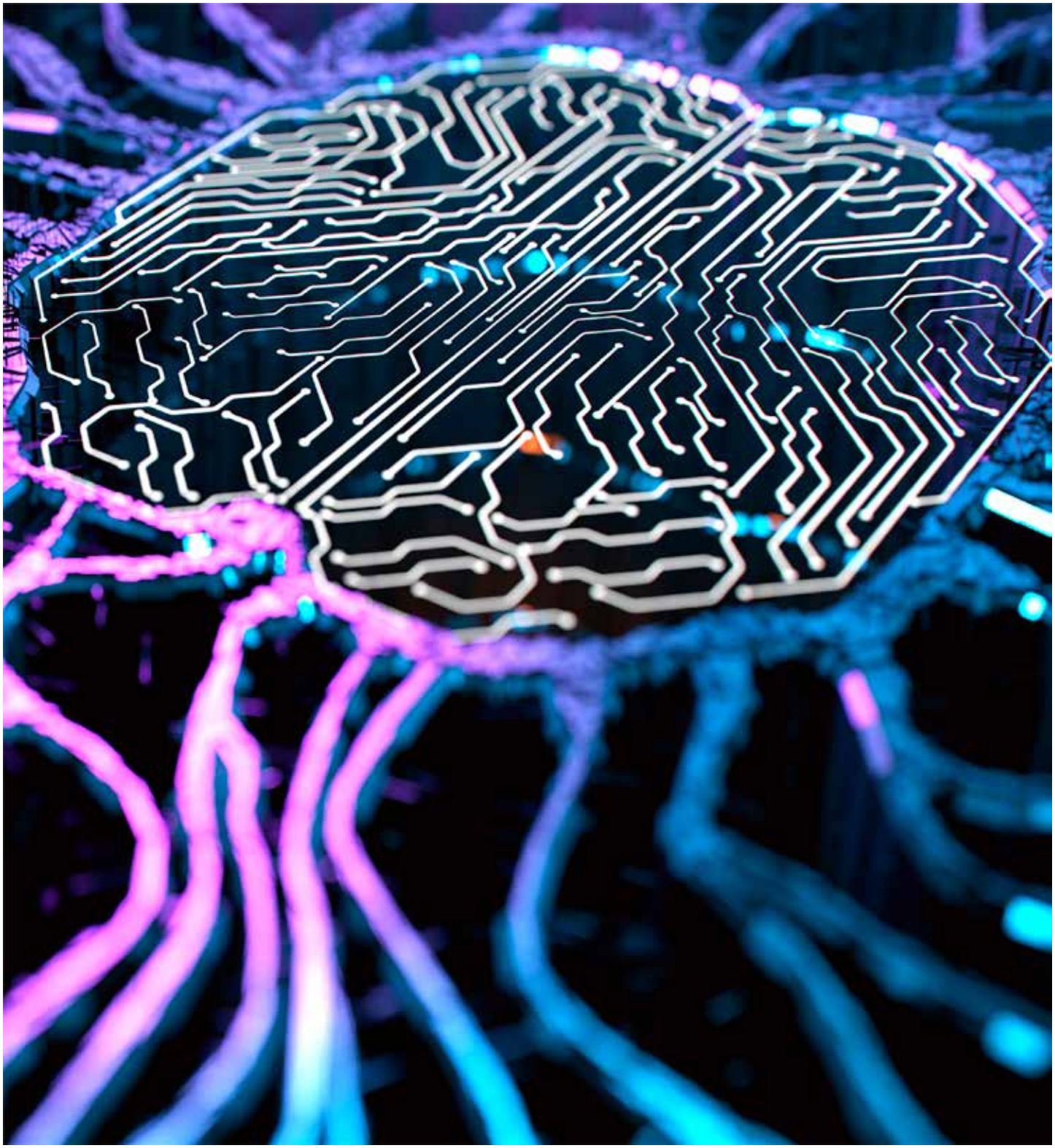
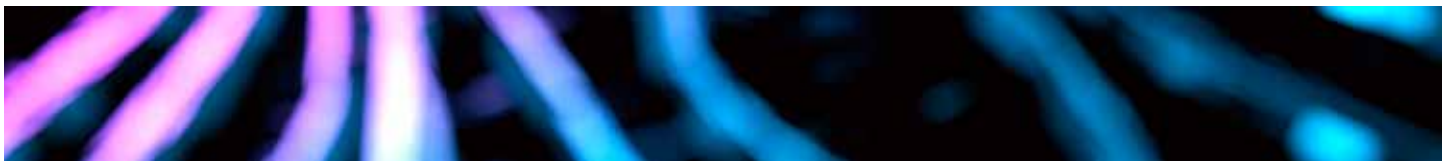
सुरक्षित AI सिस्टम विकसित करने के लिए दिशा-निर्देश





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

इस दस्तावेज के बारे में

यह दस्तावेज यूके नेशनल साइबर सिक्योरिटी सेंटर (NCSC), यूएस साइबर सिक्योरिटी एंड इंफ्रास्ट्रक्चर सिक्योरिटी एजेंसी (CISA) और निम्नलिखित अंतरराष्ट्रीय साझेदारों द्वारा प्रकाशित किया गया है:

- नेशनल सिक्योरिटी एजेंसी (NSA)
- फेडरल ब्यूरो ऑफ इन्वेस्टिगेशन्स (FBI)
- ऑस्ट्रेलियन सिग्नल्स डायरेक्टोरेट का ऑस्ट्रेलियन साइबर सिक्योरिटी सेंटर (ACSC)
- कैनेडियन सेंटर फॉर साइबर सिक्योरिटी (CCCS)
- न्यू ज़ीलैंड नेशनल साइबर सिक्योरिटी सेंटर (NCSC-NZ)
- चिली की सरकार का CSIRT
- चेकिया की नेशनल साइबर एंड इन्फोर्मेशन सिक्योरिटी एजेंसी (NUKIB)
- इन्फोर्मेशन सिस्टम ऑथोरिटी ऑफ एस्टोनिया (RIA) एवं नेशनल साइबर सिक्योरिटी सेंटर ऑफ एस्टोनिया (NCSC-EE)
- फ्रेंच साइबरसिक्योरिटी एजेंसी (ANSSI)
- जर्मनी का फेडरल ऑफिस फॉर इन्फोर्मेशन सिक्योरिटी (BSI)
- इजरायल नेशनल साइबर डायरेक्टोरेट (INCD)
- इटैलियन नेशनल साइबरसिक्योरिटी एजेंसी (ACN)
- जापान का नेशनल सेंटर ऑफ इंसिडेंट रेडिनेस एंड स्ट्रैटजी फॉर साइबरसिक्योरिटी (NISC)
- जापान का सेक्रटेरिएट ऑफ साइंस, टेक्नोलॉजी एंड इनोवेशन पॉलिसी, कैबिनेट ऑफिस
- नाइजीरिया की नेशनल इन्फोर्मेशन टेक्नोलॉजी डेवलेपमेंट एजेंसी (NITDA)
- नॉर्वेजियन नेशनल साइबर सिक्योरिटी सेंटर (NCSC-NO)
- पोलैंड मिनिस्ट्री ऑफ डिजिटल एफेयर्स
- पोलैंड का NASK नेशनल रिसर्च इंस्टिट्यूट (NASK)
- कोरिया गणराज्य की नेशनल इंटेलिजेंस सर्विस (NIS)
- सिंगापुर की साइबर सिक्योरिटी एजेंसी (CSA)

अभिस्वीकृतियाँ

इन दिशा-निर्देशों के डेवलपमेंट में निम्नलिखित संगठनों ने योगदान दिया है:

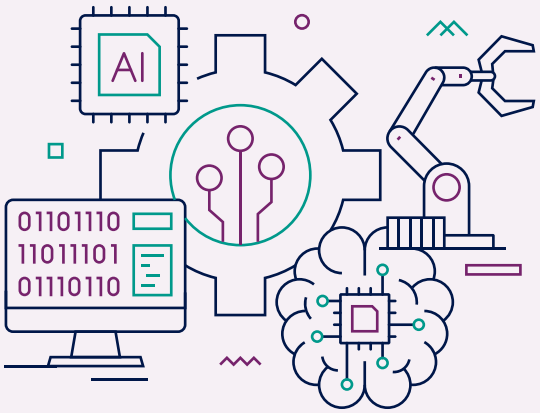
- एलन ट्यूरिंग इंस्टिट्यूट
- एंथ्रोपिक
- डेटाब्रिक्स
- जॉर्जटाउन विश्वविद्यालय का सेंटर फॉर सिक्योरिटी एंड इमर्जिंग टेक्नोलॉजी
- गूगल
- गूगल डीपमाइंड
- IBM
- इमब्यू
- माइक्रोसॉफ़्ट
- ओपनAI
- पैलांटिर
- रैंड
- स्केल AI
- कार्नेगी मेलन विश्वविद्यालय में सॉफ़्टवेयर इंजीनियरिंग इंस्टिट्यूट
- स्टैन्फर्ड सेंटर फॉर AI सिक्योरिटी
- स्टैन्फर्ड प्रोग्राम ऑन जियोपॉलिटिक्स, टेक्नोलॉजी एंड गवर्नेंस

अस्वीकरण

इस दस्तावेज में दी गई जानकारी को NCSC और लेखक संगठनों द्वारा "यथावत" रूप में उपलब्ध कराया गया है और वे इसके उपयोग से होने वाली किसी भी प्रकार की हानि, शारीरिक नुकसान या क्षति के लिए उत्तरदायी नहीं होंगे, जिसमें कानूनी आवश्यकता के लिए अपवाद है। इस दस्तावेज में दी गई जानकारी NCSC और प्राधिकृत एजेंसियों द्वारा किसी भी तृतीय पक्ष के संगठन, उत्पाद या सेवा के लिए समर्थन या सिफारिश या संकेत नहीं देती है। वेबसाइटों और तृतीय पक्ष की सामग्री के लिंक्स और संदर्भ केवल जानकारी के उद्देश्य से ही प्रदान किए गए हैं और वे किसी अन्य की तुलना में इन संसाधनों के समर्थन या सिफारिश का प्रतिनिधित्व नहीं करते हैं।

यह दस्तावेज TLP:CLEAR आधार पर उपलब्ध कराया गया है (https://www.first.org/tlp/).

