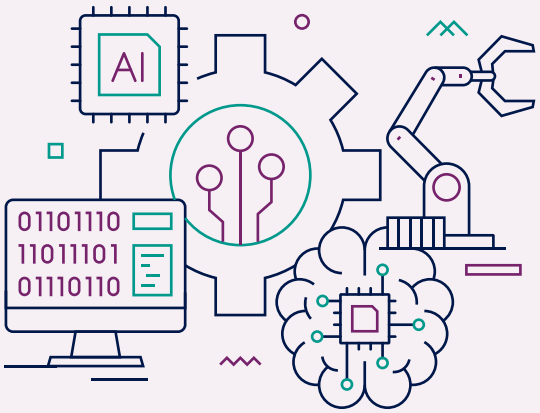
- Alan Turing Institute
- Anthropic
- Databricks
- Georgetown University's Center for Security and Emerging Technology
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute at Carnegie Mellon University
- Stanford Center for AI Safety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

Penafian

Maklumat di dalam dokumen ini disediakan 'seadanya' oleh NCSC dan pihak organisasi pengarang yang tidak akan bertanggungjawab bagi sebarang kerugian, kecederaan atau kerosakan dalam apa jua bentuk yang disebabkan oleh penggunaannya kecuali apa yang dikehendaki di bawah undang-undang. Maklumat dalam dokumen ini tidak mewakili atau mencadangkan pengendorsan atau saranan oleh mana-mana organisasi, produk, atau perkhidmatan pihak ketiga oleh NCSC dan agensi-agensi pengarang. Pautan dan rujukan kepada laman-laman web dan bahan-bahan pihak ketiga telah disediakan bagi tujuan maklumat sahaja dan tidak mewakili pengendorsan atau saranan sumber-sumber sedemikian daripada yang lain.

Dokumen ini disediakan atas dasar TLP:CLEAR (https://www.first.org/tlp/).

Isi Kandungan



Ringkasan eksekutif

Dokumen ini menyarankan garis-garis panduan kepada penyedia sebarang sistem yang menggunakan kecerdasan buatan (AI), sama ada sistem yang telah dicipta sepenuhnya atau yang dibina atas alat-alat dan perkhidmatan-perkhidmatan yang disediakan oleh pihak lain. Pelaksanaan garis-garis panduan ini akan membantu penyedia untuk membina sistem-sistem AI yang berfungsi seperti yang diingini, tersedia bila diperlukan, dan berfungsi tanpa mendedahkan data sensitif kepada pihak-pihak yang tidak dibenarkan.

Dokumen ini disasarkan terutamanya kepada penyedia sistem-sistem AI yang menggunakan model-model yang dihoskan oleh sesebuah organisasi, atau sedang menggunakan aplikasi pengantaramuka pengaturcaraan luaran (external application programming interfaces) (APIs). Kami menggesa **semua** pemegang taruh (termasuk pakar sains data, pembangun, pengurus, pembuat-keputusan dan pemilik risiko) untuk membaca garis-garis panduan ini untuk membantu mereka membuat keputusan berpengetahuan mengenai **rekabentuk**, **pembangunan**, **pengaturgerakan** dan **pengoperasian** sistem-sistem AI mereka.

Mengenai garis-garis panduan ini

Sistem-sistem AI berpotensi untuk membawa banyak manfaat kepada masyarakat. Walaubagaimanapun begitu, untuk merealisasikan peluang-peluang AI dengan sepenuhnya, ia mesti dibangun, diaturgerak dan diselenggarakan dalam satu cara yang teguh dan bertanggungjawab.

Sistem-sistem AI adalah tertakluk kepada keterdedahan-keterdedahan keselamatan terbaharu yang perlu dipertimbangkan selari dengan ancaman-ancaman keselamatan siber yang standard. Bila tahap kelajuan pembangunan ialah tinggi – seperti dalam kes dengan AI – keselamatan boleh sering menjadi pertimbangan sampingan. Keselamatan mesti menjadi satu keperluan teras, bukan sahaja dalam fasa pembangunan, tetapi di sepanjang kitaran hidup sistem itu.

Bagi tujuan ini, garis-garis panduan berkenaan dibahagikan kepada empat bidang utama di dalam kitaran hidup pembangunan sistem AI: **rekabentuk teguh**, **pembangunan teguh**, **pengaturgerakan teguh**, dan **pengoperasian dan penyelenggaraan teguh**. Untuk setiap bahagian, kami akan cadangkan pertimbangan dan pemitigasian yang akan membantu mengurangkan keseluruhan risiko kepada proses pembangunan sistem AI bagi sesebuah organisasi.

1. **Rekabentuk teguh**
   Bahagian ini mengandungi garis-garis panduan yang terpakai bagi tahap perekabentukan kitaran hidup pembangunan sistem AI. Ia meliputi pemahaman mengenai risiko dan pemodelan ancaman, selain topik-topik khusus dan tukar-ganti untuk dipertimbangkan mengenai rekabentuk sistem dan model.

2. **Pembangunan teguh**
   Bahagian ini mengandungi garis-garis panduan yang terpakai bagi tahap perekabentukan kitaran hidup pembangunan sistem AI, termasuk keselamatan rantaian bekalan, pendokumentasian, serta pengurusan hutang aset dan teknikal.

3. **Pengaturgerakan teguh**
   Bahagian ini mengandungi garis-garis panduan yang terpakai bagi tahap perekabentukan kitaran hidup pembangunan sistem AI, termasuk melindungi infrastruktur dan model-model daripada pengkompromian, ancaman atau kehilangan, membangunkan proses-proses pengurusan insiden, dan pengeluaran bertanggungjawab.

4. **Pengoperasian dan penyelenggaraan teguh**
   Bahagian ini mengandungi garis-garis panduan yang terpakai bagi tahap pengoperasian dan penyelenggaraan teguh bagi kitaran hidup pembangunan sistem AI. Ia menyampaikan garis-garis panduan mengenai tindakan-tindakan yang relevan khususnya bila sesebuah sistem telah diaturgerakkan, termasuk pengelogan dan pemantauan, pengurusan pengemaskinian dan perkongsian maklumat.

Garis-garis panduan ini menurut sebuah pendekatan 'teguh secara tereka', dan diselaraskan secara rapat kepada amalan-amalan yang didefinisikan dalam Panduan pembangunan dan pengaturgerakanNCSC, Kerangka Kerja Pembangunan Perisian Teguh NIST, dan 'prinsip-prinsip teguh secara tereka' yang diterbitkan oleh CISA, NCSC dan agensi-agensi siber antarabangsa. Mereka mengutamakan:

- Mengambil milik hasil keputusan keselamatan bagi pelanggan
- Mencakupi ketelusan dan kebertanggungjawaban radikal
- Membina struktur organisasi dan kepemimpinan agar teguh secara tereka merupakan sebuah keutamaan tinggi perniagaan

Pengenalan

Sistem-sistem kecerdasan buatan (AI) berpotensi untuk membawa banyak manfaat kepada masyarakat. Walaubagaimanapun begitu, untuk merealisasikan peluang-peluang AI dengan sepenuhnya, ia mesti dibangun, diaturgerak dan diselenggarakan dalam satu cara yang teguh dan bertanggungjawab. Keselamatan siber ialah satu pra-syarat yang diperlukan bagi keselamatan, daya tahan, privasi, keadilan, keberkesanan dan kebolehpercayaan sistem-sistem AI.

