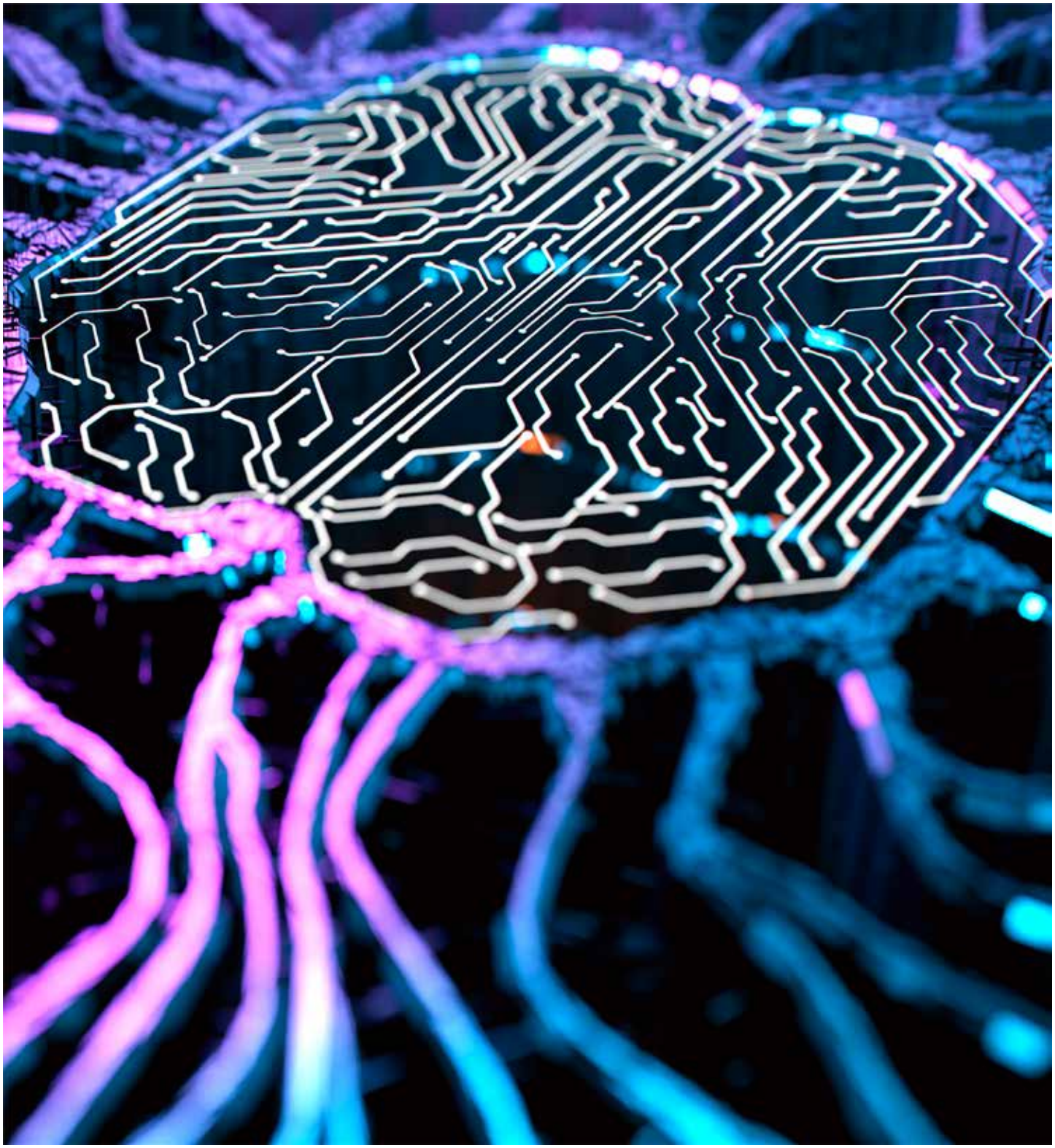
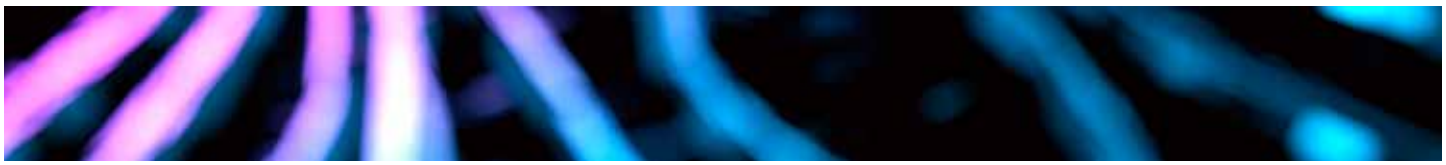
Οδηγίες για ασφαλή ανάπτυξη συστήματος AI