सामग्री



कार्यकारी सारांश

यह दस्तावेज आर्टिफिशियल इंटेलिजेंस (AI) का उपयोग करने वाले किसी भी सिस्टम के प्रोवाइडर्स के लिए दिशा-निर्देशों की सलाह देता है, चाहे वे सिस्टम बिल्कुल नए बनाए गए हों या दूसरों द्वारा प्रदान किए गए उपकरणों और सेवाओं पर आगे निर्मित किए गए हों। इन दिशा-निर्देशों के लागूकरण से प्रोवाइडर्स को ऐसे AI सिस्टम्स बनाने में सहायता मिलेगी, जो इरादे के अनुरूप कार्य करते हैं, आवश्यकता पड़ने पर उपलब्ध रहते हैं, और अनधिकृत पक्षों के समक्ष संवेदनशील डेटा प्रकट किए बिना कार्य करते हैं।

यह दस्तावेज मुख्य रूप से AI सिस्टम्स प्रोवाइडर्स के प्रति लक्षित है, जो किसी संगठन द्वारा होस्ट किए गए मॉडल का उपयोग कर रहे हैं, या एक्स्टर्नल एप्लिकेशन प्रोग्रामिंग इंटरफेसेज़ (APIs) का उपयोग कर रहे हैं। हम **सभी** हितधारकों से (जिसमें डेटा वैज्ञानिक, डेवलपर्स, प्रबंधक, निर्णयकर्ता और खतरा उठाने वाले लोग शामिल हैं) आग्रह करते हैं कि वे इन दिशा-निर्देशों को पढ़ें, ताकि उन्हें अपने AI सिस्टम्स के **डिज़ाइन**, **डेवलपमेंट**, **डिप्लॉयमेंट** और **ऑपरेशन** के बारे में सूचित निर्णय लेने में सहायता मिल सके।

दिशा-निर्देशों के बारे में

AI सिस्टम्स के पास समाज को कई लाभ देने की क्षमता होती है। किंतु AI के अवसरों को पूरी तरह से हासिल करने के लिए इसे एक सुरक्षित और जिम्मेदार तरीके से डेवलप, डिप्लॉय तथा ऑपरेट किया जाना चाहिए।

AI सिस्टम्स सिक्योरिटी से संबंधित नई कमजोरियों से प्रभावित हो सकते हैं, जिनके बारे में मानक साइबर सिक्योरिटी खतरों के साथ-साथ विचार किए जाने की आवश्यकता है। जब डेवलपमेंट की गति तेज होती है - जैसा कि AI के मामले में है - सिक्योरिटी अक्सर एक द्वितीयक विचार हो सकती है। सिक्योरिटी एक प्रमुख आवश्यकता होनी चाहिए, न केवल डेवलपमेंट के चरण में, बल्कि सिस्टम के पूरे लाइफ साइकल में।

इस कारण से AI सिस्टम डेवलपमेंट लाइफ साइकल के अंदर दिशा-निर्देशों को चार प्रमुख क्षेत्रों में विभाजित किया गया है: **सिक्योर डिज़ाइन**, **सिक्योर डेवलपमेंट**, **सिक्योर डिप्लॉयमेंट**, और **सिक्योर ऑपरेशन एंड मेन्टनेंस**। हम प्रत्येक अनुभाग के लिए विचार और खतरे कम करने के सुझाव देते हैं, जो संगठनात्मक AI सिस्टम डेवलपमेंट प्रक्रिया के लिए समग्र खतरे को कम करने में सहायता करेंगे।

1. **सिक्योर डिज़ाइन**
   इस खंड में AI सिस्टम डेवलपमेंट लाइफ साइकल के डिज़ाइन चरण में लागू होने वाले दिशा-निर्देश दिए गए हैं। इसमें जोखिम और खतरे की मॉडलिंग को समझना, और साथ ही सिस्टम व मॉडल के डिज़ाइन पर विचार करने के लिए विशिष्ट विषयों और ट्रेड-ऑफ़्स को सम्मिलित करना शामिल है।

2. **सिक्योर डेवलपमेंट**
   इस खंड में शामिल दिशा-निर्देश AI सिस्टम डेवलपमेंट लाइफ साइकल के डेवलपमेंट चरण पर लागू होते हैं, जिसमें आपूर्ति श्रृंखला सिक्योरिटी, दस्तावेजीकरण, और एसेट व टेक्निकल डेट मैनेजमेंट शामिल है।

3. **सिक्योर डिप्लॉयमेंट**
   इस खंड में शामिल दिशा-निर्देश AI सिस्टम डेवलपमेंट लाइफ साइकल के डिप्लॉयमेंट चरण पर लागू होते हैं, जिसमें बुनियादी ढांचे और मॉडल को भेदन, खतरे या नुकसान से बचाना, प्रकरण प्रबंधन प्रक्रियाओं को विकसित करना और जिम्मेदारीपूर्वक रिलीज़ करना शामिल है।

4. **सिक्योर ऑपरेशन एंड मेन्टनेंस**
   इस खंड में शामिल दिशा-निर्देश AI सिस्टम डेवलपमेंट लाइफ साइकल के सिक्योर ऑपरेशन एंड मेन्टनेंस चरण पर लागू होते हैं। यह एक बार सिस्टम डिप्लॉय कर दिए जाने के बाद विशेष रूप से प्रासंगिक कार्यों पर दिशा-निर्देश प्रदान करता है, जिसमें लॉगिंग और मॉनिटरिंग, अपडेट मैनेजमेंट और इन्फोर्मेशन शेयरिंग शामिल है।

दिशा-निर्देश 'सिक्योर बाई डिफ़ॉल्ट' एप्रोच का पालन करते हैं, और ये NCSC के सिक्योर सॉफ़्टवेयर डेवलपमेंट एंड डिप्लॉयमेंट गाइडेंस, NIST के सिक्योर सॉफ़्टवेयर डेवलपमेंट फ्रेमवर्क, और CISA, NCSC तथा अंतर्राष्ट्रीय साइबर एजेंसियों द्वारा प्रकाशित 'सिक्योर बाई डिज़ाइन प्रिंसिपल्स' में परिभाषित कार्यप्रथाओं के साथ निकटता से संरेखित हैं। ये इन बातों को प्राथमिकता देते हैं:

- सेवार्थियों के लिए सिक्योरिटी परिणामों का स्वामित्व ग्रहण करना
- मौलिक पारदर्शिता और जवाबदेयता अपनाना
- इस प्रकार से संगठनात्मक संरचना और नेतृत्व का निर्माण करना ताकि सिक्योर बाई डिज़ाइन शीर्ष व्यावसायिक प्राथमिकता हो

परिचय

आर्टिफिशियल इंटेलिजेंस (AI) सिस्टम के पास समाज को अनेक लाभ देने की क्षमता होती है। किंतु AI के अवसरों को पूरी तरह से हासिल करने के लिए इसे एक सुरक्षित और जिम्मेदार तरीके से डेवलप, डिप्लॉय तथा ऑपरेट किया जाना चाहिए। साइबर सिक्योरिटी AI सिस्टम्स की सिक्योरिटी, अनुकूलनीयता, गोपनीयता, निष्पक्षता, प्रभाविता और विश्वसनीयता के लिए एक आनिवार्य पूर्व-शर्त है।

किंतु AI सिस्टम्स सिक्योरिटी से संबंधित नई कमजोरियों से प्रभावित हो सकते हैं, जिनके बारे में मानक साइबर सिक्योरिटी खतरों के साथ-साथ विचार किए जाने की आवश्यकता है। जब डेवलपमेंट की गति तेज होती है - जैसा कि AI के मामले में है - तो सिक्योरिटी अक्सर एक द्वितीयक विचार हो सकती है। सिक्योरिटी एक प्रमुख आवश्यकता होनी चाहिए, न केवल डेवलपमेंट के चरण में, बल्कि सिस्टम के पूरे लाइफ साइकल में।

यह दस्तावेज AI का उपयोग करने वाले किसी भी सिस्टम के प्रोवाइडर्स[1] के लिए दिशा-निर्देशों की सलाह देता है, चाहे वे सिस्टम बिल्कुल नए बनाए गए हों या दूसरों द्वारा प्रदान किए गए उपकरणों और सेवाओं पर आगे निर्मित किए गए हों। इन दिशा-निर्देशों के लागूकरण से प्रोवाइडर्स को ऐसे AI सिस्टम्स बनाने में सहायता मिलेगी, जो इरादे के अनुरूप कार्य करते हैं, आवश्यकता पड़ने पर उपलब्ध रहते हैं, और अनधिकृत पक्षों के समक्ष संवेदनशील डेटा प्रकट किए बिना कार्य करते हैं।

इन दिशा-निर्देशों के बारे में पूर्वस्थापित साइबर सिक्योरिटी, जोखिम प्रबंधन, और प्रकरण फीडबैक की सर्वोत्तम कार्यप्रथाओं के साथ संयोजित रूप में विचार किया जाना चाहिए। विशेष रूप से, हम प्रोवाइडर्स से यह आग्रह करते हैं कि वे यूएस साइबरसिक्योरिटी एंड इंफ्रास्ट्रक्चर सिक्योरिटी एजेंसी (CISA), यूके नेशनल साइबर सिक्योरिटी सेंटर (NCSC), और हमारे सभी अंतरराष्ट्रीय साझेदारों द्वारा विकसित किए गए 'सिक्योर बाई डिज़ाइन'[2] सिद्धांतों का पालन करें। ये सिद्धांत इन बातों को प्राथमिकता देते हैं:

- सेवार्थियों के लिए सिक्योरिटी परिणामों का स्वामित्व ग्रहण करना
- मौलिक पारदर्शिता और जवाबदेयता अपनाना
- इस प्रकार से संगठनात्मक संरचना और नेतृत्व का निर्माण करना ताकि सिक्योर बाई डिज़ाइन शीर्ष व्यावसायिक प्राथमिकता हो।

'सिक्योर बाई डिज़ाइन' सिद्धांतों का पालन करने के लिए सिस्टम के लाइफ साइकल में महत्वपूर्ण संसाधनों की आवश्यकता होती है। इसका अर्थ है कि डेवलपर्स को उन उपकरणों के **फीचर्स**, **मेकेनिज़्म्स**, और **इंप्लिमेंटेशन** को प्राथमिकता देने में निवेश करने की आवश्यकता है, जो सिस्टम डिज़ाइन की प्रत्येक परत पर और डेवलपमेंट लाइफ साइकल के सभी चरणों में सेवार्थियों की सुरक्षा करते हैं। ऐसा करने से बाद में महंगे रीडिज़ाइन्स की रोकथाम की जा सकेगी, साथ ही निकट अवधि में सेवार्थियों और उनके डेटा का संरक्षण भी किया जा सकेगा।

AI सुरक्षा अलग क्यों है?