Namun begitu, sistem-sistem AI adalah tertakluk kepada keterdedahan-keterdedahan keselamatan terbaharu yang perlu dipertimbangkan selari dengan ancaman-ancaman keselamatan siber yang standard. Bila tahap kelajuan pembangunan ialah tinggi – seperti dalam kes dengan AI – keselamatan boleh sering menjadi pertimbangan sampingan. Keselamatan mesti menjadi satu keperluan teras, bukan sahaja dalam fasa pembangunan, tetapi di sepanjang kitaran hidup sistem itu.

Dokumen ini menyarankan garis-garis panduan kepada penyedia[1] sebarang sistem yang menggunakan kecerdasan buatan (AI), sama adasistem yang telah dicipta sepenuhnya atau yang dibina atas alat-alat dan perkhidmatan-perkhidmatan yang disediakan oleh pihak lain. Pelaksanaan garis-garis panduan ini akan membantu penyedia untuk membina sistem-sistem AI yang berfungsi seperti yang diingini, tersedia bila diperlukan, dan berfungsi tanpa mendedahkan data sensitif kepada pihak-pihak yang tidak dibenarkan.

Garis-garis panduan ini patut dipertimbangkan bersama dengan keselamatan siber, pengurusan risiko, dan amalan terbaik respons insiden yang sudah ditetapkan. Khususnya, kami menggesa para penyedia untuk menurut prinsip-prinsip 'teguh secara tereka'[2] yang dibangunkan oleh Agensi Keselamatan Siber dan Keselamatan Infrastruktur AS (principles developed by the US Cybersecurity and Infrastructure Security Agency) (CISA), Pusat Keselamatan Siber Kebangsaan UK (UK National Cyber Security Centre) (NCSC), dan semua rakan kongsi antarabangsa kami. Prinsip-prinsip ini mengutamakan:

- Mengambil milik hasil keputusan keselamatan bagi pelanggan
- Mencakupi ketelusan dan kebertanggungjawaban radikal
- Membina struktur organisasi dan kepemimpinan agar teguh secara tereka merupakan sebuah keutamaan tinggi perniagaan

Menurut prinsip-prinsip 'teguh secara tereka' memerlukan sumber-sumber signifikan sepanjang kitaran hidup sesebuah sistem. Ia bererti para pembangun mesti melabur dalam mengutamakan **ciri-ciri**, **mekanisme-mekanisme**, dan **pelaksanaan** peralatan yang melindungi pelanggan di setiap lapisan rekabentuk sistem tersebut, dan di serata tahap kitaran hidup pembangunannya. Berbuat demikian akan menghalang rekabentuk kembali yang menelan belanja nanti, selain menjamin pengawalan pelanggan serta data mereka dalam jangkamasa terdekat.

Mengapa keselamatan AI berbeza?

Dalam dokumen ini kami menggunakan 'AI' untuk merujuk secara khusus kepada aplikasi-aplikasi pembelajaran mesin (machine learning) (ML)[3]. Semua jenis ML terkandung dalam skop ini. Kami mendefinisikan aplikasi ML sebagai aplikasi yang:

- Melibatkan komponen perisian (model-model) yang membenarkan komputer untuk mengenali dan membawa konteks kepada corak-corak dalam data tanpa peraturan-peraturan yang diaturcarakan secara nyata oleh seorang manusia.
- Menjanakan ramalan, saranan-saranan, atau keputusan-keputusan berdasarkan penaakulan secara statistik.

Selain ancaman keselamatan siber sedia ada, sistem-sistem AI adalah tertakluk kepada jenis-jenis keterdedahan baharu. Frasa 'pembelajaran mesin secara berlawanan' (adversarial machine learning) (AML), adalah diguna untuk menerangkan pengeksploitasian keterdedahan asas dalam komponen-komponen ML, termasuk perkakasan, perisian, aliran kerja dan rantaian-rantaian bekalan. AML membenarkan penyerang untuk menyebabkan tingkahlaku yang tidak diingini dalam sistem-sistem ML yang boleh termasuk:

- Menjejaskan pengklasifikasian model itu atau prestasi pemunduran.
- Membenarkan pengguna untuk melakukan tindakan yang tidak dibenarkan.
- Mengekstrak maklumat model sensitif.

Terdapat pelbagai cara untuk mencapai kesan-kesan ini, misalnya sebagai serangan suntikan pantas dalam domain model bahasa yang besar (large language model) (LLM), atau dengan sengaja mencemarkan data latihan atau maklumbalas pengguna (yang dikenali sebagai 'peracunan data').

Siapa yang patut baca dokumen ini?

Dokumen ini disasarkan terutamanya kepada penyedia sistem-sistem AI yang menggunakan model-model yang dihoskan oleh sesebuah organisasi, atau sedang menggunakan aplikasi pengantaramuka pengaturcaraan luaran (external application programming interfaces) (APIs). Walaubagaimanapun begitu, kami menggesa **semua** pemegang taruh (termasuk pakar sains data, pembangun, pengurus, pembuat keputusan dan pemilik risiko) untuk membaca garis-garis panduan ini untuk membantu mereka membuat keputusan berpengetahuan mengenai **rekabentuk**, **pengaturgerakan** dan **pengoperasian** sistem-sistem pembelajaran mesin AI mereka.

Biarpun begitu, tidak semua garis-garis panduan ini akan ada kaitan secara langsung kepada semua organisasi. Tahap kecanggihan dan kaedah-kaedah serangan akan berbeza bergantung kepada pihak lawan yang menyasarkan sistem AI itu, jadi garis-garis panduan ini patut dipertimbangkan sejajar dengan kes-kes penggunaan dan profil ancaman organisasi anda.

Siapa yang bertanggungjawab bagi pembangunan AI yang teguh?

Seringkali terdapat ramai pelaku dalam rantaian-rantaian bekalan AI moden. Satu pendekatan yang ringkas adalah untuk membahagikan mereka kepada dua entiti:

- Pihak 'penyedia' yang bertanggungjawab ke atas kurasi data, pembangunan algoritma, rekabentuk, pengaturgerakan dan penyelenggaraan
- Pihak 'pengguna' yang menyampaikan input-input dan menerima output-output

Walaupun pendekatan penyedia-pengguna ini diguna dalam banyak aplikasi, ia semakin jarang digunakan [4], kerana penyedia mungkin akan menimbangkan kemasukan perisian, data, model-model dan/atau perkhidmatan jarak jauh yang disediakan oleh pihak ketiga ke dalam sistem-sistem mereka sendiri. Rantaian-rantaian bekalan kompleks ini akan menjadikannya lebih sukar bagi pengguna akhir untuk memahami di manakah kebertanggungjawaban AI yang teguh patut ditanggung.

