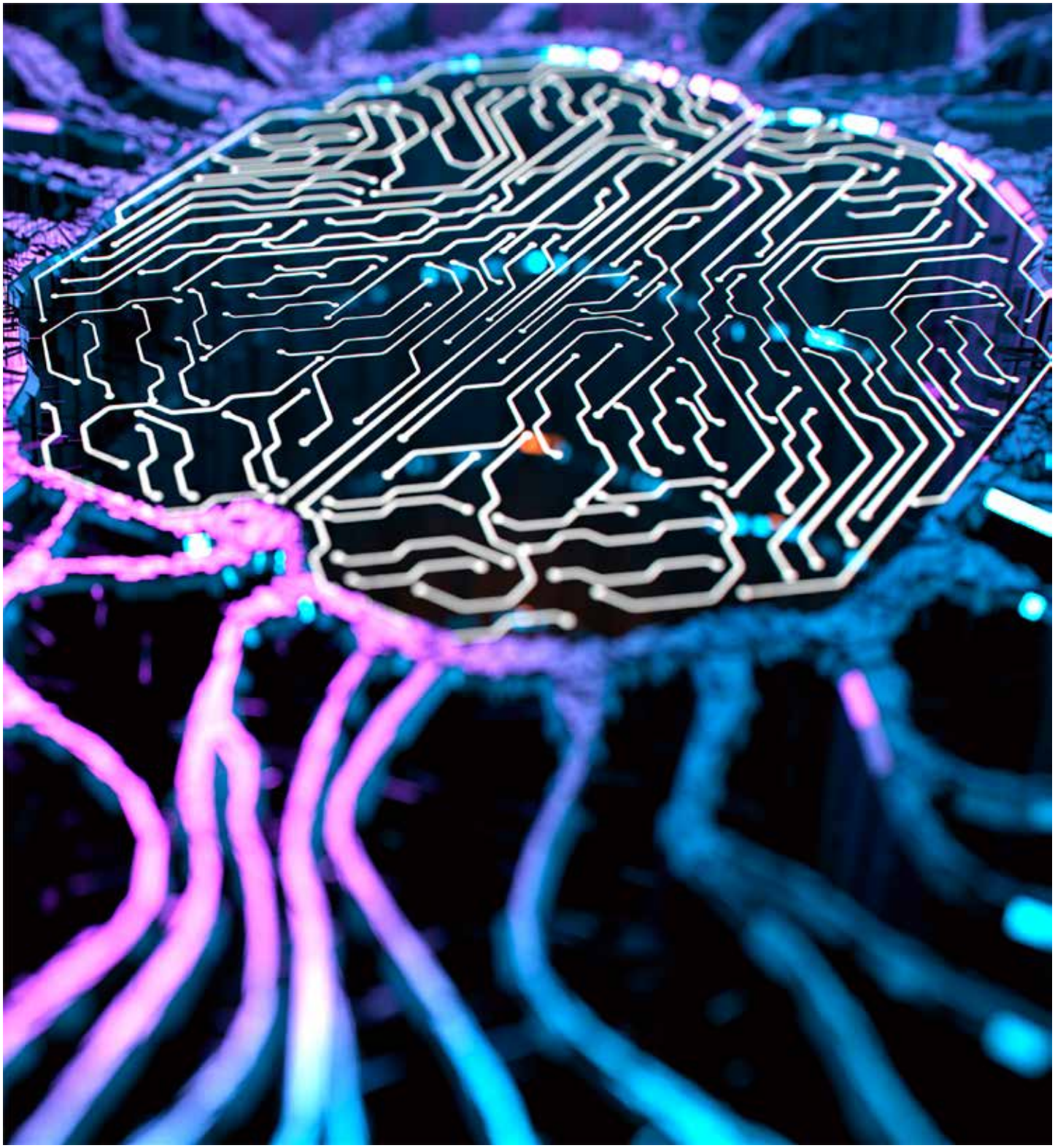
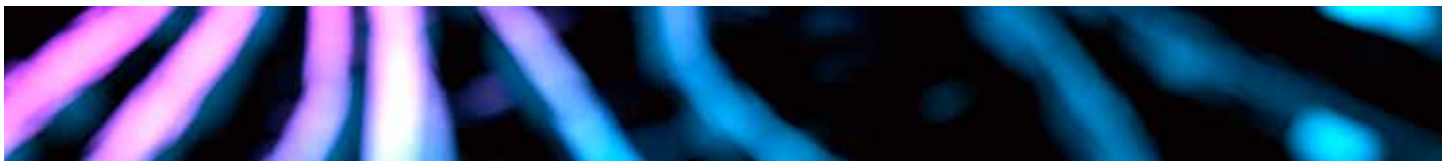
إرشادات لتطوير نظام الذكاء الاصطناعي الآمن





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

حول هذه الوثيقة

تم نشر هذه الوثيقة من قبل المركز الوطني للأمن السيبراني في المملكة المتحدة (NCSC)، والوكالة الأمريكية للأمن السيبراني وأمن البنية التحتية (CISA)، والشركاء الدوليين التاليين:

- وكالة الأمن القومي (NSA)
- مكتب التحقيقات الفيدرالي (FBI)
- مركز الأمن السيبراني الأسترالي التابع لمديرية الإشارات الأسترالية (ACSC)
- المركز الكندي للأمن السيبراني (CCCS)
- المركز الوطني النيوزيلندي للأمن السيبراني (NCSC-NZ)
- CSIRT التابع لحكومة شيلي
- الوكالة الوطنية لأمن المعلومات والإنترنت في التشيك (NUKIB)
- هيئة نظام المعلومات في إستونيا (RIA) والمركز الوطني للأمن السيبراني في إستونيا (NCSC-EE)
- الوكالة الفرنسية للأمن السيبراني (ANSSI)
- المكتب الاتحادي الألماني لأمن المعلومات (BSI)
- المديرية السيبرانية الوطنية الإسرائيلية (INCD)
- الوكالة الوطنية الإيطالية للأمن السيبراني (ACN)
- المركز الوطني الياباني للاستعداد للحوادث والاستراتيجية الخاصة بالأمن السيبراني (NISC)
- أمانة اليابان لسياسة العلوم والتكنولوجيا والابتكار، مكتب مجلس الوزراء
- الوكالة الوطنية لتطوير تكنولوجيا المعلومات في نيجيريا (NITDA)
- المركز الوطني النرويجي للأمن السيبراني (NCSC-NO)
- وزارة الشؤون الرقمية في بولندا
- معهد NASK الوطني للبحوث في بولندا (NASK)
- جهاز المخابرات الوطنية لجمهورية كوريا (NIS)
- وكالة الأمن السيبراني في سنغافورة (CSA)

شكر وتقدير

ساهمت المنظمات التالية في تطوير هذه المبادئ التوجيهية:

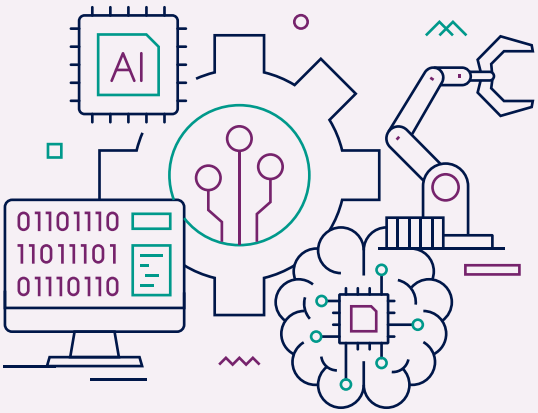
- معهد الآن تورينغ Alan Turing Institute
- أنثروبيك Anthropic
- داتابريكس Databricks
- مركز جامعة جورج تاون للأمن والتكنولوجيا الناشئة Georgetown University's Center for Security and Emerging Technology
- غوغل Google
- غوغل ديب ماند Google DeepMind
- آي بي إم IBM
- إمبيو ImBue
- مايكروسوفت Microsoft
- أوبن إي آي OpenAI
- بالانتير Palantir
- راند RAND
- سكيل إي آي Scale AI
- معهد هندسة البرمجيات في جامعة كارنيغي ميلون Software Engineering Institute at Carnegie Mellon University
- مركز ستانفورد لسلامة الذكاء الاصطناعي Stanford Center for AI Safety
- برنامج ستانفورد للجغرافيا السياسية والتكنولوجيا والحوكمة Stanford Program on Geopolitics, Technology and Governance

إبراء ذمة

يتم توفير المعلومات الواردة في هذه الوثيقة "كما هي" من قبل NCSC والمنظمات المؤلفة التي لن تكون مسؤولة عن أي خسارة أو إصابة أو ضرر من أي نوع ناتج عن استخدامها باستثناء ما يقتضيه القانون. المعلومات الواردة في هذا المستند لا تشكل أو تعني تأييدًا أو توصية لأي منظمة أو منتج أو خدمة تابعة لجهة خارجية من قبل NCSC ووكالات التأليف. يتم توفير الروابط والمراجع إلى مواقع الإنترنت ومواد الطرف الثالث للحصول على معلومات فقط ولا تمثل تأييدًا أو توصية لهذه الموارد على غيرها.