इस दस्तावेज में हम विशेष रूप से मशीन लर्निंग (ML) एप्लिकेशन्स के संदर्भ में 'AI' का उपयोग करते हैं[3]। स्कोप में सभी प्रकार की ML आती हैं। हम ML एप्लिकेशन्स की परिभाषा ऐसी एप्लिकेशन्स के रूप में देते हैं, जो:

- मानव द्वारा स्पष्ट रूप से प्रोग्राम किए जाने वाले नियमों के बिना कंप्यूटरों को डेटा में पैटर्न की पहचान करने और संदर्भ लाने की अनुमति देने वाले सॉफ़्टवेयर कॉम्पोनेंट्स को शामिल करती हैं
- सांख्यिकीय तर्क के आधार पर पूर्वानुमान, संस्तुतियाँ या निर्णय उत्पन्न करती हैं

मौजूदा साइबर सिक्योरिटी खतरों के साथ-साथ AI सिस्टम्स नई प्रकार की कमजोरियों से भी प्रभावित हो सकते हैं। 'एडवर्सरियल मशीन लर्निंग' (AML) शब्दों का उपयोग ML कॉम्पोनेंट्स में मौलिक कमजोरियों से फायदा उठाने का वर्णन करने के लिए किया जाता है, जिसमें हार्डवेयर, सॉफ़्टवेयर, वर्कफ़्लो और आपूर्ति श्रृंखलाएँ भी शामिल हैं। AML एटैकर्स को ML सिस्टम्स में अनपेक्षित व्यवहार करने में सक्षम बनाता है, जिसमें शामिल हो सकते हैं:

- मॉडल के वर्गीकरण या रिग्रेशन प्रदर्शन को प्रभावित करना
- यूज़र्स को अनधिकृत कार्य करने की अनुमति देना
- संवेदनशील मॉडल जानकारी को निकालना

इन प्रभावों को हासिल करने के कई तरीके हैं, जैसे लार्ज लर्निंग मॉडल (LLM) डोमेन में प्रॉम्प्ट इंजेक्शन एटैक्स, या जानबूझकर प्रशिक्षण डेटा या यूज़र फीडबैक (जिसे 'डेटा पॉयज़निंग' के रूप में जाना जाता है) को करप्ट बनाना।

इस दस्तावेज को किसे पढ़ना चाहिए?

यह दस्तावेज मुख्य रूप से AI सिस्टम्स प्रोवाइडर्स के प्रति लक्षित है, चाहे वे किसी संगठन द्वारा होस्ट किए गए मॉडल पर आधारित हों या वे एक्स्टर्नल एप्लिकेशन प्रोग्रामिंग इंटरफेसेज़ (APIs) का उपयोग कर रहे हों। हम **सभी** हितधारकों से (जिसमें डेटा वैज्ञानिक, डेवलपर्स, प्रबंधक, निर्णयकर्ता और खतरा उठाने वाले लोग शामिल हैं) आग्रह करते हैं कि वे इन दिशा-निर्देशों को पढ़ें, ताकि उन्हें अपने AI सिस्टम्स के **डिज़ाइन**, **डिप्लॉयमेंट** और **ऑपरेशन** के बारे में सूचित निर्णय लेने में सहायता मिल सके।

यह बताने के बावजूद भी, सभी दिशा-निर्देश सभी संगठनों पर सीधे लागू नहीं होंगे। परिष्कार का स्तर और एटैक के तरीके AI सिस्टम्स को लक्षित करने वाले विरोधी के आधार पर अलग-अलग होंगे, इसलिए इन दिशा-निर्देशों के बारे में आपके संगठन के यूज़ केसेज़ और खतरा प्रोफाइल के साथ संयोजन में विचार किया जाना चाहिए।

सुरक्षित AI विकसित करने के लिए कौन जिम्मेदार है?

आधुनिक AI आपूर्ति श्रृंखलाओं में अक्सर कई एक्टर्स शामिल होते हैं। एक आसान एप्रोच के तहत दो निकायों को मान्यता दी जाती है:

- 'प्रोवाइडर' जो डेटा क्यूरेशन, एल्गोरिथम डेवलपमेंट, डिज़ाइन, डिप्लॉयमेंट और मेन्टनेंस के लिए जिम्मेदार है
- 'यूज़र', जो इनपुट प्रदान करता है और आउटपुट प्राप्त करता है

इस प्रोवाइडर-यूज़र एप्रोच का उपयोग कई एप्लिकेशन्स में किया जाता है, लेकिन यह तेजी से असामान्य बनता जा रहा है[4], क्योंकि प्रोवाइडर्स अपने खुद के सिस्टम्स में तृतीय पक्षों द्वारा उपलब्ध कराए गए सॉफ़्टवेयर, डेटा, मॉडल्स और/या रिमोट सर्विसेज़ को शामिल करने के इच्छुक हो सकते हैं। ये जटिल आपूर्ति श्रृंखलाएँ एंड यूज़र्स के लिए यह समझना कठिन बनाती हैं कि सिक्योर AI की जिम्मेदारी कहाँ स्थित है।

यूज़र्स के पास (चाहे वे 'एंड यूज़र्स', या एक्स्टर्नल AI कॉम्पोनेंट को शामिल करने वाले प्रोवाइडर्स[5] हों) आम-तौर पर उनके द्वारा उपयोग किए जा रहे सिस्टम्स से जुड़े खतरों को पूरी तरह से समझने, उनका आकलन करने या उन्हें संबोधित करने के लिए पर्याप्त दृश्यता और/या विशेषज्ञता नहीं होती है। इसलिए, 'सिक्योर बाई डिज़ाइन' सिद्धांतों के अनुरूप, **AI कॉम्पोनेंट्स के प्रोवाइडर्स को आपूर्ति श्रृंखला में नीचे आने वाले यूज़र्स के सिक्योरिटी परिणामों की जिम्मेदारी लेनी चाहिए।**

प्रोवाइडर्स को यथासंभव अपने मॉडल्स, पाइपलाइन्स और/या सिस्टम्स के अंदर सिक्योरिटी नियंत्रण और मिटिगेशन्स लागू करने चाहिए, और जहाँ सेटिंग्स का उपयोग किया जाता है, वहाँ डिफ़ॉल्ट के रूप में सबसे सिक्योर विकल्प लागू करने चाहिए। जहाँ जोखिम कम नहीं किए जा सकते हैं, वहाँ प्रोवाइडर को इन बातों के लिए जिम्मेदारी लेनी चाहिए:

- आपूर्ति श्रृंखला में नीचे आने वाले यूज़र्स को उन जोखिमों की के बारे में सूचित करना, जिन्हें वे खुद और (यदि लागू हो) उनके अपने यूज़र्स उठा रहे हैं
- उन्हें सलाह देना कि कॉम्पोनेंट का उपयोग सिक्योर रूप से कैसे किया जाए

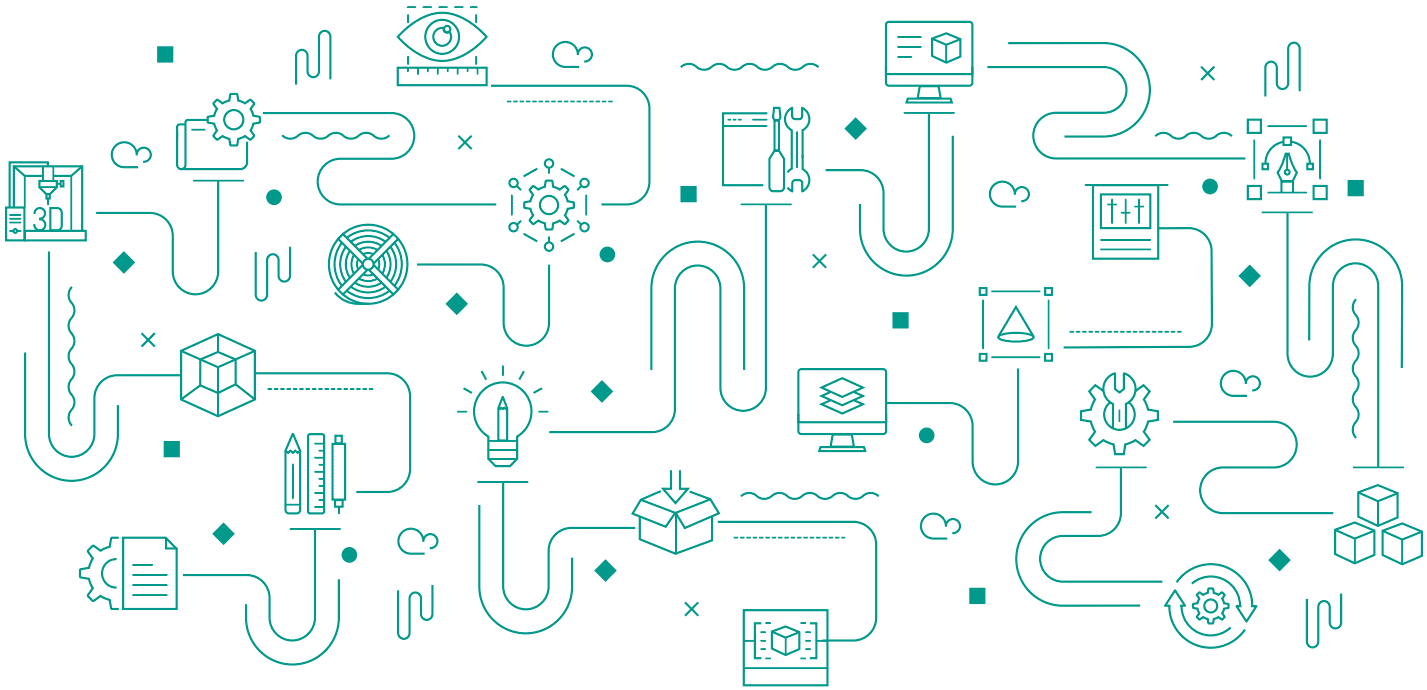
यदि सिस्टम में भेदन मूर्त या व्यापक भौतिक या प्रतिष्ठात्मक क्षति का कारण बन सकता है, व्यावसायिक ऑपरेशन को गंभीर नुकसान पहुंचा सकता है, संवेदनशील या गोपनीय जानकारी और/या कानूनी निहितार्थों को प्रकट कर सकता है, तो AI साइबर सिक्योरिटी जोखिमों को **अतिमहत्वपूर्ण** माना जाना चाहिए।