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

Σχετικά με αυτό το έγγραφο

Αυτό το έγγραφο δημοσιεύεται από το Εθνικό Κέντρο Κυβερνοασφάλειας του Ηνωμένου Βασιλείου (NCSC), την Υπηρεσία Κυβερνοασφάλειας και Ασφάλειας Υποδομών των ΗΠΑ (CISA) και τους ακόλουθους διεθνείς εταίρους:

- Υπηρεσία Εθνικής Ασφάλειας (NSA)
- Ομοσπονδιακό Γραφείο Ερευνών (FBI)
- Αυστραλιανό Κέντρο Κυβερνοασφάλειας (ACSC) της Αυστραλιανής Διεύθυνσης Σημάτων
- Καναδικό Κέντρο για την Ασφάλεια στον Κυβερνοχώρο (CCCS)
- Εθνικό Κέντρο Κυβερνοασφάλειας Νέας Ζηλανδίας (NCSC-NZ)
- CSIRT της Κυβέρνησης της Χιλής
- Εθνική Υπηρεσία Ασφάλειας Κυβερνοχώρου και Πληροφοριών της Τσεχίας (NUKIB)
- Αρχή Πληροφοριακών Συστημάτων της Εσθονίας (RIA) και Εθνικό Κέντρο Κυβερνοασφάλειας της Εσθονίας (NCSC-EE)
- Γαλλική Υπηρεσία Κυβερνοασφάλειας (ANSSI)
- Ομοσπονδιακή Υπηρεσία Ασφάλειας Πληροφοριών της Γερμανίας (BSI)
- Ισραηλινή Εθνική Διεύθυνση Κυβερνοχώρου (INCD)
- Ιταλική Εθνική Υπηρεσία Κυβερνοασφάλειας (ACN)
- Εθνικό κέντρο ετοιμότητας συμβάντων και στρατηγικής για την ασφάλεια στον κυβερνοχώρο (NISC) της Ιαπωνίας
- Γραμματεία Επιστήμης, Τεχνολογίας και Πολιτικής Καινοτομίας της Ιαπωνίας, Γραφείο Υπουργικού Συμβουλίου
- Εθνική Υπηρεσία Ανάπτυξης Τεχνολογίας Πληροφορικής της Νιγηρίας (NITDA)
- Νορβηγικό Εθνικό Κέντρο Κυβερνοασφάλειας (NCSC-NO)
- Υπουργείο Ψηφιακών Υποθέσεων της Πολωνίας
- Εθνικό Ινστιτούτο Ερευνών NASK της Πολωνίας (NASK)
- Εθνική Υπηρεσία Πληροφοριών της Δημοκρατίας της Κορέας (NIS)
- Υπηρεσία Κυβερνοασφάλειας της Σιγκαπούρης (CSA)

Ευχαριστίες

Οι ακόλουθοι οργανισμοί συνέβαλαν στην ανάπτυξη αυτών των κατευθυντήριων γραμμών:

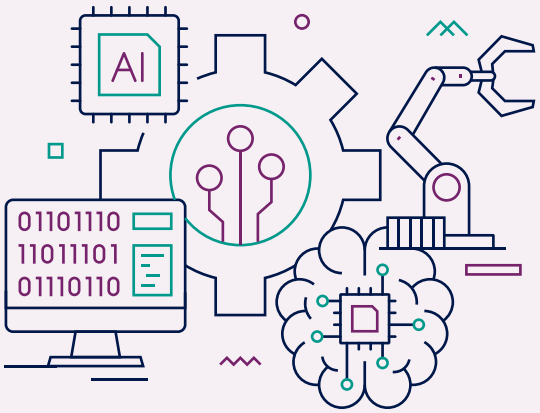
- Ινστιτούτο Alan Turing
- Anthropic
- Databricks
- Κέντρο Ασφάλειας και Αναδυόμενης Τεχνολογίας του Πανεπιστημίου Georgetown
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Ινστιτούτο Μηχανικής Λογισμικού στο Πανεπιστήμιο Carnegie Mellon
- Stanford Center for AI Safety
- Πρόγραμμα Stanford για τη Γεωπολιτική, την Τεχνολογία και τη Διακυβέρνηση