تم توفير هذا المستند على أساس TLP:CLEAR (https://www.first.org/tlp/).

المحتويات



موجز تنفيذي

توصي هذه الوثيقة بإرشادات لمقدمي أي أنظمة تستخدم الذكاء الاصطناعي (AI)، سواء تم إنشاء هذه الأنظمة من الصفر أو تم بناؤها بالاعتماد على الأدوات والخدمات المقدمة من قبل الآخرين. سيساعد تنفيذ هذه الإرشادات مقدمي الخدمة على بناء أنظمة الذكاء الاصطناعي التي تعمل على النحو المنشود، وتكون متاحة عند الحاجة، وتعمل دون الكشف عن البيانات الحساسة لأطراف غير مصرح لها.

تستهدف هذه الوثيقة في المقام الأول مقدمي أنظمة الذكاء الاصطناعي الذين يستخدمون النماذج التي تستضيفها إحدى المؤسسات، أو الذين يستخدمون واجهات برمجة التطبيقات الخارجية (APIs). نحن نحث **جميع** أصحاب المصلحة (بما في ذلك علماء البيانات والمطورين والمديرين وصناع القرار وأصحاب المخاطر) على قراءة هذه الإرشادات لمساعدتهم على اتخاذ قرارات مستنيرة بشأن **التصميم** و**التطوير**، و**النشر** و**التشغيل** لـ أنظمة الذكاء الاصطناعي الخاصة بهم.

حول المبادئ التوجيهية

تمتلك أنظمة الذكاء الاصطناعي القدرة على تحقيق العديد من الفوائد للمجتمع. ومع ذلك، لكي يتم تحقيق فرص الذكاء الاصطناعي بالكامل، يجب تطويرها ونشرها وتشغيلها بطريقة آمنة ومسؤولة.

تخضع أنظمة الذكاء الاصطناعي إلى ثغرات أمنية جديدة يجب أخذها في الاعتبار جنبًا إلى جنب مع التهديدات الأمنية السيبرانية العادية. عندما تكون وتيرة التطوير عالية – كما هو الحال مع الذكاء الاصطناعي – كثيرًا ما يكون الأمن اعتبارًا ثانويًا. يجب أن يكون الأمن متطلبًا أساسيًا، ليس فقط في مرحلة التطوير، ولكن طوال دورة حياة النظام.

ولهذا السبب، تم تقسيم الإرشادات إلى أربعة مجالات رئيسية ضمن دورة حياة تطوير نظام الذكاء الاصطناعي: **التصميم الآمن**، و**التطوير الآمن**، و**النشر الآمن**، **والتشغيل والصيانة الآمنة**. نقترح لكل قسم الاعتبارات والتخفيفات التي من شأنها أن تساعد في تقليل المخاطر الإجمالية لعملية تطوير نظام الذكاء الاصطناعي التنظيمي.

1. **التصميم الآمن**
   يحتوي هذا القسم على إرشادات تنطبق على مرحلة التصميم في دورة حياة تطوير نظام الذكاء الاصطناعي. يغطي فهم المخاطر ونمذجة التهديدات، بالإضافة إلى موضوعات محددة ومقايضات يجب أخذها في الاعتبار عند تصميم النظام والنموذج.

2. **التطوير الآمن**
   يحتوي هذا القسم على إرشادات تنطبق على مرحلة التطوير في دورة حياة تطوير نظام الذكاء الاصطناعي، بما في ذلك أمان سلسلة التوريد والتوثيق وإدارة الأصول والديون الفنية.

3. **النشر الآمن**
   يحتوي هذا القسم على إرشادات تنطبق على مرحلة النشر الخاصة بدورة حياة تطوير نظام الذكاء الاصطناعي، بما في ذلك حماية البنية التحتية والنماذج من الاختراق أو التهديد أو الخسارة، وتطوير عمليات إدارة الحوادث، والإطلاق المسؤول.

4. **التشغيل والصيانة الآمنة**
   يحتوي هذا القسم على إرشادات تنطبق على مرحلة التشغيل والصيانة الآمنة في دورة حياة تطوير نظام الذكاء الاصطناعي.
   يوفر هذا القسم إرشادات حول الإجراءات ذات الصلة بشكل خاص بمجرد نشر النظام، بما في ذلك التسجيل والمراقبة وإدارة التحديث ومشاركة المعلومات.

تتبع الإرشادات منهجًا "آمنًا افتراضيًا"، وتتوافق بشكل وثيق مع الممارسات المحددة في إرشادات التطوير والنشر الآمن الخاصة بـ NCSC، وإطار عمل تطوير البرامج الآمنة الخاص بـ NIST، و'مبادئ التصميم الآمن' التي تم نشرها بواسطة CISA وNCSC والوكالات السيبرانية الدولية. يعطون الأولوية:

- للحصول على ملكية النتائج الأمنية للعملاء
- لتبني الشفافية والمساءلة الجذرية
- لبناء الهيكل التنظيمي والقيادة الآمنين بحيث يكون التصميم الآمن من أهم أولويات العمل

مقدمة

تتمتع أنظمة الذكاء الاصطناعي (AI) بالقدرة على تحقيق العديد من الفوائد للمجتمع. ومع ذلك، لكي يتم تحقيق فرص الذكاء الاصطناعي بالكامل، يجب تطويره ونشره وتشغيله بطريقة آمنة ومسؤولة. يعد الأمن السيبراني شرطًا مسبقًا ضروريًا لسلامة ومرونة وخصوصية وعدالة وفعالية وموثوقية أنظمة الذكاء الاصطناعي.

ومع ذلك، تخضع أنظمة الذكاء الاصطناعي لثغرات أمنية جديدة يجب أخذها في الاعتبار إلى جانب التهديدات الأمنية السيبرانية العادية. عندما تكون وتيرة التطوير عالية – كما هو الحال مع الذكاء الاصطناعي – فإن الأمن كثيرًا ما يكون اعتبارًا ثانويًا. يجب أن يكون الأمن متطلبًا أساسيًا، ليس فقط في مرحلة التطوير، ولكن طوال دورة حياة النظام.

يوصي هذا المستند بإرشادات لموفري[1] أي أنظمة تستخدم الذكاء الاصطناعي، سواء تم إنشاء هذه الأنظمة من الصفر أو تم إنشاؤها بالاعتماد على الأدوات والخدمات المقدمة من قبل الآخرين. سيؤدي تنفيذ هذه الإرشادات إلى مساعدة مقدمي الخدمة على إنشاء أنظمة ذكاء اصطناعي تعمل على النحو المنشود، وتتوفر عند الحاجة، وتعمل دون الكشف عن بيانات حساسة لأطراف غير مصرح لها.

ينبغي النظر في هذه الإرشادات جنبًا إلى جنب مع أفضل ممارسات الأمن السيبراني وإدارة المخاطر والاستجابة للحوادث. وعلى وجه الخصوص، نحث مقدمي الخدمة على اتباع مبادئ "التصميم الآمن"[2] التي طورتها وكالة الأمن السيبراني وأمن البنية التحتية الأمريكية (CISA)، والمركز الوطني للأمن السيبراني في المملكة المتحدة (NCSC)، وجميع شركائنا الدوليين. تعطي المبادئ الأولوية:

- للحصول على ملكية النتائج الأمنية للعملاء
- لتبني الشفافية والمساءلة الجذرية
- لبناء الهيكل التنظيمي والقيادة الآمنين حسب التصميم من أهم أولويات العمل.