सुरक्षित AI सिस्टम विकसित करने के लिए दिशा-निर्देश

दिशा-निर्देशों को AI सिस्टम्स डेवलपमेंट लाइफ साइकल के अंदर चार प्रमुख क्षेत्रों में विभाजित किया गया है: **सिक्योर डिज़ाइन**, **सिक्योर डेवलपमेंट**, **सिक्योर डिप्लॉयमेंट**, और **सिक्योर ऑपरेशन एंड मेन्टनेंस**। हम प्रत्येक क्षेत्र के लिए विचारों और मिटिगेशन्स के सुझाव देते हैं, जो संगठनात्मक AI सिस्टम्स डेवलपमेंट प्रक्रिया के लिए समग्र खतरे को कम करने में सहायता देंगे।

इस दस्तावेज में दिए गए दिशा-निर्देश निम्नलिखित दस्तावेजों में परिभाषित सॉफ़्टवेयर डेवलपमेंट लाइफ साइकल प्रथाओं के साथ निकटता से संरेखित हैं:

- NCSC का सिक्योर डेवलपमेंट एंड डिप्लॉयमेंट गाइडेंस
- राष्ट्रीय मानक और प्रौद्योगिकी इंस्टिट्यूट (NIST) का सिक्योर सॉफ़्टवेयर डेवलपमेंट फ्रेमवर्क (SSDF)[6]



1. सिक्योर डिज़ाइन

इस खंड में AI सिस्टम डेवलपमेंट लाइफ साइकल के **डिज़ाइन** चरण में लागू होने वाले दिशा-निर्देश दिए गए हैं। इसमें जोखिम और खतरे की मॉडलिंग को समझना, और साथ ही सिस्टम व मॉडल के डिज़ाइन पर विचार करने के लिए विशिष्ट विषयों और ट्रेड-ऑफ़्स को सम्मिलित करना शामिल है।

खतरों और जोखिमों के बारे में कर्मचारियों की जागरूकता बढ़ाएँ

सिस्टम के मालिक और वरिष्ठ अग्रणी सुरक्षित AI के लिए खतरों और उनके मिटिगेशन्स को समझते हैं। आपके डेटा वैज्ञानिक और डेवलपर्स प्रासंगिक सिक्योरिटी खतरों और फेलियर मोड्स के बारे में जागरूकता बनाए रखते हैं और जोखिम मालिकों को सूचित निर्णय लेने में सहायता देते हैं। आप यूज़र्स को AI सिस्टम्स के सामने आने वाले अनन्य सिक्योरिटी जोखिमों पर मार्गदर्शन प्रदान करते/करती हैं (उदाहरण के लिए, मानक InfoSec प्रशिक्षण के हिस्से के रूप में) और डेवलपर्स को सिक्योर कोडिंग तकनीकों एवं सिक्योर व जिम्मेदारीपूर्ण AI कार्यप्रथाओं में प्रशिक्षित करते/करती हैं।

अपने सिस्टम के लिए खतरों का मॉडल बनाएँ

आप अपनी जोखिम प्रबंधन प्रक्रिया के हिस्से के रूप में अपने सिस्टम के प्रति खतरों का आकलन करने के लिए एक समग्र प्रक्रिया लागू करते/करती हैं, जिसमें किसी AI कॉम्पोनेंट में भेदन होने या इसके द्वारा अप्रत्याशित व्यवहार किए जाने के कारण सिस्टम, यूज़र्स, संगठनों और व्यापक समाज के लिए संभावित प्रभावों को समझना भी शामिल है[7]। इस प्रक्रिया में AI-विशिष्ट खतरों के प्रभाव का आकलन करना और आपके निर्णय लेने का दस्तावेजीकरण करना शामिल है[8]।

आप इस बात की पहचान करते/करती हैं कि आपके सिस्टम में उपयोग किए जाने वाले डेटा की संवेदनशीलता और इनके प्रकार किसी एटैकर के लक्ष्य के रूप में इसके मूल्य को प्रभावित कर सकते हैं। आपके आकलन में इस बारे में विचार किया जाना चाहिए कि जैसे-जैसे AI सिस्टम्स तेजी से उच्च मूल्य लक्ष्यों के रूप में देखे जाते हैं और जैसे-जैसे AI स्वयं नए, ऑटोमेटेड एटैक वेक्टर्स को सक्षम बनाता है, वैसे-वैसे कुछ खतरे बढ़ सकते हैं।

अपने सिस्टम को सिक्योरिटी के साथ-साथ कार्यक्षमता और प्रदर्शन के लिए भी डिज़ाइन करें

आप इस बारे में आश्वस्त हैं कि आपके हाथ में जो काम है, उसे AI का उपयोग करके सबसे समुचित रूप से संबोधित किया जाता है। यह तय कर लेने के बाद आप अपने AI-विशिष्ट डिज़ाइन विकल्पों की उपयुक्तता का आकलन करते/करती हैं। आप अन्य विचारों के साथ-साथ कार्यात्मकता, यूज़र अनुभव, डिप्लॉयमेंट परिवेश, प्रदर्शन, आश्वासन, निरीक्षण, नैतिक और कानूनी आवश्यकताओं के बारे में अपने खतरे के मॉडल और संबंधित सिक्योरिटी मिटिगेशन्स पर विचार करते/करती हैं। उदाहरण के लिए:

- आप खुद विकसित करते समय या एक्स्टर्नल कॉम्पोनेंट्स के उपयोग का चयन करते समय आपूर्ति श्रृंखला सिक्योरिटी पर विचार करते/करती हैं, जैसे:
    - नए मॉडल को प्रशिक्षित करने, मौजूदा मॉडल का उपयोग करने (फाइन-ट्यूनिंग के साथ या इसके बिना) या एक्स्टर्नल API के माध्यम से मॉडल को एक्सेस करने का आपका चयनित विकल्प आपकी आवश्यकताओं के अनुरूप है
    - एक्स्टर्नल मॉडल प्रोवाइडर के साथ काम करने के आपके चयनित विकल्प में उस प्रोवाइडर के अपने सिक्योरिटी पोस्चर का उचित डिलिजेंस आकलन शामिल है
    - यदि आप किसी एक्स्टर्नल लाइब्रेरी का उपयोग कर रहे/रही हैं, तो आप एक समुचित डिलिजेंस आकलन पूरा करते/करती हैं (उदाहरण के लिए, लाइब्रेरी में ऐसे नियंत्रण सुनिश्चित करने के लिए, जो मनमाने कोड एक्ज़िक्यूशन के कारण सिस्टम द्वारा अविश्वसनीय मॉडल्स को तुरंत प्रकट किए बिना उन्हें लोड करने की रोकथाम करते हैं[9])
    - आप तृतीय-पक्ष मॉडल या सीरियलाइज़्ड वेट्स की इंपोर्टिंग के समय स्कैनिंग और आइसोलेशन/सैंडबॉक्सिंग को इंप्लिमेंट करते/करती हैं, जिसे अविश्वसनीय तृतीय-पक्ष कोड के रूप में माना जाना चाहिए और जो रिमोट कोड एक्ज़िक्यूशन करने में सक्षम हो सकता है

- यदि आप किसी एक्स्टर्नल API का उपयोग कर रहे/रही हैं, तो आपको ऐसे डेटा पर उचित नियंत्रण लागू करने होंगे जिन्हें आपके संगठन के नियंत्रण से बाहर की सेवाओं के पास भेजा जा सकता है, जैसे कि यूज़र्स को संभावित रूप से संवेदनशील जानकारी भेजने से पहले लॉग इन करने और पुष्टि करने की आवश्यकता
- आप डेटा और इनपुट के लिए समुचित परीक्षण और संशोधन लागू करते/करती हैं; इसमें आपके मॉडल में यूज़र फीडबैक या निरंतर रूप से सीखने के लिए डेटा को सम्मिलित करते समय ऐसा करना शामिल है, जबकि इस बात की पहचान की जाती है कि प्रशिक्षण डेटा सिस्टम व्यवहार को परिभाषित करता है

- आप मौजूदा सिक्योर डेवलपमेंट और ऑपरेशन्स सर्वोत्तम कार्यप्रथाओं में AI सॉफ़्टवेयर सिस्टम डेवलपमेंट को एकीकृत करते/करती हैं; AI सिस्टम के सभी तत्व कोडिंग प्रथाओं और भाषाओं का उपयोग करके उपयुक्त परिवेश में लिखे जाते हैं, जो यथासंभव रूप से कमजोरियों के ज्ञात वर्गों का न्यूनन या उन्मूलन करते हैं

- यदि AI कॉम्पोनेंट्स को प्रक्रियाएँ ट्रिगर करने की आवश्यकता है, उदाहरण के लिए फ़ाइल्स का संशोधन करना या आउटपुट को एक्स्टर्नल सिस्टम्स के प्रति निर्देशित करना, तो आप संभावित कार्यों के लिए उचित प्रतिबंध लागू करते/करती हैं (यदि आवश्यक हो तो इसमें एक्स्टर्नल AI और non-AI फेल-सेफ़्स शामिल हैं)

- यूज़र इंटरएक्शन के बारे में निर्णय AI-विशिष्ट जोखिमों द्वारा सूचित होते हैं, उदाहरण के लिए:

  - आपका सिस्टम संभावित एटैकर को विवरण का अनावश्यक स्तर प्रकट किए बिना यूज़र्स को उपयोग करने योग्य आउटपुट प्रदान करता है
  - यदि आवश्यक हो, तो आपका सिस्टम मॉडल आउटपुट के आसपास प्रभावी सुरक्षा प्रदान करता है
  - यदि बाहरी सेवार्थियों या सहयोगियों को API प्रस्तावित किया जाता है, तो आप API के माध्यम से AI सिस्टम पर एटैक्स को कम करने वाले उपयुक्त नियंत्रण लागू करते/करती हैं
  - आप डिफ़ॉल्ट रूप से सिस्टम में सर्वाधिक सिक्योर सेटिंग्स का एकीकरण करते/करती हैं
  - आप किसी सिस्टम की कार्यात्मकता की एक्सेस को सीमित करने के लिए न्यूनतम विशेषाधिकार सिद्धांत लागू करते/करती हैं
  - आप यूज़र्स को अधिक जोखिम वाली क्षमताएँ समझाते/समझाती हैं और यूज़र्स के लिए इनके उपयोग का चयन करने की आवश्यकता नियत करते/करती है; आप निषिद्ध यूज़ केसेज़ संचरित करते/करती हैं, और, जहाँ संभव हो, यूज़र्स को वैकल्पिक समाधानों के बारे में सूचित करते/करती हैं