Para pengguna (sama ada 'pengguna akhir', atau penyedia-penyedia yang sedang memasukkan sebuah komponen AI luaran[5]) lazimnya tidak mempunyai keterlihatan dan/atau kepakaran yang mencukupi untuk memahami, menilai atau menangani dengan sepenuhnya risiko-risiko yang dikaitkan dengan sistem-sistem yang mereka sedang gunakan. Oleh yang demikian, selaras dengan prinsip-prinsip 'teguh secara tereka', **para penyedia komponen-komponen AI patut bertanggungjawab ke atas hasil-hasil keputusan keselamatan pengguna yang berada jauh di hujung rantaian bekalan.**

Para penyedia patut melaksanakan kawalan-kawalan keselamatan dan pemitigasian di mana yang boleh di dalam model-model, saluran-saluran pain dan/atau sistem-sistem mereka, dan di mana pengesetan diguna, laksanakan pilihan yang paling teguh sebagai pilihan lalainya. Di mana risiko tidak boleh dimitigasi, penyedia patut bertanggungjawab bagi:

- Memaklumkan pengguna yang berada jauh di hujung rantaian bekalan mengenai risiko yang sedang diterima oleh mereka dan (jika berkenaan) pengguna mereka sendiri
- Menasihati mereka tentang cara bagaimana untuk menggunakan komponen itu secara teguh

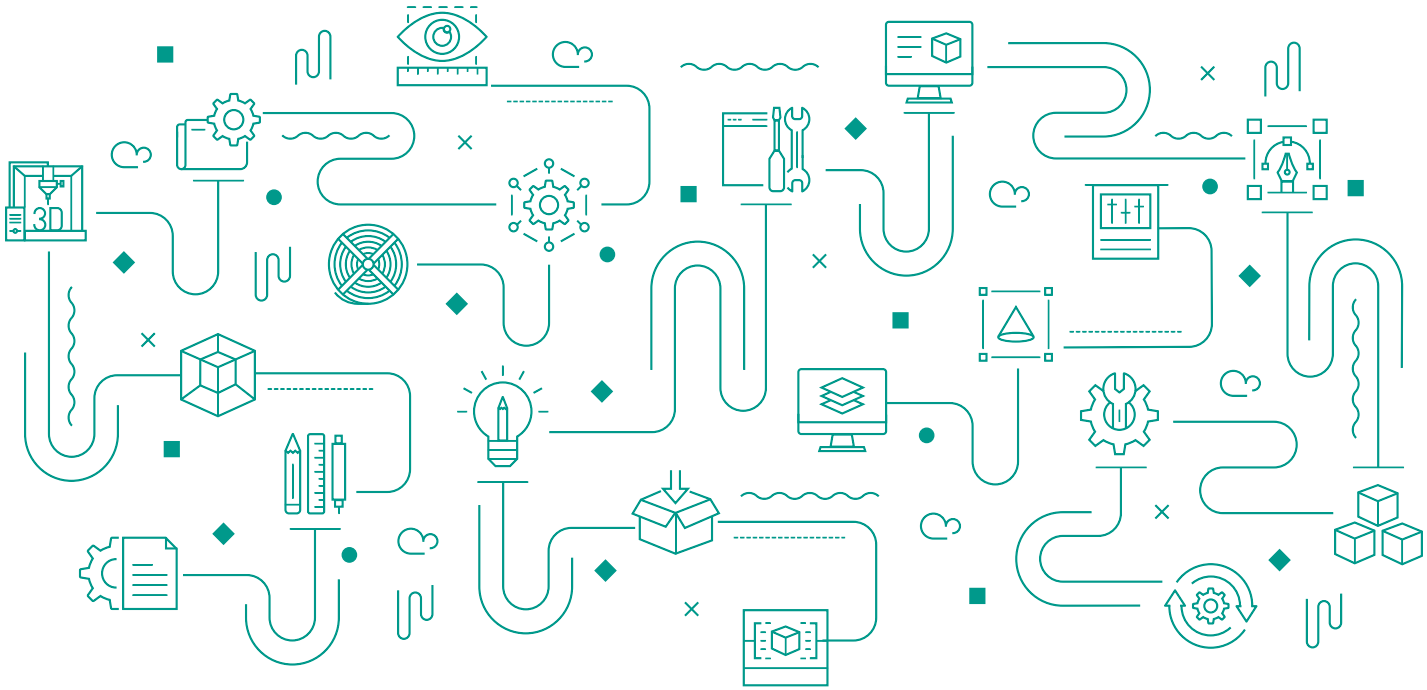
Dalam keadaan di mana pengkompromian sistem yang boleh membawa kepada kerosakan fizikal atau kepada reputasi, kehilangan signifikan pengoperasian perniagaan, kebocoran maklumat sensitif atau sulit yang ketara atau meluas dan/atau implikasi perundangan, risiko keselamatan siber AI patut ditangani sebagai **kritikal.**

Garis panduan bagi pembangunan sistem AI yang teguh

Garis-garis panduan ini dibahagikan kepada empat bidang utama di dalam kitaran hidup pembangunan sistem AI: **rekabentuk teguh**, **pembangunan teguh**, **pengaturgerakan teguh**, dan **pengoperasian dan penyelenggaraan teguh**. Untuk setiap bidang, kami akan mencadangkan pertimbangan dan pemitigasian yang akan membantu mengurangkan keseluruhan risiko kepada proses pembangunan sistem AI organisasi tersebut.

Garis-garis panduan yang dinyatakan di dalam dokumen ini adalah berkait rapat dengan amalan-amalan kitaran hidup pembangunan perisian yang didefinisikan di dalam:

- Panduan pembangunan dan pengaturgerakan teguh NCSC
- Kerangka Kerja Pembangunan Perisian Teguh (Secure Software Development Framework) (SSDF)[6] Institut Kebangsaan Kepiawaian dan Teknologi) (National Institute of Standards and Technology) (NIST)



1. Rekabentuk Teguh

Bahagian ini mengandungi garis panduan yang terpakai ke atas tahap **perekabentukan** bagi kitaran hidup pembangunan sistem AI. Ia meliputi memahami pemodelan risiko dan ancaman, selain topik-topik khusus dan tukar ganti ("trade offs") untuk mempertimbangkan rekaan sistem dan model.

Tingkatkan kesedaran kakitangan mengenai ancaman dan risiko

Pemilik sistem dan pemimpin kanan memahami ancaman untuk menawan AI dan pemitigasiannya. Pakar sains dan pembangun data anda mengekalkan sebuah kesedaran tentang ancaman keselamatan dan mod-mod kegagalan yang relevan dan bantu pemilik risiko untuk membuat keputusan yang berpengetahuan. Anda menyediakan pengguna dengan panduan mengenai risiko keselamatan unik yang dihadapi sistem-sistem AI (contohnya, sebagai sebahagian daripada latihan InfoSec standard) dan latih pembangun dalam teknik-teknik pengekodan teguh dan amalan-amalan AI yang teguh dan bertanggungjawab.

Modelkan ancaman terhadap sistem anda

Sebagai sebahagian daripada proses pengurusan risiko anda, anda menerapkan sebuah proses menyeluruh untuk mengakses ancaman-ancaman kepada sistem anda, yang termasuk memahami kesan-kesan yang berpotensi berlaku kepada sistem itu, pengguna, organisasi-organisasi, dan masyarakat secara umumnya jika sesuatu komponen AI dikompromi atau berkelakuan secara tidak dijangka[7]. Proses ini melibatkan penilaian kesan ancaman khusus-AI [8] dan mendokumentasikan pengambilan keputusan anda.

Anda mengakui bahawa kesensitivian dan jenis-jenis data yang diguna dalam sistem anda mungkin mempengaruhi nilainya sebagai sebuah sasaran bagi seorang penyerang. Penilaian anda patut mempertimbangkan bahawa sesetengah ancaman mungkin akan berkembang apabila sistem-sistem AI semakin dilihat sebagai sasaran bernilai tinggai, dan apabila AI itu sendiri membolehkan vektor-vektor serangan automatik baharu.

Rekabentukkan sistem anda untuk keselamatan serta kefungsian dan prestasi.