يتطلب اتباع مبادئ "التصميم الآمن" موارد كبيرة طوال دورة حياة النظام. يعني هذا أنه يجب على المطورين الاستثمار في وضع **ميزات**، و**آليات**، و**تنفيذ** الأدوات التي تحمي العملاء في كل طبقة من طبقات تصميم النظام، وعبر كافة مستويات مراحل دورة الحياة التطوير كأولوية. سيؤدي القيام بذلك إلى منع عمليات إعادة التصميم المكلفة لاحقًا، بالإضافة إلى حماية العملاء وبياناتهم على المدى القريب.

لماذا أمن الذكاء الاصطناعي مختلف؟

نستخدم في هذا المستند "AI" للإشارة تحديدًا إلى تطبيقات تعلم الآلة (ML)[3]. جميع أنواع تعلم الآلة ML موجودة في النطاق. نحن نعرّف تطبيقات ML بأنها التطبيقات التي:

- تتضمن مكونات برمجية (نماذج) تسمح لأجهزة الحاسوب بالتعرف على أنماط البيانات وإدخال سياق لها دون الحاجة إلى برمجة القواعد بشكل صريح بواسطة إنسان
- إنشاء تنبؤات أو توصيات أو قرارات بناءً على المنطق الإحصائي

بالإضافة إلى تهديدات الأمن السيبراني الحالية، تتعرض أنظمة الذكاء الاصطناعي لأنواع جديدة من نقاط الضعف. يُستخدم مصطلح "تعلم الآلة المضاد" (AML) لوصف استغلال نقاط الضعف الأساسية في مكونات تعلم الآلة، بما في ذلك الأجهزة والبرامج وسير العمل وسلاسل التوريد. تمكن AML المهاجمين من التسبب في سلوكيات غير مقصودة في أنظمة تعلم الآلة والتي يمكن أن تشمل:

- التأثير على تصنيف النموذج أو تراجع الأداء
- السماح للمستخدمين بتنفيذ إجراءات غير مصرح بها
- استخراج معلومات النموذج الحساسة

هناك العديد من الطرق لتحقيق هذه التأثيرات، مثل هجمات الحقن السريع في مجال نموذج اللغة الكبير (LLM)، أو إتلاف بيانات التدريب أو تعليقات المستخدمين عمدًا (المعروف باسم "تسمم البيانات").

من يجب عليه قراءة هذه الوثيقة؟

تستهدف هذه الوثيقة في المقام الأول مقدمي أنظمة الذكاء الاصطناعي، سواء بناءً على النماذج التي تستضيفها مؤسسة ما أو الاستفادة من واجهات برمجة التطبيقات الخارجية (APIs). ومع ذلك، فإننا نحث **جميع** أصحاب المصلحة (بما في ذلك علماء البيانات والمطورين والمديرين وصناع القرار وأصحاب المخاطر) على قراءة هذه الإرشادات لمساعدتهم على اتخاذ قرارات مستنيرة بشأن **التصميم**، **النشر والتشغيل** لأنظمة الذكاء الاصطناعي الخاصة بتعلم الآلة.

ومع ذلك، لن تنطبق جميع المبادئ التوجيهية بشكل مباشر على جميع المنظمات. سيختلف مستوى التعقيد وأساليب الهجوم اعتمادًا على الخصم الذي يستهدف نظام الذكاء الاصطناعي، لذلك يجب أخذ الإرشادات في الاعتبار جنبًا إلى جنب مع حالات استخدام مؤسستك وملف تعريف التهديد.

من هو المسؤول عن تطوير الذكاء الاصطناعي الآمن؟

غالبًا ما يكون هناك العديد من الجهات الفاعلة في سلاسل التوريد الحديثة للذكاء الاصطناعي. النهج البسيط يفترض كيانين:

- "المزود" المسؤول عن تنظيم البيانات وتطوير الخوارزميات والتصميم والنشر والصيانة
- "المستخدم" الذي يقدم المدخلات ويستقبل المخرجات

على الرغم من استخدام هذا النهج بين الموفر والمستخدم في العديد من التطبيقات، إلا أنه أصبح غير شائع على نحو متزايد[4]، حيث قد يتطلع مقدمو الخدمة إلى دمج البرامج و/أو البيانات و/أو النماذج و/أو الخدمات عن بُعد المقدمة من جهات خارجية في أنظمتهم. تجعل سلاسل التوريد المعقدة هذه من الصعب على المستخدمين النهائيين فهم أين تقع مسؤولية الذكاء الاصطناعي الآمن.

لا يتمتع المستخدمون (سواء كانوا "مستخدمين نهائيين" أو مقدمي خدمات يدمجون مكونًا خارجيًا للذكاء الاصطناعي[5]) عادةً بالرؤية و/أو الخبرة الكافية لفهم المخاطر المرتبطة بالأنظمة التي يستخدمونها أو تقييمها أو معالجتها بشكل كامل. على هذا النحو، وتماشيًا مع مبادئ "التصميم الآمن"، **يجب على موفري مكونات الذكاء الاصطناعي تحمل مسؤولية النتائج الأمنية للمستخدمين في سلسلة التوريد.**

يجب على مقدمي الخدمة تنفيذ ضوابط الأمان وعمليات تخفيف التأثير حيثما أمكن ذلك ضمن نماذجهم و/أو مساراتهم و/أو أنظمتهم، وتنفيذ الخيار الأكثر أمانًا كإعداد افتراضي حيثما يتم استخدام الإعدادات. عندما لا يمكن تخفيف المخاطر، يجب أن يكون مقدم الخدمة مسؤولاً عن:

- إعلام المستخدمين في أسفل سلسلة التوريد بالمخاطر التي يقبلونها هم ومستخدموهم (إذا وجدوا)
- تقديم المشورة لهم حول كيفية استخدام المكون بشكل آمن

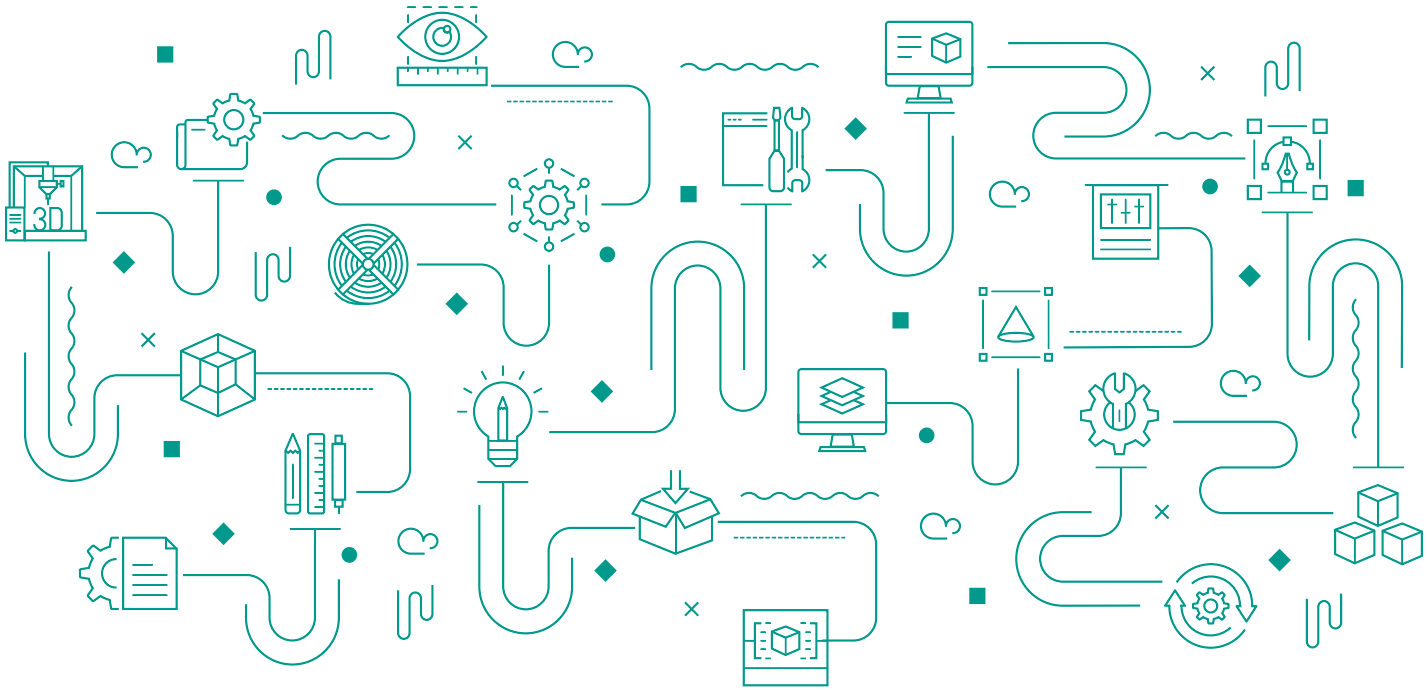
عندما يؤدي اختراق النظام إلى ضرر ملموس أو واسع النطاق على المستوى المادي أو السمعة، أو خسارة كبيرة في العمليات التجارية، أو تسرب معلومات حساسة أو سرية و/أو آثار قانونية، يجب التعامل مع مخاطر الأمن الإلكتروني للذكاء الاصطناعي على أنها **خطيرة**.