अपने AI मॉडल का चयन करते समय सिक्योरिटी से संबंधित लाभों और ट्रेड-ऑफ़्स के बारे में विचार करें

आपके AI मॉडल के चयन में अनेकानेक आवश्यकताओं को संतुलित किया जाना शामिल होगा। इसमें मॉडल आर्किटेक्चर, कॉन्फ़िगरेशन, प्रशिक्षण डेटा, प्रशिक्षण एल्गोरिदम और हाइपरपैरामीटर्स के विकल्प शामिल हैं। आपके निर्णय आपके खतरे के मॉडल द्वारा सूचित होते हैं, और AI सिक्योरिटी अनुसंधान में प्रगति और खतरे की समझ विकसित होने के साथ-साथ इनका नियमित रूप से पुनराकलन किया जाता है।

AI मॉडल का चयन करते समय आपके विचारों में संभावित रूप से निम्नलिखित शामिल होंगे, लेकिन ये केवल इन्हीं तक सीमित नहीं हैं:

- आपके द्वारा उपयोग किए जा रहे मॉडल की जटिलता, यानि चयनित आर्किटेक्चर और पैरामीटर्स की संख्या; अन्य कारकों के साथ-साथ आपके मॉडल के चुने हुए आर्किटेक्चर और पैरामीटर्स की संख्या इस बात को प्रभावित करेगी कि आपके मॉडल को कितने प्रशिक्षण डेटा की आवश्यकता है और वह उपयोग किए जाने पर इनपुट डेटा में परिवर्तन के प्रति कितना सशक्त है

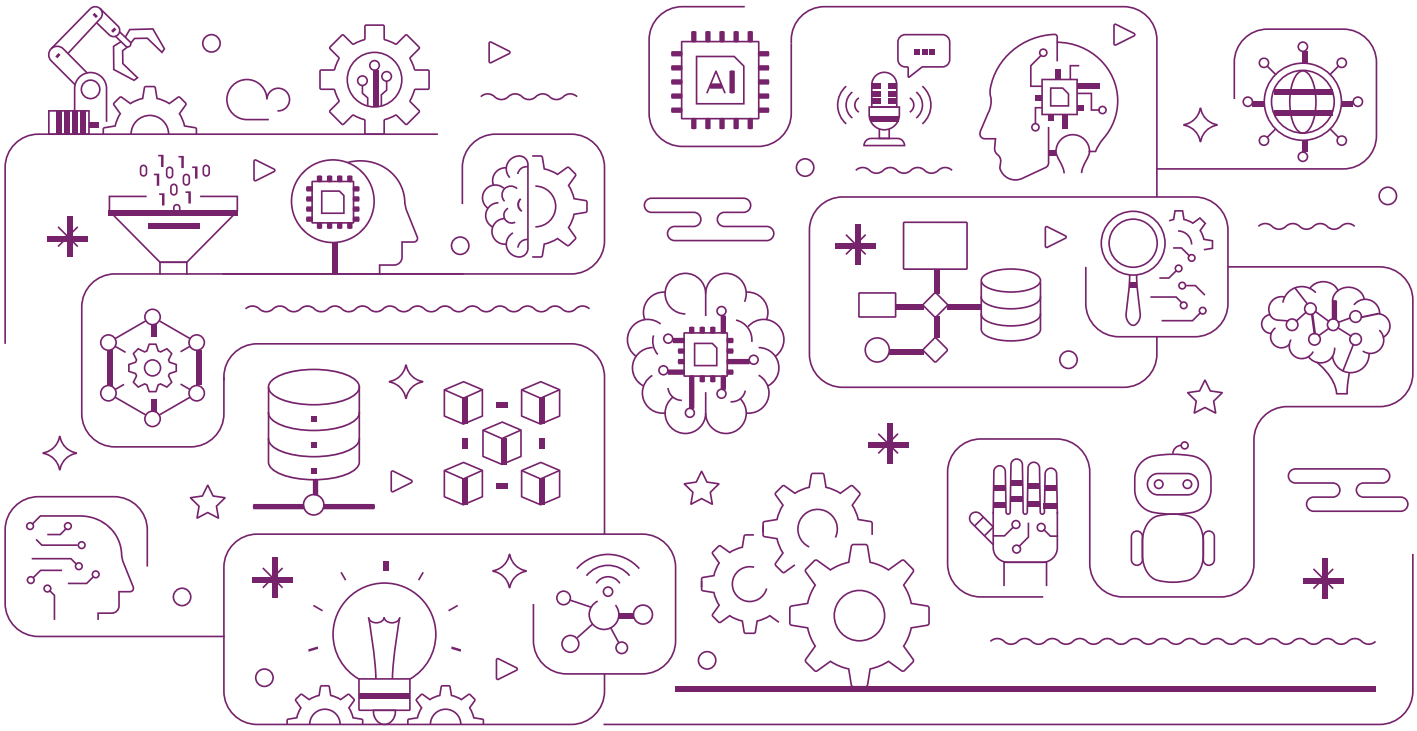
- आपके यूज़ केस के लिए मॉडल की उपयुक्तता और/या इसे आपकी विशिष्ट आवश्यकता के अनुकूल बनाने की व्यवहार्यता (उदाहरण के लिए फाइन-ट्यूनिंग के माध्यम से)

- अपने मॉडल के आउटपुट को संरेखित करने, उसका विश्लेषण करने और उसे समझाने की क्षमता (उदाहरण के लिए डिबगिंग, ऑडिट या नियामक अनुपालन); विश्लेषण करने में अधिक कठिन, बड़े और जटिल मॉडल्स की तुलना में आसान एवं और अधिक पारदर्शी मॉडल्स का उपयोग करने के लाभ हो सकते हैं

- प्रशिक्षण डेटासेट(टों) की विशेषताएँ, जिसमें आकार, अखंडता, गुणवत्ता, संवेदनशीलता, आयु, प्रासंगिकता और विविधता शामिल है

- मॉडल हार्डनिंग (जैसे एडवर्सीरियल प्रशिक्षण), नियमितीकरण और/या गोपनीयता को संवृद्ध करने वाली तकनीकों के उपयोग का मूल्य
- कॉम्पोनेंट्स की सिद्धता और आपूर्ति श्रृंखलाएँ, जिसमें मॉडल या फाउंडेशन मॉडल, प्रशिक्षण डेटा और संबंधित उपकरण शामिल हैं

इनमें से कितने कारक सुरक्षा परिणामों को प्रभावित करते हैं, इस बारे में और अधिक जानकारी के लिए NCSC का 'मशीन लर्निंग की सिक्योरिटी के लिए सिद्धांत' देखें, विशेषकर डिज़ाइन फ़ॉर सिक्योरिटी (मॉडल आर्किटेक्चर)।



2. सिक्योर डेवलपमेंट

इस खंड में शामिल दिशा-निर्देश AI सिस्टम डेवलपमेंट लाइफ साइकल के **डेवलपमेंट** चरण पर लागू होते हैं, जिसमें आपूर्ति श्रृंखला सिक्योरिटी, दस्तावेजीकरण, और एसेट व टेक्निकल डेट मैनेजमेंट शामिल है।

अपनी आपूर्ति श्रृंखला को सुरक्षित बनाएँ

आप सिस्टम के लाइफ साइकल में अपनी AI आपूर्ति श्रृंखलाओं की सुरक्षा का आकलन और निगरानी करते/करती हैं, और आपूर्तिकर्ताओं के लिए उन्हीं मानकों का पालन करने की आवश्यकता निर्धारित करते/करती हैं जो आपका अपना संगठन अन्य सॉफ़्टवेयर के लिए लागू करता है। यदि आपूर्तिकर्ता आपके संगठन के मानकों का पालन नहीं कर सकते हैं, तो आप अपनी मौजूदा जोखिम प्रबंधन नीतियों के अनुसार कार्य करते/करती हैं।

यदि हार्डवेयर और सॉफ़्टवेयर कॉम्पोनेंट्स का इन-हाउस उत्पादन नहीं किया जाता है, तो आप अपने सिस्टम में मजबूत सुरक्षा सुनिश्चित करने के लिए सत्यापित किए गए कमर्शियल, ओपन सोर्स और अन्य तृतीय-पक्ष डेवलपर्स की ओर से अच्छी तरह से सुरक्षित और अच्छी तरह से प्रलेखित हार्डवेयर और सॉफ़्टवेयर कॉम्पोनेंट्स (उदाहरण के लिए, मॉडल्स, डेटा, सॉफ़्टवेयर लाइब्रेरीज़, मॉड्यूल्स, मिडलवेयर, फ्रेमवर्क्स और एक्स्टर्नल APIs) प्राप्त करते/करती हैं और उनका रख-रखाव करते/करती हैं।

यदि सुरक्षा मानदंड पूरे नहीं होते हैं, तो आप मिशन के लिए महत्वपूर्ण सिस्टम्स के वैकल्पिक समाधानों में फेलओवर के लिए तैयार रहते/रहती हैं। आप आपूर्ति श्रृंखला और सॉफ़्टवेयर डेवलपमेंट लाइफ साइकल्स के एटेस्टेशन्स को ट्रैक करने के लिए संसाधनों का उपयोग करते/करती हैं, जैसे NCSC का आपूर्ति श्रृंखला मार्गदर्शन और सॉफ़्टवेयर आर्टिफैक्ट्स के लिए सप्लाई चेन लेवल्स (SLSA)[10] जैसे फ्रेमवर्क्स।

अपने एसेट्स की पहचान करें, उन्हें ट्रैक करें और उनकी सुरक्षा करें

आप अपने संगठन के लिए अपने AI-संबंधी एसेट्स का मूल्य समझते/समझती हैं, और इस बात की पहचान करते/करती हैं कि वे कहाँ महत्वपूर्ण निवेश का प्रतिनिधित्व करते हैं और कहाँ पर उनकी एक्सेस एटैकर को सक्षम बनाती है, जिसमें मॉडल्स, डेटा (यूज़र फीडबैक सहित), प्रॉम्प्ट्स, सॉफ़्टवेयर, दस्तावेजीकरण, लॉग्स और एसेसमेंट्स भी (संभावित असुरक्षित क्षमताओं और फेलियर मोड्स के बारे में जानकारी सहित) शामिल हैं। आप लॉग्स को संवेदनशील डेटा के रूप में मानते/मानती हैं और उनकी गोपनीयता, अखंडता और उपलब्धता के संरक्षण के लिए नियंत्रण लागू करते/करती हैं।