Anda yakin bahawa cara yang paling baik untuk menangani tugasan yang sedang dilakukan ialah dengan menggunakan AI. Setelah menentukan perkara ini, anda menilai kesesuaian pilihan-pilihan rekaan khusus-AI anda. Anda pertimbangkan model ancaman anda dan pemitigasian keselamatan yang berkaitan selari dengan kefungsian, pengalaman pengguna, persekitaran pengaturgerakan, prestasi, pengesyoran, pengawasan, keperluan etika dan perundangan, selain pertimbangan-pertimbangan lain. Contohnya:

- Anda pertimbangkan keselamatan rantaian bekalan bila memilih sama ada untuk membangun secara dalaman atau gunakan komponen-komponen luaran, contohnya:
    - Pilihan anda sama ada untuk melatih sebuah model baharu, gunakan sebuah model sedia ada (dengan atau tanpa talaan halus) atau mengakses sebuah model melalui sebuah API luaran yang sesuai dengan keperluan anda
    - Pilihan anda sama ada untuk bekerja dengan suatu penyedia model luaran termasuk sebuah penilaian usaha wajar terhadap postur keselamatan penyedia itu sendiri
    - Jika anda menggunakan sebuah perpustakaan luaran, anda lengkapkan sebuah penilaian usaha wajar (contohnya, untuk memastikan perpustakaan itu memiliki kawalan-kawalan yang menghalang sistem itu daripada memuatnaik model-model yang tidak diyakini tanpa mendedahkan diri mereka secara langsung kepada pelaksanaan kod arbitrari[9])
    - Anda laksanakan pengimbasan dan pemencilan/"sandboxing" sewaktu mengimpot model-model pihak ketiga atau berat-berat bersiri, yang patut ditangani sebagai kod pihak ketiga yang tidak diyakini dan mungkin berupaya membolehkan pelaksanaan kod kawalan jauh

- Jika anda menggunakan API luaran, anda perlu terapkan kawalan sewajarnya ke atas data yang boleh dihantar ke perkhidmatan-perkhidmatan di luar kawalan organisasi anda, misalnya dengan memerlukan pengguna untuk daftar masuk dan mengesahkannya sebelum menghantar maklumat yang berpotensi sensitif
- Anda terapkan pemeriksaan dan pesanitasian data dan input sewajarnya; ini termasuk sewaktu memasukkan maklumbalas pengguna atau data pembelajaran berterusan ke dalam modal anda, dengan mengiktiraf bahawa data latihan mendefinisikan tingkahlaku sistem

- Anda integrasikan pembangunan sistem perisian AI ke dalam amalan-amalan terbaik pembangunan dan pengoperasian teguh sedia ada; semua elemen sistem AI telah ditulis dalam persekitaran yang wajar dengan menggunakan amalan-amalan dan bahasa-bahasa pengodan yang mengurangkan atau menyingkirkan kelas-kelas keterdedahan yang dikenali di mana ia berpatutan
- Jika komponen-komponen AI perlu mencetus tindakan, contohnya mengubah fail-fail atau mengarahkan output ke sistem-sistem luaran, anda terapkan larangan-larangan yang wajar terhadap tindakan-tindakan yang mungkin berlaku (ini termasuk AI luaran dan jaminan daripada kegagalan bukan-AI jika perlu)
- Keputusan-keputusan yang berlegar atas soal interaksi pengguna dimaklumkan oleh risiko khusus-AI, contohnya:
  - Sistem anda menyampaikan pengguna dengan output yang boleh diguna tanpa mendedahkan tahap pembutiran yang tidak perlu kepada bakal penyerang
  - Jika perlu, sistem anda menyampaikan susur kawalan berkesan di sekitar output model-model
  - Jika menawarkan sesuatu AI kepada pelanggan atau pengusahasama luaran, and terapkan kawalan wajar yang akan memitigasikan serangan terhadap sistem AI melalui API
  - Anda integrasikan pengesetan yang paling teguh ke dalam sistem itu dengan lalai
  - Anda menerapkan prinsip-prinsip hak keistimewaan yang paling kurang untuk mengehadkan akses kepada pengefungsian sesebuah sistem
  - Anda menjelaskan keupayaan-keupayaan yang lebih berisiko kepada pengguna dan memerlukan penggunaan untuk memilih untuk memasuki dan menggunakannya; anda memberitahu mengenai kes-kes penggunaan yang dilarang, dan di mana berkenaan, maklumkan kepada pengguna mengenai penyelesaian alternatif

Pertimbangkan manfaat keselamatan dan tukar-ganti sewaktu memilih model AI anda

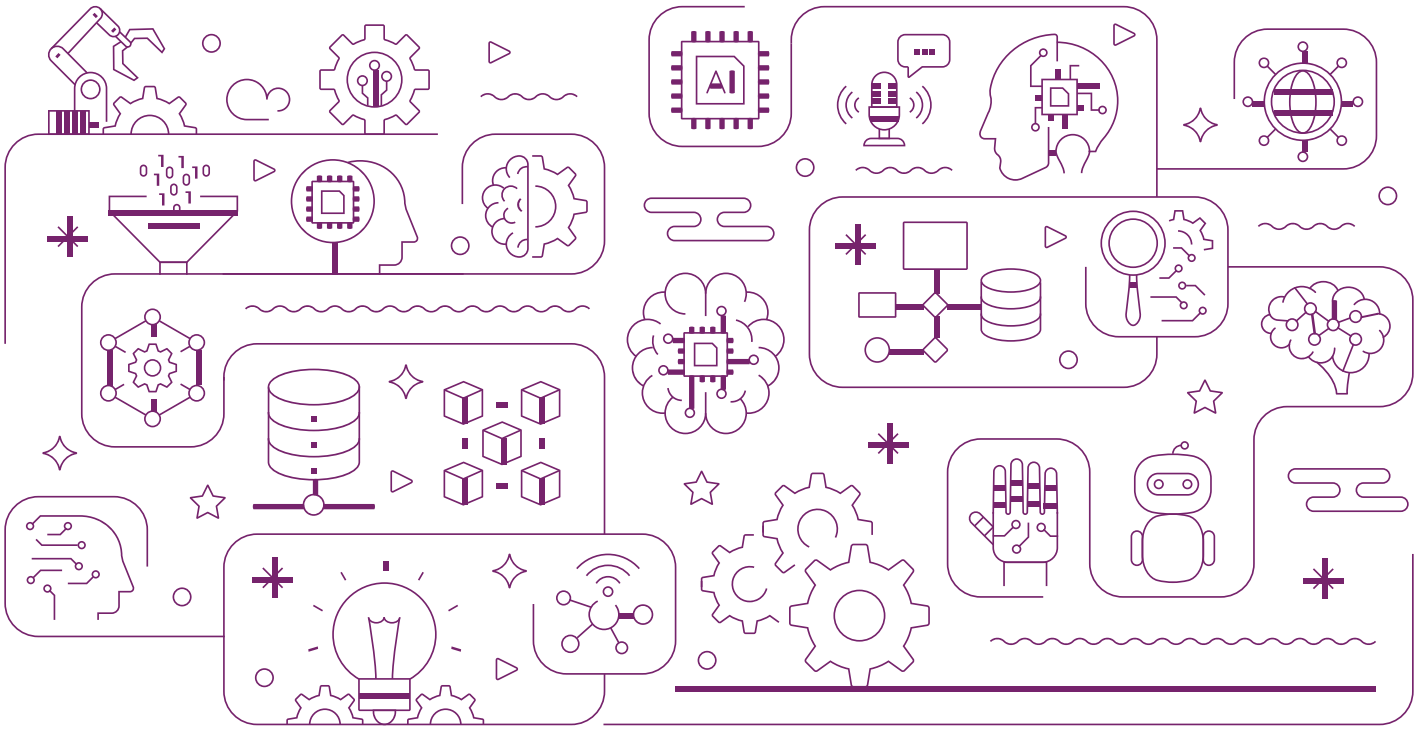
Pilihan model AI anda akan melibatkan pengimbangan pelbagai keperluan. Ini termasuk pilihan arkitektur model, pengkonfigurasian, data latihan, algoritma latihan dan parameterhiper. Keputusan anda dimaklumkan oleh model ancaman anda, dan dinilai kembali secara berkala selaras dengan kemajuan penyelidikan keselamatan AI dan evolusi pemahaman tentang ancaman berkenaan.