إرشادات لتطوير نظام الذكاء الاصطناعي الآمن

تم تقسيم الإرشادات إلى أربعة مجالات رئيسية ضمن دورة حياة تطوير نظام الذكاء الاصطناعي: **التصميم الآمن**، و**التطوير الآمن**، و**النشر الآمن**، و**التشغيل والصيانة الآمنة**. بالنسبة لكل مجال، نقترح الاعتبارات والتخفيفات التي من شأنها أن تساعد في تقليل المخاطر الإجمالية لعملية تطوير نظام الذكاء الاصطناعي التنظيمي.

تتوافق الإرشادات الواردة في هذه الوثيقة بشكل وثيق مع ممارسات دورة حياة تطوير البرمجيات المحددة في:

- إرشادات التطوير والنشر الآمن الخاصة بـ NCSC
- المعهد الوطني للمعايير والتكنولوجيا (NIST) إطار عمل تطوير البرمجيات الآمنة (SSDF)[6]



1. التصميم الآمن

يحتوي هذا القسم على إرشادات تنطبق على مرحلة **التصميم** في دورة حياة تطوير نظام الذكاء الاصطناعي. يغطي فهم المخاطر ونمذجة التهديدات، بالإضافة إلى موضوعات محددة ومقايضات يجب مراعاتها عند تصميم النظام والنموذج.

رفع مستوى وعي الموظفين حول التهديدات والمخاطر

يفهم مالكو النظام وكبار القادة التهديدات التي تواجه تأمين الذكاء الاصطناعي وطرق تخفيفها. يحافظ علماء ومطورو البيانات لديك على وعي بالتهديدات الأمنية ذات الصلة وأنماط الأعطال ويساعدون أصحاب المخاطر على اتخاذ قرارات مستنيرة. أنت تزود المستخدمين بالإرشادات حول المخاطر الأمنية الفريدة التي تواجه أنظمة الذكاء الاصطناعي (على سبيل المثال، كجزء من تدريب InfoSec المتعارف عليه) وتدريب المطورين على تقنيات التشفير الآمنة وممارسات الذكاء الاصطناعي الآمنة والمسؤولة.

نمذج التهديدات التي يتعرض لها نظامك

كجزء من عملية إدارة المخاطر لديك، يمكنك تطبيق عملية شاملة لتقييم التهديدات التي يتعرض لها نظامك، والتي تتضمن فهم التأثيرات المحتملة على النظام والمستخدمين والمؤسسات والمجتمع الأوسع في حالة تعرض أحد مكونات الذكاء الاصطناعي للاختراق أو التصرف بشكل غير متوقع[7]. تتضمن هذه العملية تقييم تأثير التهديدات الخاصة بالذكاء الاصطناعي[8] وتوثيق عملية اتخاذ القرار.

أنت تدرك أن حساسية وأنواع البيانات المستخدمة في نظامك قد تؤثر على قيمتها كهدف للمهاجم. يجب أن يأخذ تقييمك في الاعتبار أن بعض التهديدات قد تنمو مع تزايد النظر إلى أنظمة الذكاء الاصطناعي على أنها أهداف ذات قيمة عالية، وبما أن الذكاء الاصطناعي نفسه يتيح نواقل هجوم آلية جديدة.

صمم نظامك من أجل الأمن بالإضافة إلى الوظيفة والأداء

أنت واثق من أن المهمة المطروحة يتم معالجتها بشكل مناسب باستخدام الذكاء الاصطناعي. بعد تحديد ذلك، يمكنك تقييم مدى ملاءمة اختيارات التصميم الخاصة بالذكاء الاصطناعي. عليك أن تأخذ في الاعتبار نموذج التهديد الخاص بك وعمليات الأمن لتخفيف الأثر المرتبطة به جنبًا إلى جنب مع الوظائف وتجربة المستخدم وبيئة النشر والأداء والضمان والإشراف والمتطلبات الأخلاقية والقانونية، من بين اعتبارات أخرى. على سبيل المثال:

- يجب أن تأخذ في الاعتبار أمان سلسلة التوريد عند اختيار ما إذا كنت تريد التطوير داخليًا أو استخدام مكونات خارجية، على سبيل المثال:
  - يعد اختيارك لتدريب نموذج جديد، أو استخدام نموذج موجود (مع أو بدون ضبط دقيق) أو الوصول إلى نموذج عبر واجهة برمجة تطبيقات خارجية مناسبًا لمتطلباتك
  - يتضمن اختيارك للعمل مع مزود نموذج خارجي تقييم العناية الواجبة للوضع الأمني الخاص بهذا المزود
  - إذا كنت تستخدم مكتبة خارجية، فإنك تقوم بإكمال تقييم العناية الواجبة (على سبيل المثال، للتأكد من أن المكتبة لديها ضوابط تمنع النظام من تحميل نماذج غير موثوقة دون تعريض نفسها على الفور لتنفيذ تعليمات برمجية عشوائية[9])
  - تقوم بتنفيذ المسح والعزل/وضع الحماية عند استيراد نماذج جهة خارجية أو أوزان متسلسلة، والتي يجب التعامل معها على أنها تعليمات برمجية غير موثوق بها من جهة خارجية ويمكن أن تتيح تنفيذ التعليمات البرمجية عن بُعد

- في حالة استخدام واجهات برمجة التطبيقات الخارجية، فإنك تطبق عناصر التحكم المناسبة على البيانات التي يمكن إرسالها إلى خدمات خارجة عن سيطرة مؤسستك، مثل مطالبة المستخدمين بتسجيل الدخول والتأكيد قبل إرسال معلومات قد تكون حساسة
- تقوم بتطبيق عمليات التحقق والتنقية المناسبة للبيانات والمدخلات؛ يتضمن ذلك عند دمج تعليقات المستخدمين أو بيانات التعلم المستمر في نموذجك، مع إدراك أن بيانات التدريب تحدد سلوك النظام

- تقوم بدمج تطوير نظام برمجيات الذكاء الاصطناعي في أفضل ممارسات التطوير والعمليات الآمنة الحالية؛ تتم كتابة جميع عناصر نظام الذكاء الاصطناعي في بيئات مناسبة باستخدام ممارسات الترميز واللغات التي تقلل أو تقضي على فئات معروفة من نقاط الضعف حيثما كان ذلك ممكنًا

- إذا كانت مكونات الذكاء الاصطناعي بحاجة إلى تشغيل إجراءات، على سبيل المثال تعديل الملفات أو توجيه المخرجات إلى أنظمة خارجية، فيمكنك تطبيق القيود المناسبة على الإجراءات المحتملة (وهذا يشمل الذكاء الاصطناعي الخارجي وعمليات الأمان غير المتعلقة بالذكاء الاصطناعي إذا لزم الأمر)

- يتم اتخاذ القرارات المتعلقة بتفاعل المستخدم من خلال المخاطر الخاصة بالذكاء الاصطناعي، على سبيل المثال:

- يوفر نظامك للمستخدمين مخرجات قابلة للاستخدام دون الكشف عن مستويات غير ضرورية من التفاصيل لمهاجم محتمل
- إذا لزم الأمر، يوفر نظامك حواجز حماية فعالة حول مخرجات النموذج
- إذا كنت تقدم واجهة برمجة التطبيقات (API) للعملاء أو المتعاونين الخارجيين، فإنك تطبق عناصر التحكم المناسبة التي تخفف من الهجمات على نظام الذكاء الاصطناعي عبر واجهة برمجة التطبيقات (API).
- تقوم بدمج الإعدادات الأكثر أمانًا في النظام بشكل افتراضي
- تقوم بتطبيق مبادئ الامتيازات الأقل لتقييد الوصول إلى وظائف النظام
- تشرح للمستخدمين القدرات الأكثر خطورة وتطلب من المستخدمين الاشتراك لاستخدامها؛ تقوم بالإبلاغ عن حالات الاستخدام المحظورة، وإبلاغ المستخدمين، حيثما أمكن ذلك، بالحلول البديلة

ضع في اعتبارك المزايا الأمنية والمقايضات عند اختيار نموذج الذكاء الاصطناعي الخاص بك

سيتضمن اختيارك لنموذج الذكاء الاصطناعي تحقيق التوازن بين مجموعة من المتطلبات، يتضمن ذلك اختيار بنية النموذج والتكوين وبيانات التدريب وخوارزمية التدريب والمعلمات الفائقة. يتم اتخاذ قراراتك بناءً على نموذج التهديد الخاص بك، ويتم إعادة تقييمها بانتظام مع تقدم أبحاث أمان الذكاء الاصطناعي وتطور فهم التهديد.