आप जानते/जानती हैं कि आपके एसेट्स कहाँ मौजूद हैं और आपने सभी संबंधित खतरों का आकलन किया है और उन्हें स्वीकार कर लिया है। आपके पास अपने एसेट्स को ट्रैक, ऑथेंटिकेट, वर्ज़न कंट्रोल और सिक्योर करने के लिए प्रक्रियाएँ और टूल्स उपलब्ध हैं, और भेदन की स्थिति में आप एक ज्ञात अच्छी स्टेट को रिस्टोर कर सकते/सकती हैं।

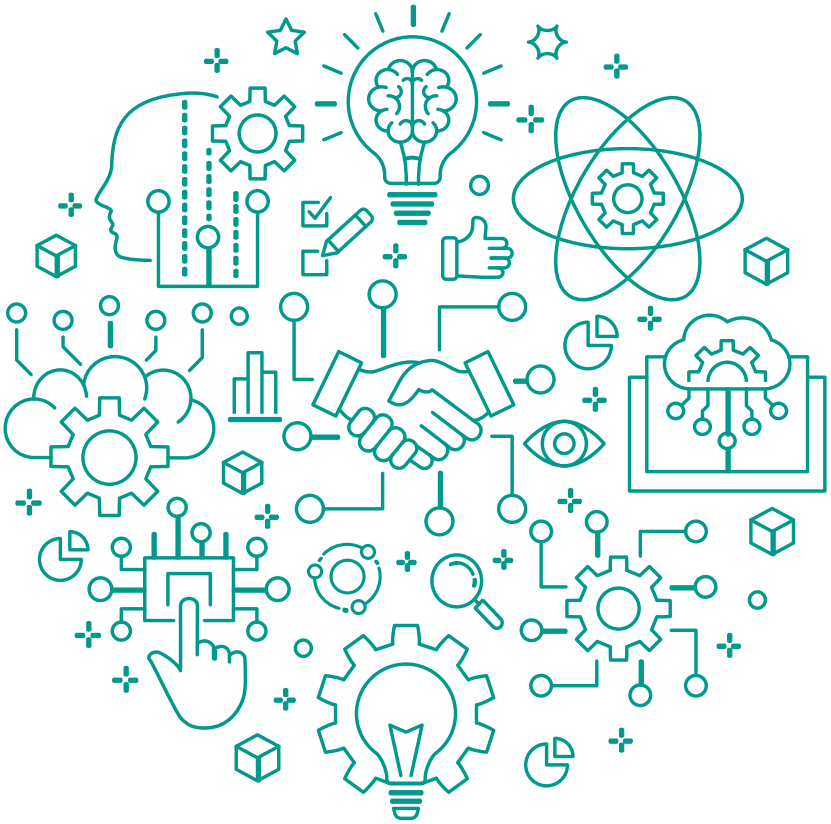
AI सिस्टम्स कौन से डेटा को एक्सेस कर सकते हैं, और AI द्वारा जेनरेट किए गए कन्टेंट को उसकी संवेदनशीलता (और इसे जेनरेट करने में प्रयुक्त इनपुट की संवेदनशीलता) के अनुसार प्रबंधित करने के लिए आपके पास प्रक्रियाएँ और नियंत्रण उपलब्ध हैं।

अपने डेटा, मॉडल्स और प्रॉम्प्ट्स का दस्तावेजीकरण करें

आप किसी भी मॉडल, डेटासेट्स और मेटा-अथवा सिस्टम-प्रॉम्प्ट्स के निर्माण, ऑपरेशन और लाइफ साइकल प्रबंधन का दस्तावेजीकरण करते/करती हैं। आपके दस्तावेजीकरण में सुरक्षा से प्रासंगिक जानकारी शामिल है, जैसे प्रशिक्षण डेटा के स्रोत (जिसमें डेटा की फाइन-ट्यूनिंग और मानवीय या अन्य ऑपरेशनल फीडबैक भी शामिल है), इच्छित दायरे और सीमाएं, गार्डरेल्स, क्रिप्टोग्राफिक हैशेज़ या सिग्नेचर्स, रिटेंशन अवधि, सुझावित समीक्षा आवृत्ति और संभावित फेलियर मोड्स। इसमें सहायता देने वाली उपयोगी संरचनाओं में मॉडल कार्ड्स, डेटा कार्ड्स और सॉफ़्टवेयर बिल्स ऑफ मटीरियल्स (SBOMs) शामिल हैं। व्यापक दस्तावेजीकरण का उत्पादन पारदर्शिता और जवाबदेही का समर्थन करता है[11]।

अपने टेक्निकल डेट का प्रबंधन करें

किसी भी अन्य सॉफ़्टवेयर सिस्टम के समान ही आप AI सिस्टम्स के लाइफ साइकल के दौरान अपने 'टेक्निकल डेट' की पहचान, ट्रैकिंग और प्रबंधन करते/करती हैं (टेक्निकल डेट तब होता है, जहाँ ऐसे इंजीनियरिंग निर्णय लिए जाते हैं जो अल्पकालिक परिणाम प्राप्त करने की सर्वोत्तम कार्यप्रथाओं से न्यून होते हैं और जिन्हें दीर्घकालिक लाभों के स्थान पर लिया जाता है)। वित्तीय ऋण के समान ही टेक्निकल डेट स्वत: बुरा नहीं होता है, लेकिन इसका प्रबंधन डेवलपमेंट के सबसे आरंभिक चरणों से ही किया जाना चाहिए[12]। आप इस बात की पहचान करते/करती हैं कि मानक सॉफ़्टवेयर की तुलना में AI के संदर्भ में ऐसा करना अधिक चुनौतीपूर्ण हो सकता है, और यह कि तेज डेवलपमेंट साइकल्स और अच्छी तरह से स्थापित प्रोटोकॉल्स व इंटरफेसेज़ की कमी के कारण आपके टेक्निकल डेट का स्तर अधिक होने की संभावना है। आप यह सुनिश्चित करते/करती हैं कि आपकी लाइफ साइकल योजनाएँ (जिसमें AI सिस्टम्स को डिकमिशन करने की प्रक्रियाएँ भी शामिल हैं) भविष्य में समान प्रणालियों के जोखिमों को एसेस, एकनॉलेज और मिटिगेट करती हैं।



3. सिक्योर डिप्लॉयमेंट

इस खंड में शामिल दिशा-निर्देश AI सिस्टम डेवलपमेंट लाइफ साइकल के **डिप्लॉयमेंट** चरण पर लागू होते हैं, जिसमें बुनियादी ढांचे और मॉडल को भेदन, खतरे या नुकसान से बचाना, प्रकरण प्रबंधन प्रक्रियाओं को विकसित करना और जिम्मेदारीपूर्वक रिलीज़ करना शामिल है।

अपने बुनियादी ढांचे को सिक्योर बनाएँ

आप अपने सिस्टम के लाइफ साइकल के हरेक हिस्से में उपयोग किए जाने वाले बुनियादी ढांचे के लिए अच्छे इंफ्रास्ट्रक्चर सिक्योरिटी सिद्धांत लागू करते/करती हैं। आप अनुसंधान और डेवलपमेंट के साथ-साथ डिप्लॉयमेंट में भी अपने APIs, मॉडल्स और डेटा, और उनके प्रशिक्षण व प्रोसेसिंग पाइपलाइनों के लिए उपयुक्त एक्सेस नियंत्रण लागू करते/करती हैं। इसमें संवेदनशील कोड या डेटा स्टोर करने वाले परिवेशों का समुचित आइसोलेशन शामिल है। इससे मानक साइबर सिक्योरिटी एटैक्स को कम करने में भी सहायता मिलेगी, जिनका उद्देश्य मॉडल को चुराना या उसके प्रदर्शन को नुकसान पहुंचाना होता है।

अपने मॉडल को लगातार रूप से सिक्योर बनाए रखें

मॉडल को प्रत्यक्ष रूप से एक्सेस करके (मॉडल वेट्स प्राप्त करके) या अप्रत्यक्ष रूप से एक्सेस करके (किसी एप्लिकेशन या सेवा के माध्यम से मॉडल की क्वेरी करके) एटैकर्स मॉडल की कार्यक्षमता[13] को या उस डेटा को, जिसके लिए उसे प्रशिक्षित किया गया था[14], फिर से निर्मित करने में सक्षम हो सकते हैं। एटैकर्स प्रशिक्षण के दौरान या उसके बाद मॉडल, डेटा या संकेतों के साथ छेड़छाड़ भी कर सकते हैं, जिससे आउटपुट अविश्वसनीय हो जाता है।

आप निम्नलिखित माध्यमों से मॉडल और डेटा को क्रमशः प्रत्यक्ष और अप्रत्यक्ष एक्सेस से बचा सकते/सकती हैं:

- मानक साइबर सिक्योरिटी की सर्वोत्तम प्रथाएँ लागू करके
- गोपनीय जानकारी को एक्सेस करने, संशोधित करने और उसे बाहर निकालने के प्रयासों का पता लगाने और उसे रोकने के लिए क्वेरी इंटरफ़ेस पर नियंत्रण लागू करके

उपभोग करने वाले सिस्टम्स द्वारा मॉडल्स का वैलिडेटशन सुनिश्चित करने के लिए, आप मॉडल के प्रशिक्षित होते ही मॉडल फ़ाइल्स (उदाहरण के लिए, मॉडल वेट्स) और डेटासेट्स के (जिसमें चेकप्वॉइंट्स भी शामिल हैं) क्रिप्टोग्राफिक हैशेज़ और/या सिग्नेचर्स को कंप्यूट और शेयर करते/करती हैं। बेहतरीन कुंजी प्रबंधन अनिवार्य है, जैसाकि क्रिप्टोग्राफी के साथ हमेशा ही होता है[15]।

गोपनीयता जोखिम कम करने के लिए आपकी एप्रोच यूज़ केस और खतरे के मॉडल पर काफी निर्भर करेगी। कुछ एप्लिकेशन्स के लिए, उदाहरण के लिए जिनमें अत्यधिक संवेदनशील डेटा शामिल होता है, सैद्धांतिक गारंटी की आवश्यकता हो सकती है, जिन्हें लागू करना कठिन या महंगा हो सकता है। यदि उचित हो, तो गोपनीयता में वृद्धि करने वाली प्रौद्योगिकियों (जैसे डिफरेंशियल प्राइवेसी या होमोमॉर्फिक एन्क्रिप्शन) का उपयोग उपभोक्ताओं, यूज़र्स और एटैकर्स के साथ जुड़े जोखिम के स्तर का पता लगाने या आश्वासन के लिए किया जा सकता है।