Bila memilih sebuah model AI, pertimbangan anda mungkin termasuk, tetapi tidak terhad kepada:

- Kerumitan model yang anda gunakan, iaitu, arkitektur yang dipilih dan bilangan parameter; arkitektur pilihan dan bilangan parameter model anda akan, antara faktor-faktor lain, membawa kesan ke atas berapa banyak data latihan yang ia perlukan dan bagaimana daya ketahanannya terhadap perubahan di dalam data input apabila ia sedang diguna
- Kewajaran model tersebut kepada kes penggunaan anda dan/atau kebolehlaksanaan untuk menyesuaikannya kepada keperluan khusus anda (contohnya dengan talaan halus)
- keupayaan untuk menyelari, menafsir dan menjelaskan output-output model anda (contohnya untuk menyah pepijat, pematuhan audit atau pengawalseliaan); barangkali terdapat manfaat untuk menggunakan model-model yang ringkas, dan lebih telus, daripada model-model yang besar dan rumit yang lebih sukar untuk ditafsir
- Sifat-sifat dataset(-dataset) latihan, termasuk saiz, integriti, kualiti, kesensitivian, umur, kerelevanan dan kepelbagaian

- nilai penggunaan pengerasan model (misalnya latihan berlawanan), pembiasaan dan/atau teknik-teknik peningkatan-privasi
- provenans dan rantaian bekalan komponen-komponen termasuk model atau model asas, data latihan dan alat-alat berkaitan

Untuk maklumat lanjut mengenai berapa banyak daripada faktor ini membawa kesan ke atas hasil keputusan keselamatan, sila rujuk kepada 'Prinsip-prinsip bagi Keselamatan Pembelajaran Mesin' ('Principles for the Security of Machine Learning') NCSC, khususnya Rekabentuk bagi keselamatan (arkitektur model).



2. Pembangunan teguh

Bahagian ini mengandungi garis-garis panduan yang terpakai ke atas tahap **pembangunan** kitaran hidup pembangunan sistem AI, termasuk keselamatan rantaian bekalan, pendokumentasian, serta pengurusan aset dan hutang teknikal.

Teguhkan rantaian bekalan anda

Anda mengakses dan memantau keselamatan bekalan rantaian AI anda di serata kitaran hidup sesuatu sistem, dan memerlukan pembekal untuk mematuhi tahap-tahap kepiawaian yang sama dengan apa yang organisasi anda terapkan ke atas perisian lain. Jika pembekal tidak boleh mematuhi tahap-tahap kepiawaian organisasi anda, anda bertindak secara berpatutan dengan dasar-dasar pengurusan risiko sedia ada anda.

Bila ia tidak dikeluarkan secara dalaman, anda perolehi dan selenggarakan komponen-komponen perkakasan dan perisian yang diteguhkan dengan baik dan didokumentasikan dengan baik (contohnya, model-model, data, perpustakaan-perpustakaan perisian, modul-modul, perkakasan pertengahan (middleware), kerangka-kerangka kerja, dan API-API luaran) daripada pembangun komersial, sumber terbuka dan pihak ketiga lain yang sudah ditentusahkan untuk memastikan keselamatan yang teguh dan kuat di dalam sistem-sistem anda.

Anda bersedia untuk berulang gagal kepada penyelesaian-penyelesaian lain untuk sistem-sistem misi-kritikal lain, jika kriteria keselamatan tidak dipenuhi. Anda gunakan sumber-sumber seperti Panduan Rantaian Bekalan NCSC dan kerangka-kerangka kerja seperti Tahap-Tahap Rantaian Bekalan bagi Artifak-Artifak Perisian (Supply Chain Levels for Software Artifacts) (SLSA)[10] untuk menjejaki pengakusaksian rantaian bekalan dan kitaran hidup pembangunan perisian.

Kenalpasti, jejaki dan lindungi aset-aset anda

Anda memahami nilai aset-aset berkaitan-AI anda kepada organisasi anda termasuk model-model, data (termasuk maklumbalas pengguna), pembantu ingat, perisian, pendokumentasian, log-log dan penilaian (termasuk maklumat mengenai keupayaan tidak selamat dan mod-mod kegagalan yang berpotensi berlaku), dan mengakui di mana mereka mewakili pelaburan signifikan dan di mana akses kepada mereka akan membolehkan seseorang penyerang. Anda menangani log sebagai data sensitif dan melaksanakan kawalan-kawalan untuk melindungi kerahsiaan, integriti dan adanya mereka.

Anda tahu di mana aset-aset anda berada dan telah menilai dan menerima apa-apa risiko berkaitan. Anda mempunyai proses-proses dan alat-alat untuk menjejaki, mengesahkan ketulenan, mengawal versi dan meneguhkan aset-aset anda, dan boleh kembali kepada satu keadaan baik yang diketahui jika sebuah kejadian pengkompromian berlaku.

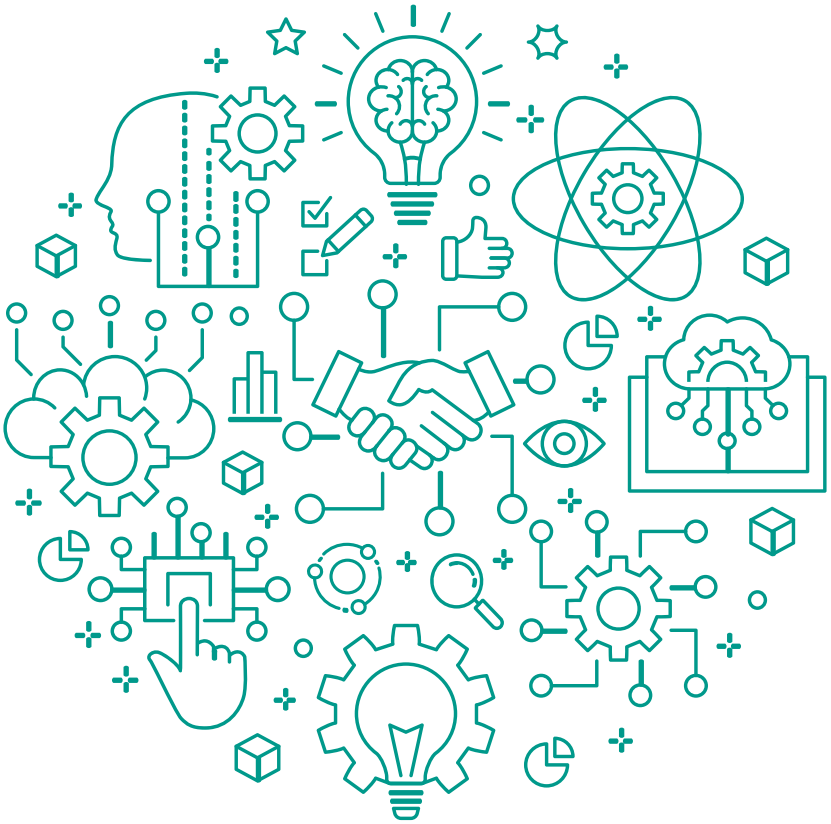
Anda mempunyai proses-proses dan kawalan-kawalan yang sudah siap ditempatkan untuk mengurus data apa yang boleh diakses sistem-sistem AI, dan untuk menguruskan kandungan yang dijana AI menurut kesensitiviannya (dan kesensitivian input-input yang disertakan untuk menjanakannya).