عند اختيار نموذج الذكاء الاصطناعي، من المحتمل أن تتضمن اعتباراتك، على سبيل المثال لا الحصر، ما يلي:

- مدى تعقيد النموذج الذي تستخدمه، أي البنية المختارة وعدد المعلمات؛ ستؤثر البنية المختارة لنموذجك وعدد المعلمات، من بين عوامل أخرى، على مقدار بيانات التدريب التي يتطلبها ومدى قوة التغييرات في بيانات الإدخال عند الاستخدام

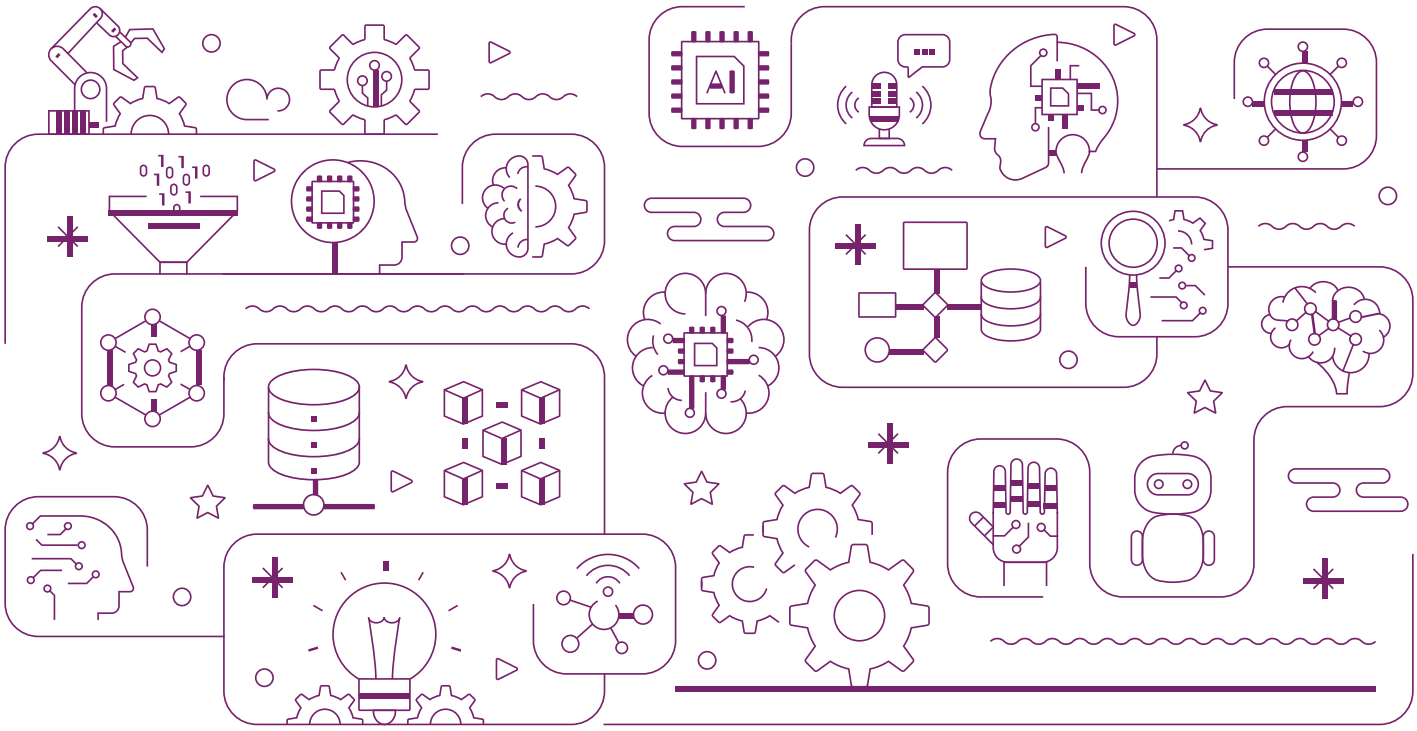
- مدى ملاءمة النموذج لحالة الاستخدام الخاصة بك و/أو جدوى تكييفه مع احتياجاتك المحددة (على سبيل المثال عن طريق الضبط الدقيق)

- القدرة على محاذاة وتفسير وشرح مخرجات نموذجك (على سبيل المثال لتصحيح الأخطاء، التدقيق أو الامتثال التنظيمي)؛ قد تكون هناك فوائد لاستخدام نماذج أبسط وأكثر شفافية مقارنة بالنماذج الكبيرة والمعقدة التي يصعب تفسيرها

- خصائص مجموعة (مجموعات) بيانات التدريب، بما في ذلك الحجم والاكتمال والجودة والحساسية والعمر والأهمية والتنوع

- قيمة استخدام تقوية النموذج (مثل التدريب المضاد)، و/أو التنظيم و/أو تقنيات تعزيز الخصوصية
- مصدر وسلاسل التوريد للمكونات بما في ذلك النموذج أو النموذج الأساسي وبيانات التدريب والأدوات المرتبطة بها

لمزيد من المعلومات حول عدد هذه العوامل التي تؤثر على نتائج الأمان، راجع "مبادئ أمان تعلم الآلة / "Principles for the Security of Machine Learning" الصادرة عن NCSC، وعلى وجه الخصوص التصميم من أجل الأمن (بنية النموذج).



2. التطوير الآمن

يحتوي هذا القسم على إرشادات تنطبق على مرحلة **التطوير** من دورة حياة تطوير نظام الذكاء الاصطناعي، بما في ذلك أمن سلسلة التوريد والتوثيق وإدارة الأصول والديون الفنية.

تأمين سلسلة توريدك

أنت تقوم بتقييم ومراقبة أمان سلاسل توريد الذكاء الاصطناعي الخاصة بك عبر دورة حياة النظام، وتطلب من الموردين الالتزام بنفس المعايير التي تطبقها مؤسستك على البرامج الأخرى. إذا لم يتمكن الموردون من الالتزام بمعايير مؤسستك، فإنك تتصرف وفقًا لسياسات إدارة المخاطر الحالية لديك.

عندما لا يتم إنتاجها داخليًا، فإنك تحصل على مكونات الأجهزة والبرامج وتحافظ عليها آمنة وموثقة جيدًا (على سبيل المثال، النماذج والبيانات ومكتبات البرامج والوحدات النمطية والبرامج الوسيطة والأطر وواجهات برمجة التطبيقات الخارجية) من مصادر تجارية مفتوحة المصدر ومعتمدة، ومطوري الطرف الثالث الآخرين لضمان أمان قوي في أنظمتك.

أنت جاهز للانتقال إلى حلول بديلة للأنظمة ذات المهام الحرجة، إذا لم يتم استيفاء معايير الأمان. يمكنك استخدام موارد مثل إرشادات سلسلة التوريد / Supply Chain Guidance الخاصة بـ NCSC وأطر العمل مثل مستويات سلسلة التوريد لعناصر البرامج (SLSA) / [10] Supply Chain Levels for Software Artifacts (SLSA) لتتبع شهادات دورات حياة تطوير سلسلة التوريد والبرامج.