प्रकरण प्रबंधन प्रक्रियाएँ विकसित करें

आपके AI सिस्टम्स को प्रभावित करने वाले सिक्योरिटी प्रकरणों की अनिवार्यता आपकी प्रकरण प्रतिक्रिया, आगे उठाने की और समाधान योजनाओं में परिलक्षित होती है। जैसे-जैसे सिस्टम और व्यापक शोध का विकास होता है, आपकी योजनाएँ अलग-अलग परिदृश्यों को दर्शाती हैं और नियमित रूप से पुनराकलित की जाती हैं। आप कंपनी के महत्वपूर्ण डिजिटल संसाधनों को ऑफ़लाइन बैकअप में स्टोर करते/करती हैं। उत्तरदाताओं को AI-संबंधी प्रकरणों का आकलन करने और उन्हें संबोधित करने के लिए प्रशिक्षित किया गया है। आप सेवार्थियों और यूज़र्स को बिना किसी अतिरिक्त शुल्क के उच्च-गुणवत्ता वाले ऑडिट लॉग्स और अन्य सिक्योरिटी सुविधाएँ या जानकारी प्रदान करते/करती हैं, ताकि उनकी प्रकरण फीडबैक प्रक्रियाओं को सक्षम बनाया जा सके।

AI को जिम्मेदारीपूर्वक रिलीज़ करें

आप मॉडल्स, एप्लिकेशन्स या सिस्टम्स का समुचित और प्रभावी सिक्योरिटी आकलन करने के बाद ही उन्हें रिलीज़ करते/करती हैं, उदाहरण के लिए बेंचमार्किंग और रेड टीमिंग (और साथ ही इन दिशा-निर्देशों के दायरे से बाहर अन्य परीक्षण, जैसे सुरक्षा या निष्पक्षता) और आप अपने यूज़र्स के साथ ज्ञात सीमाओं या संभावित फेलियर मोड्स के बारे में स्पष्ट रहते/रहती हैं। इस दस्तावेज के अंत में दिए गए आगे पढ़ें अनुभाग में ओपन-सोर्स सिक्योरिटी परीक्षण टेस्टिंग लाइब्रेरीज़ का विवरण दिया गया है।

यूज़र्स के लिए सही काम करना आसान बनाएँ

आप इस बात की पहचान करते/करती हैं कि प्रत्येक नई सेटिंग या कॉन्फ़िगरेशन विकल्प का आकलन उससे प्राप्त होने वाले व्यावसायिक लाभ और उसके माध्यम से प्रस्तुत होने वाले किसी भी सिक्योरिटी खतरे के साथ किया जाना है। आदर्श रूप से, सबसे सिक्योर सेटिंग को एकमात्र विकल्प के रूप में सिस्टम में एकीकृत किया जाएगा। जब कॉन्फ़िगरेशन आवश्यक हो, तो डिफ़ॉल्ट विकल्प सामान्य खतरों के प्रति व्यापक रूप से सिक्योर होना चाहिए (यानी, डिफ़ॉल्ट रूप से सिक्योर)। आप मैलिशियस तरीकों से अपने सिस्टम के उपयोग या डिप्लॉयमेंट को रोकने के लिए नियंत्रण लागू करते/करती हैं।

आप यूज़र्स को अपने मॉडल या सिस्टम के समुचित उपयोग पर मार्गदर्शन प्रदान करते/करती हैं, जिसमें सीमाओं और संभावित फेलियर मोड्स को प्रकट करना भी शामिल है। आप यूज़र्स को स्पष्ट रूप से बताते/बताती हैं कि वे सिक्योरिटी के किन पहलुओं के लिए जिम्मेदार हैं, और आप इस बारे में पारदर्शी रहते/रहती हैं कि कहाँ (और कैसे) उनके डेटा को इस्तेमाल, एक्सेस या स्टोर किया जा सकता है (उदाहरण के लिए, यदि इसका उपयोग मॉडल की रीट्रेनिंग के लिए किया जाता है, या कर्मचारियों या साझेदारों द्वारा इसकी समीक्षा की जाती है)।

4. सिक्योर ऑपरेशन एंड मेन्टनेंस

इस खंड में शामिल दिशा-निर्देश AI सिस्टम डेवलपमेंट लाइफ साइकल के **सिक्योर ऑपरेशन एंड मेन्टनेंस** चरण पर लागू होते हैं। यह सिस्टम के डिप्लॉय कर दिए जाने के बाद विशेष रूप से प्रासंगिक कार्यों पर दिशा-निर्देश प्रदान करता है, जिसमें लॉगिंग और मॉनिटरिंग, अपडेट मैनेजमेंट और इन्फोर्मेशन शेयरिंग शामिल है।

अपने सिस्टम के व्यवहार की निगरानी करें 

आप अपने मॉडल और सिस्टम के आउटपुट्स व प्रदर्शन का मापन करते/करती हैं, ताकि आप सुरक्षा को प्रभावित करने वाले व्यवहार में अचानक और क्रमिक परिवर्तन देख सकें। आप संभावित अनधिकृत प्रवेशों व भेदन, और साथ ही प्राकृतिक डेटा बहाव के लिए जिम्मेदारी और पहचान स्थापित कर सकते/सकती हैं।

अपने सिस्टम के इनपुट की निगरानी करें 

गोपनीयता और डेटा सिक्योरिटी आवश्यकताओं के अनुरूप, आप भेदन या दुरुपयोग की परिस्थिति में अनुपालन दायित्वों, ऑडिट, जांच और समाधान को सक्षम बनाने के लिए अपने सिस्टम के इनपुट्स (जैसे इन्फ़रेंस रिक्वेस्ट्स, क्वेरीज़ या प्रॉम्प्टस) की निगरानी करते/करती हैं और उनके लॉग बनाते/बनाती हैं। इसमें आउट-ऑफ़-डिस्ट्रिब्युशन और/या एडवर्सरियल इनपुट का स्पष्ट रूप से पता लगाना शामिल हो सकता है, जिसमें वे भी शामिल हैं जिनका उद्देश्य डेटा को तैयार करने के चरणों में फायदा उठाना है (जैसे इमेजेस की क्रॉपिंग और रीसाइज़िंग)।

अपडेट्स के लिए सिक्योर बाई डिज़ाइन एप्रोच का पालन करें 

आप प्रत्येक उत्पाद में डिफ़ॉल्ट रूप से ऑटोमेटड अपडेट्स शामिल करते/करती हैं और उन्हें उपलब्ध कराने के लिए सिक्योर, मॉड्यूलर अपडेट प्रक्रियाओं का उपयोग करते/करती हैं। आपकी अपडेट प्रक्रियाएँ (जिसमें परीक्षण और आकलन व्यवस्थाएँ शामिल हैं) इस तथ्य को दर्शाती हैं कि डेटा, मॉडल्स या संकेतों में परिवर्तनों से सिस्टम के व्यवहार में परिवर्तन हो सकता है (उदाहरण के लिए, आप प्रमुख अपडेट्स को नए संस्करणों की तरह मानते/मानती हैं)। आप यूज़र्स को मॉडल परिवर्तनों का आकलन करने और उनका उत्तर देने में समर्थित करते/करती हैं (उदाहरण के लिए, प्रिव्यू एक्सेस और वर्ज़न्ड API उपलब्ध कराके)।

सीखे गए सबक एकत्र और साझा करें 

आप उद्योग, शिक्षा और सरकारों के वैश्विक इकोसिस्टम में सहयोग देते हुए सूचना-साझाकरण समुदायों में भाग लेते/लेती हैं, ताकि उपयुक्तानुसार सर्वोत्तम कार्यप्रथा को साझा किया जा सके। आप अपने संगठन में आंतरिक और बाहरी रूप से सिस्टम सिक्योरिटी के बारे में फीडबैक के लिए संचार लाइनें खुली बनाए रखते/रखती हैं, जिसमें सिक्योरिटी शोधकर्ताओं को कमजोरियों पर शोध करने और उनके बारे में सूचित करने के लिए सहमति प्रदान करना शामिल है। आप मुद्दों को आवश्यकतानुसार व्यापक समुदाय में आगे ले जाते/ती हैं, उदाहरण के लिए भेद्यता प्रकटीकरण का उत्तर देने वाली बुलेटिन्स प्रकाशित करना, जिसमें विस्तृत और संपूर्ण सामान्य भेद्यताओं को सूचीबद्ध करना भी शामिल है। आप मुद्दों को जल्दी और उचित रूप से मिटिगेट व हल करने के लिए कदम उठाते/उठाती हैं।

आगे पढ़ें

AI डेवलपमेंट

मशीन लर्निंग की सिक्योरिटी के लिए सिद्धांत
ML कॉम्पोनेंट के साथ सिस्टम को डेवलप, डिप्लॉय या ऑपरेट करने के बारे में NCSC का विस्तृत मार्गदर्शन।

सिक्योर बाई डिज़ाइन - साइबर सिक्योरिटी जोखिम के संतुलन को स्थानांतरित करना: सिक्योर बाई डिज़ाइन सॉफ़्टवेयर के लिए सिद्धांत और एप्रोचेज़
CISA, NCSC तथा अन्य एजेंसियों द्वारा सह-लिखित यह मार्गदर्शन वर्णन करता है कि सॉफ़्टवेयर सिस्टम के निर्माताओं को, जिसमें AI भी शामिल है, प्रोडक्ट डेवलपमेंट के डिज़ाइन चरण में सिक्योरिटी को कारक बनाने के लिए कदम कैसे उठाने चाहिए, और ऐसे प्रोडक्टस भेजे जाने चाहिए जो डिब्बे से निकलें तो पूरी तरह सुरक्षित हों।

संक्षेप में AI सिक्योरिटी चिंताएँ
जर्मन फेडरल ऑफिस फॉर इंफॉर्मेशन सिक्योरिटी (BSI) द्वारा निर्मित यह दस्तावेज मशीन लर्निंग सिस्टम्स पर संभावित एटैक्स और उन एटैक्स से संभावित बचावों के बारे में बताता है।