Dokumenkan data, model, dan pembantu ingat anda

Anda mendokumentasikan penciptaan, pengoperasian, dan pengurusan kitaran hidup mana-mana model, set data dan pembantu ingat meta atau sistem. Pendokumentasian anda termasuk maklumat relevan-keselamatan seperti sumber-sumber data latihan (termasuk talaan halus data dan maklum balas manusia atau pengoperasian yang lain), skop dan had-had yang diingini, susur kawalan, tanda pagar kriptografik atau tandatangan, masa penyimpanan, kekerapan penyemakan kembali yang dicadangkan, dan mod-mod kegagalan yang berpotensi berlaku. Struktur-struktur berguna yang akan membantu untuk berbuat demikian termasuklah kad-kad model, kad-kad data dan bil-bil material-material perisian (software bills of materials) (SBOMs). Pengeluaran pendokumentasian komprehensif menyokong ketelusan dan ketanggungjawaban[11].

Tangani hutang teknikal anda

Seperti juga dengan mana-mana sistem perisian lain, anda akan mengenal pasti, menjejaki dan mengurus 'hutang teknikal' anda di sepanjang kitaran hidup sesebuah sistem AI (hutang teknikal ialah di mana keputusan kejuruteraan yang tidak memenuhi amalan-amalan terbaik untuk mencapai hasil keputusan jangkamasa-pendek dibuat, dengan mengabaikan manfaat jangkmasa-panjang). Seperti hutang kewangan, hutang teknikal bukanlah semestinya tidak baik, tetapi patut diurus daripada tahap-tahap terawal pembangunannya[12]. Anda mengakui bahawa berbuat demikian mungkin akan menjadi lebih mencabar di dalam sebuah konteks AI daripada perisian standard, dan bahawa tahap-tahap hutang teknikal anda kemungkinan besarnya tinggi akibat kitaran-kitaran pembangunan yang pantas dam satu kekurangan protokol dan pengantaramuka yang sudah ditetapkan dengan baik. Anda memastikan rancangan kitaran hidup anda (termasuk proses-proses untuk menyahkomisenkan sistem-sistem AI) mengakses, mengambil maklum dan memitigasikan risiko bagi sistem-sistem serupa di masa hadapan.



3. Pengaturgerakan teguh

Bahagian ini mengandungi garis panduan yang terpakai ke atas tahap **pengaturgerakan** kitaran hidup pembangunan sistem AI, termasuk melindungi infrastruktur dan model daripada pengkompromian, ancaman atau kehilangan, membangunkan proses-proses pengurusan insiden, dan keluaran bertanggungjawab.

Teguhkan infrastruktur anda

Anda terapkan prinsip-prinsip keselamatan infrastruktur baik kepada infrastruktur yang diguna di dalam setiap bahagian kitaran hidup sistem anda. Anda terapkan kawalan akses yang sesuai kepada API-API, model-model dan data-data anda, dan kepada saluran latihan dan pemerosesan mereka, dalam penyelidikan dan pembangunan serta pengaturgerakan mereka juga. Ini termasuk pengasingan yang sesuai bagi persekitaran yang mengandungi kod atau data sensitif. Ini juga akan membantu memitigasikan serangan keselamatan siber standard yang bermatlamat untuk mencuri sesebuah model atau mendatangkan mudarat kepada prestasinya.

Lindungi model anda secara berterusan

Para penyerang mungkin boleh membina kembali kefungsian sesebuah model[13] atau data yang dilatihkan olehnya[14], dengan mengakses sesebuah model secara langsung (dengan memperolehi berat-berat model) atau secara tidak langsung (dengan menyoal model tersebut melalui sebuah aplikasi atau perkhidmatan). Penyerang juga mungkin akan mengacau model-model, data atau pembantu ingat semasa atau selepas latihan, menjadikan output itu tidak boleh diyakini.

Anda melindungi model dan data daripada akses langsung dan tidak langsung, masing-masing, dengan:

- Melaksanakan amalan-amalan terbaik keselamatan siber standard
- Melaksanakan kawalan-kawalan ke atas pengantaramuka pertanyaan untuk mengesan dan menghalang percubaan untuk mengakses, mengolah, dan meresap keluar maklumat sulit

Untuk memastikan bahawa sistem-sistem penggunaan boleh mensahihkan model-model, anda komput dan kongsikan hash kriptografik dan/atau tandatangan fail-fail model (contohnya, berat model) dan set-set data (termasuk titik pemeriksaan)sebaik sahaja model itu dilatih. Seperti kebiasaannya dengan kriptografi, pengurusan kunci yang baik adalah diperlukan[15].

Pendekatan anda kepada pemitigasian risiko kerahsiaan akan bergantung secara besar kepada kes penggunaan dan model ancaman. Sesetengah aplikasi, contohnya yang melibatkan data yang sungguh sensitif, mungkin memerlukan jaminan bersifat teori yang mungkin akan menjadi sukar atau mahal untuk diterapkan. Jika ia sesuai, teknologi-teknologi yang meningkatkan privasi (misalnya kebezaan privasi atau penyulitan homomorfik) boleh diguna untuk menerokai atau menjamin tahap-tahap risiko yang dikaitkan dengan para konsumer, pengguna dan penyerang memiliki akses kepada model-model dan output-output.

Bangunkan prosedur-prosedur pengurusan insiden

Penjejasan sistem-sistem AI anda oleh insiden-insiden keselamatan yang tidak dapat dielakkan adalah dicerminkan dalam rancangan-rancangan respons insiden, pendadakan dan pemulihan anda. Rancangan-rancangan anda mencerminkan senario-senario yang berbeza dan kerap dinilai kembali selaras dengan evolusi sistem dan penyelidikan yang meluas. Anda menyimpan sumber-sumber digital kritikal syarikat dalam rancangan sandar luar talian. Para perespons telah dilatih untuk menilai dan menangani insiden-insiden berkaitan-AI. Anda menyediakan log-log audit kualiti-tinggi dan ciri-ciri keselamatan atau maklumat lain kepada pelanggan dan pengguna tanpa dikenakan caj tambahan, untuk membolehkan proses-proses respons insiden mereka.

Keluarkan AI secara bertanggungjawab

Anda keluarkan model-model, aplikasi-aplikasi atau sistem-sistem hanya setelah mereka melalui penilaian keselamatan yang wajar dan berkesan misalnya penanda arasan dan pasukan merah (red teaming) (selain ujian-ujian lain yang terletak di luar skop garis panduan-garis panduan ini, misalnya keselamatan atau keadilan), dan anda menjelaskan kepada pengguna anda mengenai pengehadan atau mod-mod kegagalan yang berpotensi berlaku yang diketahui. Butir-butir perpustakaan ujian keselamatan sumber-terbuka diberi di dalam bahagian bacaan lanjut pada penghujung dokumen ini.