تحديد وتتبع وحماية أصولك

أنت تدرك قيمة أصولك المتعلقة بالذكاء الاصطناعي بالنسبة لمؤسستك، بما في ذلك النماذج والبيانات (بما في ذلك تعليقات المستخدمين) والمطالبات والبرامج والوثائق والسجلات والتقييمات (بما في ذلك المعلومات حول القدرات التي قد تكون غير آمنة وأوضاع الأعطال). مع إدراك أين تمثل استثمار كبير وحيث الوصول إليها يمكّن المهاجم. أنت تتعامل مع السجلات كبيانات حساسة وتنفذ ضوابط لحماية سريتها وسلامتها وتوافرها.

أنت تعرف مكان وجود أصولك وقمت بتقييم وقبول أي مخاطر مرتبطة بها. لديك عمليات وأدوات لتتبع أصولك والمصادقة عليها والتحكم في الإصدار وتأمينها. ويمكنك استعادة حالة جيدة معروفة في حالة حدوث اختراق.

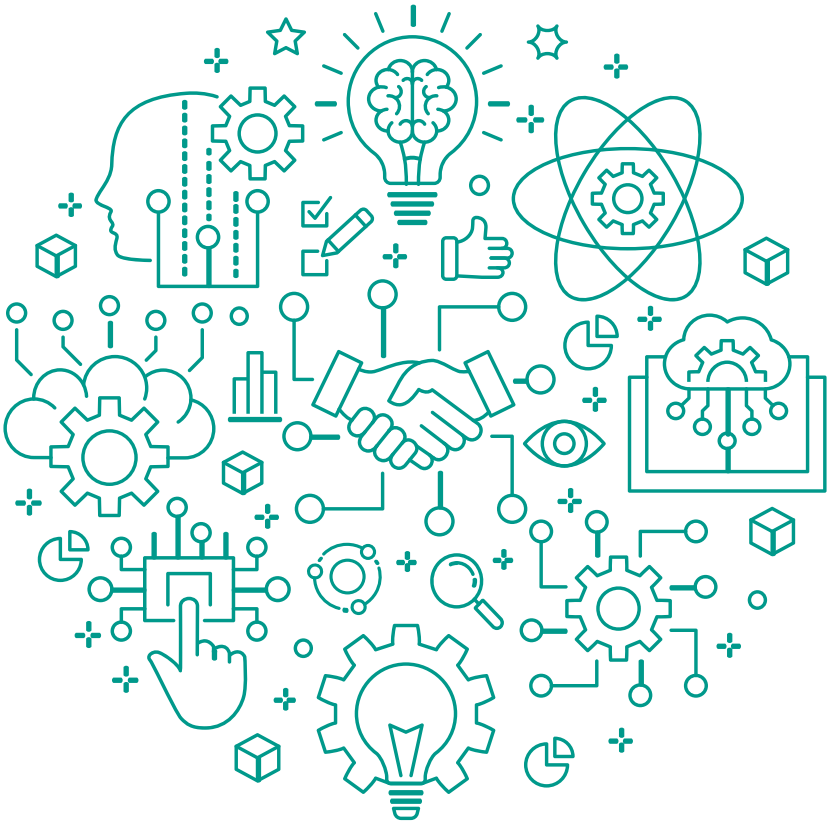
لديك عمليات وضوابط لإدارة البيانات التي يمكن لأنظمة الذكاء الاصطناعي الوصول إليها، وإدارة المحتوى الذي ينشئه الذكاء الاصطناعي وفقًا لحساسيته (وحساسية المدخلات التي دخلت في إنشائه).

قم بتوثيق بياناتك ونماذجك ومطالباتك

أنت توثق الإنشاء والتشغيل وإدارة دورة الحياة لأي نماذج ومجموعات بيانات ومطالبات تعريفية أو نظام. تتضمن وثائقك معلومات ذات صلة بالأمان مثل مصادر بيانات التدريب (بما في ذلك بيانات الضبط الدقيق والتعليقات البشرية أو غيرها من التعليقات التشغيلية)، والنطاق والقيود المقصودة، وحواجز الحماية، وتجزئة التشفير أو التوقيعات، ووقت الاحتفاظ، وتكرار المراجعة المقترحة، وأنماط الأعطال المحتملة . تشتمل الهياكل المفيدة للمساعدة في القيام بذلك على البطاقات النموذجية وبطاقات البيانات وفواتير المواد البرمجية (SBOMs). إن إنتاج وثائق شاملة يدعم الشفافية والمسؤولية[11].

إدارة ديونك الفنية

كما هو الحال مع أي نظام برمجي، يمكنك تحديد وتتبع وإدارة "دينك الفني" طوال دورة حياة نظام الذكاء الاصطناعي (الدين الفني هو حيث يتم اتخاذ القرارات الهندسية التي لا ترقى إلى مستوى أفضل الممارسات لتحقيق نتائج قصيرة المدى، على حساب فوائد طويلة الأجل). مثل الديون المالية، فإن الديون الفنية ليست سيئة بطبيعتها، ولكن يجب إدارتها منذ المراحل الأولى من التطوير[12]. أنت تدرك أن القيام بذلك قد يكون أكثر صعوبة في سياق الذكاء الاصطناعي مقارنة بالبرمجيات القياسية، وأن مستويات الديون الفنية لديك من المحتمل أن تكون مرتفعة بسبب دورات التطوير السريعة والافتقار إلى البروتوكولات والواجهات الراسخة. أنت تتأكد من أن خططك لدورة الحياة (بما في ذلك عمليات إيقاف تشغيل أنظمة الذكاء الاصطناعي) تقوم بتقييم المخاطر التي تتعرض لها الأنظمة المماثلة المستقبلية والاعتراف بها وتخفيف أثرها.



3. النشر الآمن

يحتوي هذا القسم على إرشادات تنطبق على مرحلة **النشر** من دورة حياة تطوير نظام الذكاء الاصطناعي، بما في ذلك حماية البنية الأساسية والنماذج من الاختراق أو التهديد أو الخسارة، وتطوير عمليات إدارة الحوادث، والإطلاق المسؤول.

تأمين بنيتك التحتية

يمكنك تطبيق مبادئ أمان البنية التحتية الجيدة على البنية التحتية المستخدمة في كل جزء من دورة حياة نظامك. يمكنك تطبيق ضوابط الوصول المناسبة على واجهات برمجة التطبيقات والنماذج والبيانات الخاصة بك، وعلى مسارات التدريب والمعالجة الخاصة بها، في البحث والتطوير وكذلك النشر. يتضمن ذلك الفصل المناسب بين البيئات التي تحتوي على تعليمات برمجية أو بيانات حساسة. سيساعد هذا أيضًا في التخفيف من هجمات الأمن السيبراني المتعارف عليها التي تهدف إلى سرقة نموذج أو الإضرار بأدائه.

حماية نموذجك بشكل مستمر

قد يتمكن المهاجمون من إعادة بناء وظائف النموذج[13] أو البيانات التي تم التدريب عليها[14]، من خلال الوصول إلى النموذج مباشرةً (من خلال الحصول على أوزان النموذج) أو بشكل غير مباشر (من خلال الاستعلام عن النموذج عبر تطبيق أو خدمة). وقد يتلاعب المهاجمون أيضًا بالنماذج أو البيانات أو المطالبات أثناء التدريب أو بعده، مما يجعل المخرجات غير جديرة بالثقة.

يمكنك حماية النموذج والبيانات من الوصول المباشر وغير المباشر، على التوالي، من خلال:

- تنفيذ أفضل الممارسات القياسية للأمن السيبراني
- تنفيذ ضوابط على واجهة الاستعلام لاكتشاف ومنع محاولات الوصول إلى المعلومات السرية وتعديلها وتصفيتها

للتأكد من أن الأنظمة المستهلكة يمكنها التحقق من صحة النماذج، يمكنك حساب ومشاركة متجزئات التشفير و/أو التوقيعات لملفات النموذج (على سبيل المثال، أوزان النماذج) ومجموعات البيانات (بما في ذلك نقاط التفتيش) بمجرد تدريب النموذج. كما هو الحال دائمًا مع التشفير، تعتبر الإدارة الجيدة للمفاتيح أمرًا ضروريًا[15].

سيعتمد أسلوبك في تخفيف مخاطر السرية إلى حد كبير على حالة الاستخدام ونموذج التهديد. قد تتطلب بعض التطبيقات، على سبيل المثال تلك التي تتضمن بيانات حساسة للغاية، ضمانات نظرية قد يكون تطبيقها صعبًا أو مكلفًا. إذا كان ذلك مناسبًا، يمكن استخدام تقنيات تعزيز الخصوصية (مثل الخصوصية التفاضلية أو التشفير المتماثل) لاستكشاف أو ضمان مستويات المخاطر المرتبطة بالمستهلكين والمستخدمين والمهاجمين الذين لديهم إمكانية الوصول إلى النماذج والمخرجات.