हिरोशिमा प्रक्रिया एडवांस्ड AI सिस्टम्स विकसित करने वाले संगठनों के लिए अंतर्राष्ट्रीय मार्गदर्शक सिद्धांत और हिरोशिमा प्रक्रिया एडवांस्ड AI सिस्टम्स विकसित करने वाले संगठनों के लिए अंतर्राष्ट्रीय आचार संहिता
ये दस्तावेज, जिन्हें G7 हिरोशिमा AI प्रक्रिया के हिस्से के रूप में निर्मित किया गया है, सबसे एडवांस्ड AI सिस्टम्स विकसित करने वाले संगठनों के लिए मार्गदर्शन प्रदान करते हैं, जिसमें सबसे एडवांस्ड फाउंडेशन मॉडल्स और जेनरेटिव AI शामिल हैं और इसका उद्देश्य पूरे विश्व-भर में सुरक्षित, सिक्योर और भरोसेमंद AI को बढ़ावा देना है।

AI वेरिफाई
सिंगापुर का AI गवर्नेंस टेस्टिंग फ्रेमवर्क और सॉफ़्टवेयर टूलकिट, जो मानकीकृत परीक्षणों के माध्यम से अंतरराष्ट्रीय स्तर पर मान्यता-प्राप्त सिद्धांतों के सेट के प्रति AI सिस्टम्स के प्रदर्शन को मान्यता देता है।

AI के लिए अच्छी साइबर सिक्योरिटी कार्यप्रथाओं का बहुस्तरीय फ्रेमवर्क - ENISA (europa.eu)
राष्ट्रीय सक्षम प्राधिकारियों और AI हितधारकों के कदमों का मार्गदर्शन करने के लिए एक फ्रेमवर्क, जिनका पालन करने की आवश्यकता उन्हें अपने AI सिस्टम्स, ऑपरेशन्स और प्रक्रियाओं को सिक्योर करने के लिए है।

ISO 5338: AI सिस्टम लाइफ साइकल प्रक्रियाएँ (समीक्षाधीन)
AI सिस्टम्स के लाइफ साइकल का वर्णन करने के लिए प्रक्रियाओं और संबंधित अवधारणाओं का एक सेट, जो मशीन लर्निंग और ह्यूरिस्टिक सिस्टम्स पर आधारित है।

AI क्लाउड सर्विस कंप्लायेंस क्राइटीरिया कैटलॉग (AIC4)
BSI का AI क्लाउड सर्विस कंप्लायेंस क्राइटीरिया कैटलॉग AI-विशिष्ट मानदंड प्रदान करता है, जो AI सर्विस के लाइफ साइकल में उसकी सिक्योरिटी के आकलन को सक्षम बनाता है।

NIST IR 8269 (मसौदा) एडवर्सीरियल मशीन लर्निंग की टैक्सोनॉमी और टर्मिनोलॉजी
मशीन लर्निंग और ह्यूरिस्टिक सिस्टम्स पर आधारित AI सिस्टम्स के लाइफ साइकल का वर्णन करने के लिए प्रक्रियाओं और संबंधित अवधारणाओं का एक सेट।

MITRE ATLAS
मशीन लर्निंग (ML) सिस्टम्स के प्रति विरोधी रणनीतियों, तकनीकों और केस स्टडीज़ का एक ज्ञान आधार, जिसे MITRE ATT&CK फ्रेमवर्क के अनुसार मॉडल करके उसमें जोड़ा गया है।

कैटेस्ट्रॉफिक AI खतरों का अवलोकन (2023)
सेंटर फॉर AI सिक्योरिटी द्वारा निर्मित यह दस्तावेज AI से उत्पन्न जोखिम के क्षेत्रों को नियत करता है।

लार्ज लैंग्वेज मॉडल्स: उद्योग और प्राधिकरणों के लिए अवसर और खतरे
कंपनियों, अधिकारियों और डेवलपर्स के लिए BSI द्वारा निर्मित दस्तावेज, जो LLMs के डेवलपमेंट, डिप्लॉयमेंट और/या उपयोग के अवसरों और खतरों के बारे में और अधिक जानना चाहते हैं।

AI मॉडल्स का सिक्योरिटी परीक्षण करने में यूज़र्स की सहायता करने वाले ओपन-सोर्स प्रोजेक्ट्स में शामिल हैं:

- एडवर्सीरियल रोबस्टनेस टूलबॉक्स (IBM)
- क्लेवरहैंस (टोरंटो विश्वविद्यालय)
- टेक्स्टएटैक (वर्जीनिया विश्वविद्यालय)
- प्रॉम्प्ट बेंच (माइक्रोसॉफ़्ट)
- काउंटरफिट (माइक्रोसॉफ़्ट)
- AI वेरिफाई (इन्फोकॉम मीडिया डेवलपमेंट ऑथोरिटी, सिंगापुर)

साइबर सिक्योरिटी

CISA के साइबर सिक्योरिटी प्रदर्शन लक्ष्य
सिक्योरिटीज़ का एक सामान्य सेट, जिसे सभी महत्वपूर्ण बुनियादी ढांचा संस्थाओं को ज्ञात जोखिमों और विरोधी तकनीकों की संभावना और प्रभाव को सार्थक रूप से कम करने के लिए लागू करना चाहिए।

NCSC CAF फ्रेमवर्क
साइबर एसेसमेंट फ्रेमवर्क (CAF) महत्वपूर्ण रूप से अतिमहत्वपूर्ण सेवाओं और गतिविधियों के लिए जिम्मेदार संगठनों को मार्गदर्शन प्रदान करता है।

MITRE का आपूर्ति श्रृंखला सिक्योरिटी फ्रेमवर्क
आपूर्ति श्रृंखला के अंदर आपूर्तिकर्ताओं और सर्विस प्रोवाइडर्स के आकलन के लिए एक रूपरेखा।

जोखिम प्रबंधन

NIST AI जोखिम प्रबंधन फ्रेमवर्क (AI RMF)
AI RMF यह रेखांकित करता है कि AI के साथ विशिष्ट रूप से जुड़े व्यक्तियों, संगठनों और समाज के लिए सामाजिक-तकनीकी जोखिमों का प्रबंधन कैसे किया जाए।

ISO 27001: सूचना सिक्योरिटी, साइबर सिक्योरिटी और गोपनीयता संरक्षण
यह मानक संगठनों को सूचना सिक्योरिटी प्रबंधन सिस्टम के इस्टैबलिशमेंट, इंप्लिमेंटेशन और मेन्टनेंस के बारे में मार्गदर्शन प्रदान करता है।

ISO 31000: जोखिम प्रबंधन
संगठनों के अंदर जोखिम प्रबंधन के लिए संगठनों को दिशा-निर्देश और सिद्धांत प्रदान करने वाला अंतरराष्ट्रीय मानक।

NCSC जोखिम प्रबंधन मार्गदर्शन
यह मार्गदर्शन साइबर सिक्योरिटी के जोखिम पेशेवरों को अपने संगठनों को प्रभावित करने वाले साइबर सिक्योरिटी से संबंधित खतरों को बेहतर ढंग से समझने और उनका प्रबंधन करने में सहायता देता है।

नोट्स

1. यहाँ ऐसे व्यक्ति, सार्वजनिक प्राधिकरण, एजेंसी या अन्य निकाय के रूप में परिभाषित किया गया है, जो AI सिस्टम विकसित करता है (या जिसने पहले AI सिस्टम विकसित किया है) और उस सिस्टम को बाज़ार में शामिल करता है या इसे अपने नाम या ट्रेडमार्क के तहत सेवा में उपलब्ध कराता है
2. सिक्योर बाई डिज़ाइन के बारे में और अधिक जानकारी के लिए देखें: CISA का सिक्योर बाई डिज़ाइन वेब पेज और मार्गदर्शन साइबर सिक्योरिटी के जोखिम संतुलन को स्थानांतरित करना: सिक्योर बाई डिज़ाइन सॉफ़्टवेयर के लिए सिद्धांत और एप्रोचेज़
3. Non-ML AI दृष्टिकोणों के विपरीत, नियम-आधारित सिस्टम्स के समान
4. CEPS अपने प्रकाशन में सात अलग-अलग प्रकार के AI डेवलपमेंट इंटरएक्शन का वर्णन करते हैं 'यूरोपीय संघ के आर्टिफिशियल इंटेलिजेंस एक्ट के साथ AI वैल्यू चेन का मिलान'
5. ISO/IEC 22989:2022(en) इसे 'AI सिस्टम्स का निर्माण करने वाले कार्यात्मक तत्व' के रूप में परिभाषित करता है
6. NIST को आर्टिफिशियल इंटेलिजेंस (AI) के सिक्योर, सुरक्षित और भरोसेमंद डेवलपमेंट तथा उपयोग को आगे बढ़ाने के लिए दिशा-निर्देश (और अन्य कार्यवाही करने) का काम सौंपा गया है। 30 अक्टूबर, 2023 के कार्यकारी आदेश के तहत NIST की ज़िम्मेदारियाँ देखें
7. खतरे की मॉडलिंग के बारे में और अधिक जानकारी OWASP फाउंडेशन के पास उपलब्ध है
8. MITRE ATLAS एडवर्सरियल मशीन लर्निंग 101 देखें
9. GitHub: मैलिशियस लैंब्डा लेयर के उपयोग से टेंसरफ्लो के लिए RCE PoC
10. SLSA: 'किसी भी सॉफ़्टवेयर आपूर्ति श्रृंखला में आर्टिफैक्ट इंटेग्रिटी का संरक्षण'
11. METI (जापानी अर्थव्यवस्था, वाणिज्य एवं उद्योग मंत्रालय, 2023), 'सॉफ़्टवेयर प्रबंधन के प्रयोजन से सॉफ़्टवेयर बिल ऑफ़ मटीरियल्स (SBOM) के परिचय के लिए मार्गदर्शिका'
12. गूगल अनुसंधान: मशीन लर्निंग: टेक्निकल डेट का हाई इंटरेस्ट क्रेडिट कार्ड
13. ट्रेमर एट ऑल 2016, प्रिडिक्शन APIs के माध्यम से मशीन लर्निंग मॉडल्स की चोरी करना
14. बोनिश, 2020, मशीन लर्निंग गोपनीयता पर एटैक्स (भाग 1): IBM-ART फ्रेमवर्क के साथ मॉडल इन्वर्ज़न एटैक्स
15. नेशनल साइबर सिक्योरिटी सेंटर, 2020, निजी तौर पर होस्ट किए गए पब्लिक की इंफ्रास्ट्रक्चर का डिज़ाइन और निर्माण