Jadikan ia lebih mudah bagi pengguna untuk melakukan perkara yang betul

Anda mengiktiraf bahawa setiap pengesetan baharu atau pilihan pengkonfigurasian akan dinilai sejajar dengan manfaat perniagaan yang diperolehinya, dan sebarang risiko keselamatan yang ia kenalkan. Sebaiknya, pengesetan paling teguh akan diintegrasikan ke dalam sistem itu sebagai satu-satunya pilihan yang ada sahaja. Bila pengkonfigurasian diperlukan, pilihan lalai patut diteguhkan secara melebar terhadap ancaman-ancaman lazim (iaitu, teguh secara lalai). Anda terapkan kawalan-kawalan untuk menghalang penggunaan atau pengaturgerakan sistem anda dalam cara-cara yang berniat jahat.

Anda sediakan pengguna dengan panduan mengenai penggunaan sewajarnya terhadap model atau sistem anda, termasuk menggariskan had-had dan mod-mod kegagalan yang berpotensi berlaku. Anda menyatakan dengan jelas kepada pengguna aspek-aspek keselamatan mana yang menjadi tanggungjawab mereka, dan bersikap telus mengenai di mana (dan bagaimana) data mereka mungkin diguna, diakses atau disimpan (contohnya, jika ia diguna bagi latihan kembali mod, atau disemak kembali oleh para kakitangan atau rakan kongsi).

4. Pengoperasian dan penyelenggaraan teguh

Bahagian ini mengandungi garis panduan yang terpakai ke atas tahap **pengoperasian dan penyelenggaraan teguh** kitaran hidup pembangunan sistem AI. Ia menyampaikan garis panduan mengenai tindakan-tindakan yang khususnya relevant apabila sesebuah sistemtelah diaturgerakkan, termasuk pengelogan dan pemantauan, pengurusan pengemaskinian dan perkongsian maklumat.

Pantau tingkahlaku sistem anda

Ukurkan output dan prestasi model dan sistem anda dalam cara di mana anda boleh memerhatikan perubahan yang mendadak dan bertahap dalam tingkahlaku yang menjejaskan keselamatan. Anda boleh mengambil kira dan mengenal pasti potensi pencerobohan dan pengkompromian, selain hanyutan data semulajadi.

Pantau input-input sistem anda

Sejajar dengan keperluan privasi dan perlindungan data, anda memantau dan melog input-input ke dalam sistem anda (misalnya permintaan inferens, pertanyaan-pertanyaan atau pembantu ingat) untuk membolehkan kewajipan-kewajipan pematuhan, audit, penyiasatan dan pemulihan sekiranya berlaku pengkompromian atau penyalahgunaan. Ini mungkin termasuk pengesanan nyata kehabisan-pengedaran dan/atau input berlawanan, termasuk yang menyasar untuk mengeksploitasi langkah-langkah persediaan data (misalnya "cropping" dan mengubah saiz bagi imej-imej).

Turuti sebuah pendekatan teguh secara tereka kepada pengemaskinian

Anda masukkan pengemaskinian automatik secara lalai di dalam setiap produk dan gunakan prosedur-prosedur pengemaskinian teguh, dan modular untuk mengedarkannya. Proses-proses pengemaskinian anda (termasuk rejim-rejim ujian dan penilaian) mencerminkan fakta bahawa perubahan kepada data, model-model atau pembantu ingat boleh membawa kepada perubahan dalam kelakuan sistem (contohnya, anda tangani pengemaskinian utama seperti versi baharu). Anda menyokong pengguna untuk menilai dan merespons kepada perubahan model (contohnya dengan menyediakan akses pra-tonton dan API-API berversi).

Kumpul dan kongsikan teladan yang dipelajari

Anda mengambil bahagian dalam komuniti-komuniti perkongsian maklumat, berusaha sama di serata eko-sistem global industri ini, akademia dan kerajaan untuk berkongsi amalan terbaik dengan sewajarnya. Anda mengekalkan saluran-saluran komunikasiterbuka untuk maklum balas mengenai keselamatan sistem, baik secara dalaman atau luaran kepada organisasi anda, termasuk menyampaikan kebenaran kepada penyelidik keselamatan untuk menyelidik dan melaporkan keterdedahan. Bila diperlukan, anda mendadakkan isu-isu kepada komuniti yang lebih meluas, contohnya dengan menerbitkan buletin-buletin yang merespons kepada pendedahan keterdedahan, termasuk pembilangan keterdedahan biasa secara terperinci dan lengkap. Anda mengambil tindakan untuk memitigasi dan memulihkan isu-isu dengan secepat dan sewajar mungkin.

Bacaan Lanjut

Pembangunan AI

Prinsip-prinsip keselamatan bagi pembelajaran mesin
Panduan terperinci NCSC mengenai pembangunan, pengaturgerakan atau pengoperasian sesebuah sistem dengan sebuah komponen ML.

Teguh secara Tereka – Mengalih Keseimbangan Risiko Keselamatan Siber: Prinsip-prinsip dan Pendekatan-pendekatan bagi Perisian Teguh secara Tereka
Dikarang bersama oleh CISA, NCSC dan agensi-agensi lain, panduan ini menerangkan cara bagaimana pengeluar sistem-sistem perisian, termasuk AI, patut mengambil langkah untuk memfaktorkan keselamatan ke dalam peringkat perekabentukan pembangunan produk, dan mengirim produk-produk yang datang siap teguh bila ia dikeluarkan dari kotaknya.

Kebimbangan Keselamatan AI secara Ringkas
Dihasilkan oleh Pejabat Persekutuan bagi Keselamatan Maklumat Jerman (BSI), dokumen ini menyampaikan sebuah pengenalan kepada serangan-serangan yang mungkin berlaku ke atas sistem-sistem pembelajaran mesin dan pertahanan yang berpotensi diadakan terhadap serangan-serangan tersebut.

Prinsip-prinsip Panduan Antarabangsa Proses Hiroshima bagi Organisasi-Organisasi yang Membangunkan Sistem-Sistem AI Maju dan Kod Tatalaku Antarabangsa Proses Hiroshima bagi Organisasi-Organisasi yang Membangunkan Sistem-Sistem AI Maju
Dokumen-dokumen ini, dihasilkan sebagai sebahagian daripada Proses AI Hiroshima G7, menyampaikan panduan bagi organisasi-organisasi yang membangunkan sistem-sistem AI yang paling maju, termasuk model-model asas yang paling maju dan sistem-sistem AI penjana dengan matlamat untuk mempromosikan keselamatan, teguh, dan AI boleh diyakini di seluruh dunia.

Penentusahan AI
Kerangka Kerja Ujian Tadbir Urus AI Singapura dan kit peralatan Perisian yang menentusahkan prestasi Sistem-sistem AI terhadap satu set prinsip-prinsip yang diiktiraf secara antarabangsa melalui ujian-ujian yang diseragamkan.

Kerangka Kerja Lapisan Berganda untuk Amalan-Amalan Keselamatan Siber Yang Baik untuk AI (Multilayer Framework for Good Cybersecurity Practices for AI) - ENISA (europa.eu)
Satu kerangka kerja untuk memandu Pihak-Pihak Berkuasa Cekap Kebangsaan dan pemegang-pemegang taruh AI tentang langkah-langkah yang mereka perlu ikuti untuk meneguhkan sistem-sistem, operasi-operasi dan proses-proses AI mereka.