تطوير إجراءات إدارة الحوادث

تنعكس حتمية الحوادث الأمنية التي تؤثر على أنظمة الذكاء الاصطناعي لديك في خطط الاستجابة للحوادث والتصعيد والمعالجة. تعكس خططك سيناريوهات مختلفة ويتم إعادة تقييمها بانتظام مع تطور النظام والبحث على نطاق أوسع. يمكنك تخزين الموارد الرقمية المهمة للشركة في نسخ احتياطية دون اتصال بالإنترنت. تم تدريب المستجيبين على تقييم ومعالجة الحوادث المتعلقة بالذكاء الاصطناعي. أنت تقدم سجلات تدقيق عالية الجودة وميزات أو معلومات أمان أخرى للعملاء والمستخدمين دون أي رسوم إضافية، لتمكين عمليات الاستجابة للحوادث الخاصة بهم.

إطلاق الذكاء الاصطناعي بمسؤولية

لا يجوز لك إطلاق النماذج أو التطبيقات أو الأنظمة إلا بعد إخضاعها لتقييم أمني مناسب وفعال مثل قياس الأداء والفريق الأحمر ألا وهي محاكاة الهجمات (بالإضافة إلى الاختبارات الأخرى التي تقع خارج نطاق هذه الإرشادات، مثل السلامة أو العدالة). وأنك واضح في ذلك للمستخدمين لديك حول القيود المعروفة أو أوضاع الأعطال المحتملة. تتوفر تفاصيل مكتبات اختبار الأمان مفتوحة المصدر في قسم القراءة الإضافية في نهاية هذه الوثيقة.

سهّل على المستخدمين القيام بالأشياء الصحيحة

أنت تدرك أنه سيتم تقييم كل إعداد أو خيار تكوين جديد جنبًا إلى جنب مع المزايا التجارية التي يستمدها، وأي مخاطر أمنية يقدمها. ومن الناحية المثالية، سيتم دمج الإعداد الأكثر أمانًا في النظام باعتباره الخيار الوحيد. عندما يكون التكوين ضروريًا، يجب أن يكون الخيار الافتراضي آمنًا على نطاق واسع ضد التهديدات الشائعة (أي آمن افتراضيًا). يمكنك تطبيق عناصر التحكم لمنع استخدام نظامك أو نشره بطرق ضارة.

أنت تقدم للمستخدمين إرشادات حول الاستخدام المناسب لنموذجك أو نظامك، والذي يتضمن تسليط الضوء على القيود وأنماط الفشل المحتملة. أنت تقول بوضوح للمستخدمين جوانب الأمان التي يتحملون المسؤولية عنها، وتتسم بالشفافية بشأن مكان (وكيفية) استخدام بياناتهم أو الوصول إليها أو تخزينها (على سبيل المثال، إذا تم استخدامها لإعادة تدريب النماذج، أو مراجعتها من قبل الموظفين أو الشركاء).

4. التشغيل والصيانة الآمنة

يحتوي هذا القسم على إرشادات تنطبق على مرحلة **التشغيل والصيانة الآمنة** من دورة حياة تطوير نظام الذكاء الاصطناعي. يوفر هذا القسم إرشادات حول الإجراءات ذات الصلة بشكل خاص بمجرد نشر النظام ، بما في ذلك التسجيل والمراقبة وإدارة التحديث ومشاركة المعلومات.

مراقبة سلوك نظامك

يمكنك قياس مخرجات وأداء نموذجك ونظامك بحيث يمكنك ملاحظة التغيرات المفاجئة والتدريجية في السلوك التي تؤثر على الأمان. يمكنك حساب وتحديد عمليات التطفل والاختراق المحتملة، بالإضافة إلى الانجراف الطبيعي للبيانات.

راقب مدخلات نظامك

تماشيًا مع متطلبات الخصوصية وحماية البيانات، يمكنك مراقبة وتسجيل المدخلات إلى نظامك (مثل طلبات الاستدلال أو الاستعلامات أو المطالبات) لتمكين التزامات الامتثال والتدقيق والتحقيق والمعالجة في حالة الاختراق أو سوء الاستخدام. يمكن أن يشمل ذلك الكشف الصريح عن المدخلات الغريبة و/أو العدائية، بما في ذلك تلك التي تهدف إلى استغلال خطوات إعداد البيانات (مثل الاقتصاص وتغيير حجم الصور).

اتبع نهج تصميم آمنًا للتحديثات

يمكنك تضمين التحديثات التلقائية بشكل افتراضي في كل منتج واستخدام إجراءات التحديث المعيارية الآمنة لتوزيعها. تعكس عمليات التحديث الخاصة بك (بما في ذلك أنظمة الاختبار والتقييم) حقيقة أن التغييرات في البيانات أو النماذج أو المطالبات يمكن أن تؤدي إلى تغييرات في سلوك النظام (على سبيل المثال، تعامل التحديثات الرئيسية مثل الإصدارات الجديدة). أنت تدعم المستخدمين في تقييم تغييرات النموذج و الاستجابة لها (على سبيل المثال من خلال توفير الوصول للمعاينة وواجهات برمجة التطبيقات ذات الإصدارات).

جمع وتبادل الدروس المستفادة

أنت تشارك في مجتمعات تبادل المعلومات، وتتعاون عبر النظام البيئي العالمي للصناعة والأوساط الأكاديمية والحكومات لمشاركة أفضل الممارسات حسب الاقتضاء. يجب عليك الحفاظ على خطوط اتصال مفتوحة للحصول على تعليقات بشأن أمان النظام، داخليًا وخارجيًا لمؤسستك، بما في ذلك تقديم الموافقة للباحثين الأمنيين للبحث عن نقاط الضعف والإبلاغ عنها. عند الحاجة، يمكنك تصعيد المشكلات إلى المجتمع الأوسع، على سبيل المثال، نشر النشرات التي تستجيب لعمليات الكشف عن الثغرات الأمنية، بما في ذلك تعداد الثغرات المشتركة المفصل والكامل. أنت تتخذ الإجراءات اللازمة للتخفيف من حدة المشكلات ومعالجتها بسرعة وبشكل مناسب.

القراءة الإضافية

تطوير الذكاء الاصطناعي

مبادئ أمن تعلم الآلة
إرشادات مفصلة من NCSC حول تطوير أو نشر أو تشغيل نظام يحتوي على مكون تعلم الآلة ML.

مبادئ وأساليب من أجل برمجيات التصميم الآمن / Principles and Approaches for Secure by Design Software
تم تأليف هذا الدليل بالاشتراك مع CISA و NCSC ووكالات أخرى. ويصف هذا الدليل كيف ينبغي لمصنعي أنظمة البرمجيات، بما في ذلك الذكاء الاصطناعي، اتخاذ خطوات لأخذ الأمن في الاعتبار في مرحلة التصميم لتطوير المنتج، وشحن المنتجات التي تصل آمنة بادئ ذي بدء

المخاوف المتعلقة بأمن الذكاء الاصطناعي باختصار / AI Security Concerns in a Nutshell
تقدم هذه الوثيقة، التي أصدرها المكتب الفيدرالي الألماني لأمن المعلومات (BSI)، مقدمة للهجمات المحتملة على أنظمة تعلم الآلة والدفاعات المحتملة ضد تلك الهجمات.

المبادئ التوجيهية الدولية لعملية هيروشيما للمؤسسات التي تطور أنظمة ذكاء اصطناعي متقدمة / Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems و
مدونة قواعد السلوك الدولية لعملية هيروشيما للمؤسسات التي تطور أنظمة ذكاء اصطناعي متقدمة / Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems
هذه الوثائق، التي تم إنتاجها كجزء من عملية هيروشيما للذكاء الاصطناعي لمجموعة السبع، تقدم إرشادات للمؤسسات التي تعمل على تطوير أنظمة الذكاء الاصطناعي الأكثر تقدمًا، بما في ذلك النماذج الأساسية الأكثر تقدمًا وأنظمة الذكاء الاصطناعي التوليدية بهدف تعزيز الذكاء الاصطناعي الآمن والمأمون والجدير بالثقة في جميع أنحاء العالم.

إي آي فيريفاي / AI Verify
إطار عمل اختبار حوكمة الذكاء الاصطناعي في سنغافورة ومجموعة الأدوات البرمجية التي تتحقق من صحة أداء أنظمة الذكاء الاصطناعي مقابل مجموعة من المبادئ المعترف بها دوليًا من خلال اختبارات موحدة.

إطار عمل متعدد الطبقات لممارسات الأمن السيبراني الجيدة للذكاء الاصطناعي / Multilayer Framework for Good Cybersecurity Practices for AI — ENISA (europa.eu)
إطار عمل لتوجيه السلطات الوطنية المختصة وأصحاب المصلحة في مجال الذكاء الاصطناعي بشأن الخطوات التي يتعين عليهم اتباعها لتأمين أنظمة وعمليات ونهج الذكاء الاصطناعي الخاصة بهم.

آيزو / ISO 5338: عمليات دورة حياة نظام الذكاء الاصطناعي (قيد المراجعة)
مجموعة من العمليات والمفاهيم المرتبطة بها لوصف دورة حياة أنظمة الذكاء الاصطناعي استنادًا إلى تعلم الآلة والأنظمة التجريبية.

فهرس معايير الامتثال لخدمة الذكاء الاصطناعي السحابية (AIC4) / AI Cloud Service Compliance Criteria Catalogue (AIC4)
يوفر فهرس معايير الامتثال لخدمة الذكاء الاصطناعي السحابية من BSI معايير خاصة بالذكاء الاصطناعي، والتي تمكن من تقييم أمان خدمة الذكاء الاصطناعي عبر دورة حياتها.

NIST IR 8269 (مسودة) تصنيف ومصطلحات تعلم الآلة المضاد / NIST IR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning
مجموعة من العمليات والمفاهيم المرتبطة بها لوصف دورة حياة أنظمة الذكاء الاصطناعي استنادًا إلى تعلم الآلة والأنظمة التجريبية.

مايتر أطلس / MITRE ATLAS
قاعدة معرفية لتكتيكات وتقنيات ودراسات الحالة الخاصة بأنظمة تعلم الآلة (LM)، والتي تم تصميمها وربطها وفقًا لإطار عمل MITRE ATT&CK.

نظرة عامة على مخاطر الذكاء الاصطناعي الكارثية (2023) / An Overview of Catastrophic AI Risks (2023)
يحدد هذا المستند الذي أصدره مركز سلامة الذكاء الاصطناعي، مجالات المخاطر التي يشكلها الذكاء الاصطناعي.

نماذج اللغة الكبيرة: الفرص والمخاطر للصناعة والسلطات/ Opportunities and Risks for Industry and Authorities
وثيقة من إنتاج BSI للشركات والهيئات والمطورين الذين يرغبون في معرفة المزيد حول الفرص والمخاطر المتعلقة بتطوير ونشر و/أو استخدام LLMs.

تتضمن المشاريع مفتوحة المصدر لمساعدة المستخدمين على اختبار نماذج الذكاء الاصطناعي ما يلي:

- مجموعة أدوات قوة الخصومة / Adversarial Robustness Toolbox (IBM)/ آي بي إم
- كلفر هانز / CleverHans (جامعة تورنتو)
- تكست أتاك / TextAttack (جامعة فرجينيا)
- برومت بنش / Prompt Bench (مايكروسفت)
- كاونترفيت / Counterfit (مايكروسفت)
- إي آي فيريفاي / AI Verify (هيئة تطوير وسائل الإعلام Infocomm، سنغافورة)

الأمن السيبراني

أهداف أداء الأمن السيبراني الخاصة بـ CISA / CISA's Cybersecurity Performance Goals
مجموعة مشتركة من وسائل الحماية التي يجب على جميع كيانات البنية التحتية الحيوية تنفيذها لتقليل احتمالية وتأثير المخاطر المعروفة والتقنيات العدائية بشكل فعال.

إطار عمل NCSC CAF / NCSC CAF Framework
يوفر إطار عمل التقييم السيبراني (CAF) إرشادات للمؤسسات المسؤولة عن الخدمات والأنشطة ذات الأهمية الحيوية.

إطار عمل أمان سلسلة توريد MITRE / MITRE's Supply Chain Security Framework
إطار عمل لتقييم الموردين ومقدمي الخدمات داخل سلسلة التوريد.

إدارة المخاطر

إطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF) التابع لـ NIST / NIST AI Risk Management Framework (AI RMF)
يحدد إطار عمل إدارة مخاطر الذكاء الاصطناعي (AI RMF) كيفية إدارة المخاطر الاجتماعية التقنية للأفراد والمؤسسات والمجتمع المرتبطة بشكل فريد بالذكاء الاصطناعي.

آيزو / ISO 27001: أمن المعلومات والأمن السيبراني وحماية الخصوصية
يوفر هذا المعيار إرشادات للمؤسسات حول إنشاء وتنفيذ وصيانة نظام إدارة أمن المعلومات.

آيزو / ISO 31000: إدارة المخاطر
معيار دولي يزود المؤسسات بإرشادات ومبادئ لإدارة المخاطر داخل المنظمات.

إرشادات إدارة مخاطر NCSC / NCSC Risk Management Guidance
يساعد هذا التوجيه ممارسي مخاطر الأمن السيبراني على فهم وإدارة مخاطر الأمن السيبراني التي تؤثر على مؤسساتهم بشكل أفضل.

ملحوظات

1. يتم تعريفه هنا على أنه شخص أو سلطة عامة أو وكالة أو هيئة أخرى تقوم بتطوير نظام الذكاء الاصطناعي (أو لديه نظام ذكاء اصطناعي تم تطويره) ويطرح هذا النظام في السوق أو يضعه في الخدمة تحت اسمه أو علامته التجارية
2. للحصول على مزيد من المعلومات حول التصميم الآمن، راجع صفحة الإنترنت التصميم الآمن التابعة لـ CISA وإرشاداتها تحويل توازن مخاطر الأمن السيبراني: مبادئ وأساليب برمجيات التصميم الآمن
3. على عكس أساليب الذكاء الاصطناعي غير المتعلقة بتعلم الآلة، مثل الأنظمة القائمة على القواعد
4. يصف CEPS سبعة أنواع مختلفة من تفاعل تطوير الذكاء الاصطناعي في منشوره "التوفيق بين سلسلة قيمة الذكاء الاصطناعي وقانون الذكاء الاصطناعي للاتحاد الأوروبي / "Reconciling the AI Value Chain with the EU's Artificial Intelligence Act"
5. ISO/IEC 22989:2022(en) يُعرّف ذلك بأنه "عنصر وظيفي يبني نظام الذكاء الاصطناعي"
6. تم تكليف NIST بإنتاج إرشادات (واتخاذ إجراءات أخرى) لتعزيز تطوير واستخدام الذكاء الاصطناعي (AI) بطريقة آمنة ومأمونة وجديرة بالثقة. راجع مسؤوليات NIST بموجب الأمر التنفيذي الصادر في 30 تشرين الأول/أكتوبر 2023
7. يتوفر المزيد من المعلومات حول نماذج التهديدات من مؤسسة أواسب / OWASP Foundation
8. راجع MITRE ATLAS تعلم الآلة المضاد 101 / Adversarial Machine Learning 101
9. جيت هب / GitHub: RCE PoC لـ Tensorflow باستخدام طبقة Lambda الضارة
10. SLSA: "حماية سلامة المنتجات عبر أية سلسلة توريد برمجيات"
11. METI (وزارة الاقتصاد والتجارة والصناعة اليابانية، 2023). "دليل تقديم قائمة مواد البرمجيات (SBOM) لإدارة البرمجيات"
12. أبحاث غوغل / Google: تعلم الآلة: بطاقة الائتمان ذات الفائدة المرتفعة للديون الفنية
13. ترامير وآخرون 2016 / Tramèr et al 2016، سرقة نماذج تعلم الآلة عبر واجهات برمجة التطبيقات للتنبؤ
14. بوينيش / Boenisch، 2020، الهجمات ضد خصوصية تعلم الآلة (الجزء الأول): هجمات الانقلاب النموذجية باستخدام إطار عمل IBM-ART
15. المركز الوطني للأمن السيبراني، 2020، تصميم وإنشاء بنية تحتية أساسية عامة مستضافة بشكل خاص