ISO 5338: Proses-proses kitaran hidup sistem AI (Di bawah semakan kembali)
Satu set proses-proses dan konsep-konsep berkaitan bagi menerangkan kitaran hidup sistem-sistem AI berdasarkan pembelajaran mesin dan sistem-sistem heuristik.

Katalog Kriteria Pematuhan Perkhidmatan Awan AI (AI Cloud Service Compliance Criteria Catalogue) (AIC4)
Katalog Kriteria Pematuhan Perkhidmatan Awan AI BSI (BSI's AI Cloud Service Compliance Criteria Catalogue) menyampaikan kriteria AI-khusus, yang membolehkan penilaian keselamatan sesebuah perkhidmatan AI di sepanjang kitaran hidupnya.

NIST IR 8269 (Draf) Satu Taksonomi dan Terminologi Pembelajaran Mesin Berlawanan (Draft) (A Taxonomy and Terminology of Adversarial Machine Learning)
Satu set proses-proses dan konsep-konsep berkaitan bagi menerangkan kitaran hidup sistem-sistem AI berdasarkan pembelajaran mesin dan sistem-sistem heuristik.

MITRE ATLAS
Sebuah asas pengetahuan mengenai taktik-taktik pihak lawan, teknik-teknik, dan kajian-kajian kes untuk sistem-sistem pembelajaran mesin (ML), yang dimodelkan menurut dan berkaitan dengan kerangka kerja MITRE ATT&CK.

Sebuah Pengenalan Umum kepada Bencana Risiko AI (An Overview of Catastrophic AI Risks) (2023)
Dihasilkan oleh Pusat Keselamatan AI, dokumen ini menggariskan bidang-bidang risiko yang dikemukakan oleh AI.

Model-Model Bahasa Besar: Peluang-Peluang dan Risiko-Risiko bagi Industri dan Pihak-Pihak Berkuasa
Dokumen yang dikeluarkan oleh BSI untuk syarikat-syarikat, pihak-pihak berkuasa dan para pembangun yang mahu belajar lanjut mengenai peluang-peluang dan risiko-risiko membangun, mengaturgerak dan/atau menggunakan LLM.

Projek-projek sumber terbuka yang membantu pengguna menguji keselamatan model-model AI termasuk:

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

Keselamatan siber

Matlamat-matlamat Prestasi Keselamatan Siber CISA
Sebuah set umum perlindungan-perlindungan yang semua entiti infrastruktur kritikal patut laksanakan untuk mengurangkan secara kemungkinan dan kesan teknik-teknik risiko dan pihak lawan dengan bermakna.

Kerangka Kerja NCSC CAF
Kerangka Kerja Penilaian Siber (Cyber Assessment Framework) (CAF) menyediakan panduan kepada organisasi-organisasi yang bertanggungjawab bagi perkhidmatan-perkhidmatan dan kegiatan-kegiatan yang penting lagi mustahak.

Kerangka Kerja Keselamatan Rantaian Bekalan MITRE
Sebuah kerangka kerja untuk menilai para pembekal dan penyedia perkhidmatan di dalam rantaian bekalan.

Pengurusan risiko

Kerangka Kerja Pengurusan Risiko AI IST (IST AI Risk Management Framework) (AI RMF)
AI RMF menggariskan cara bagaimana untuk mengurus risiko-risiko sosio-teknikal kepada individu, organisasi, dan masyarakat yang dikaitkan secara unik dengan AI.

ISO 27001: Maklumat keselamatan, keselamatan siber dan perlindungan privasi
Standard ini menyampaikan panduan kepada organisasi-organisasi mengenai pengewujudan, pelaksanaan dan penyelenggaraan sebarang sistem pengurusan keselamatan maklumat.

ISO 31000: Pengurusan risiko
Sebuah tahap kepiawaian antarabangsa yang menyampaikan garis-garis panduan dan prinsip-prinsip kepada organisasi-organisasi bagi pengurusan risiko di dalam organisasi-organisasi.

Panduan Pengurusan Risiko NCSC
Panduan ini membantu pengamal risiko keselamatan siber untuk memahami dan menangani risko-risiko keselamatan yang menjejaskan organisasi-organisasi mereka dengan lebih baik.

Nota

1. Didefinisikan di sini sebagai seorang individu, pihak berkuasa awam, agensi atau mana-mana badan lain yang membangunkan sebuah sistem AI (atau yang telah membangunkan sebuah sistem AI) dan meletakkan sistem itu dalam pasaran atau menjalankannya di bawah nama atau jenamanya sendiri
2. Untuk maklumat lanjut mengenai teguh secara tereka, sila lihat laman web Teguh secara Tereka dan panduan Mengalih Keseimbangan Risiko Keselamatan Siber CISA: Prinsip-prinsip dan Pendekatan-pendekatan bagi Perisian Teguh secara Tereka
3. Iaitu yang bertentangan dengan pendekatan-pendekatan bukan-ML AI seperti sistem-sistem berdasarkan-peraturan
4. CEPS menerangkan tujuh jenis interaksi pembangunan AI berbeza di dalam penerbitan mereka 'Menyelaraskan Rantaian Nilai AI dengan Akta Kecerdasan Buatan EU' ('Reconciling the AI Value Chain with the EU's Artificial Intelligence Act')
5. ISO/IEC 22989:2022(en) mendefinisikan perkara ini sebagai 'sebuah elemen berfungsi yang membina sebuah sistem AI'
6. NIST diberi tugas untuk menghasilkan garis-garis panduan (dan mengambil tindakan lain) untuk memajukan pembangunan dan penggunaan Kecerdasan Buatan (AI) yang selamat, teguh dan boleh diyakini. Sila lihat Tanggungjawab NIST Di Bawah Arahan Eksekutif 30 Oktober, 2023
7. Maklumat lanjut mengenai pemodelan ancaman disediakan daripada Yayasan OWASP (OWASP Foundation)
8. Sila lihat MITRE ATLAS Pembelajaran Mesin Berlawanan 101 (Adversarial Machine Learning 101)
9. GitHub: PoC RCE untuk Tensorflow menggunakan sebuah lapisan Lambda yang berniat jahat (RCE PoC for Tensorflow using a malicious Lambda layer)
10. SLSA: 'Mengawalselamat integriti artifak di serata mana-mana rantaian bekalan perisian'
11. METI (Japanese Ministry of Economy, Trade and Industry, 2023), 'Guide of Introduction of Software Bill of Materials (SBOM) for Software Management'
12. Penyelidikan Google: Pembelajaran Mesin: Kad Kredit Bunga Tinggi Hutang Teknikal
13. Tramèr et al 2016, Mencuri model-model pembelajaran mesin melalui Ramalan API (Stealing Machine Learning Models via Prediction APIs)
14. Boenisch, 2020, Serangan terhadap Privasi Pembelajaran Mesin (Bahagian 1) (Attacks against Machine Learning Privacy (Part 1)): Serangan Penelangkupan Model dengan Kerangka Kerja IBM-ART (Model Inversion Attacks with the IBM-ART Framework)
15. Pusat Keselamatan Siber Kebangsaan (National Cyber Security Centre), 2020, Merekabentuk dan Membina sebuah Infrastruktur Utama Awam yang dihoskan secara peribadi