Αποποίηση ευθυνών

Οι πληροφορίες σε αυτό το έγγραφο παρέχονται «ως έχουν» από το NCSC και τους συντάκτες οργανισμούς, οι οποίοι δεν ευθύνονται για οποιαδήποτε απώλεια, βλάβη ή ζημιά οποιουδήποτε είδους που προκαλείται από τη χρήση του, εκτός εάν απαιτείται από τη νομοθεσία. Οι πληροφορίες σε αυτό το έγγραφο δεν συνιστούν ούτε υπονοούν έγκριση ή σύσταση οποιουδήποτε οργανισμού, προϊόντος ή υπηρεσίας τρίτου μέρους από το NCSC και τα γραφεία σύνταξης. Οι σύνδεσμοι και οι παραπομπές σε ιστότοπους και υλικό τρίτων παρέχονται μόνο για ενημέρωση και δεν αντιπροσωπεύουν έγκριση ή σύσταση τέτοιων πόρων έναντι άλλων.

Αυτό το έγγραφο διατίθεται σε βάση TLP:CLEAR (https://www.first.org/tlp/).

Περιεχόμενα



Περίληψη των κυριότερων σημείων

Αυτό το έγγραφο συνιστά οδηγίες για τους παρόχους οποιωνδήποτε συστημάτων που χρησιμοποιούν τεχνητή νοημοσύνη (AI), είτε αυτά τα συστήματα έχουν δημιουργηθεί από το μηδέν είτε έχουν κατασκευαστεί πάνω από εργαλεία και υπηρεσίες που παρέχονται από άλλους. Η εφαρμογή αυτών των κατευθυντήριων γραμμών θα βοηθήσει τους παρόχους να δημιουργήσουν συστήματα AI που λειτουργούν όπως προβλέπεται, είναι διαθέσιμα όταν χρειάζεται και λειτουργούν χωρίς να αποκαλύπτουν ευαίσθητα δεδομένα σε μη εξουσιοδοτημένα μέρη.

Αυτό το έγγραφο απευθύνεται κυρίως σε παρόχους συστημάτων AI που χρησιμοποιούν μοντέλα που φιλοξενούνται από έναν οργανισμό ή χρησιμοποιούν εξωτερικές διεπαφές προγραμματισμού εφαρμογών (API). Προτρέπουμε **όλα** τα ενδιαφερόμενα μέρη (συμπεριλαμβανομένων των επιστημόνων δεδομένων, των προγραμματιστών, των διαχειριστών, των υπευθύνων λήψης αποφάσεων και των ιδιοκτητών κινδύνου) να διαβάσουν αυτές τις οδηγίες για να τους βοηθήσουν να λάβουν τεκμηριωμένες αποφάσεις σχετικά με τον **σχεδιασμό**, **ανάπτυξη**, **εγκατάσταση** και **λειτουργία** των συστημάτων AI τους.

Σχετικά με τις οδηγίες

Τα συστήματα AI έχουν τη δυνατότητα να αποφέρουν πολλά οφέλη στην κοινωνία. Ωστόσο, για να αξιοποιηθούν πλήρως οι ευκαιρίες της AI, πρέπει να αναπτυχθεί, να εγκατασταθεί και να λειτουργήσει με ασφαλή και υπεύθυνο τρόπο.

Τα συστήματα AI υπόκεινται σε νέα τρωτά σημεία ασφαλείας που πρέπει να ληφθούν υπόψη παράλληλα με τις τυπικές απειλές για την ασφάλεια στον κυβερνοχώρο. Όταν ο ρυθμός ανάπτυξης είναι υψηλός – όπως συμβαίνει με την AI – η ασφάλεια μπορεί συχνά να αποτελεί δευτερεύουσα σημασία. Η ασφάλεια πρέπει να είναι βασική απαίτηση, όχι μόνο στη φάση ανάπτυξης, αλλά σε όλο τον κύκλο ζωής του συστήματος.

Για αυτόν τον λόγο, οι κατευθυντήριες γραμμές αναλύονται σε τέσσερις βασικούς τομείς εντός του κύκλου ζωής της ανάπτυξης του συστήματος AI: **ασφαλής σχεδιασμός**, **ασφαλής ανάπτυξη**, **ασφαλής εγκατάσταση**, και **ασφαλής λειτουργία και συντήρηση**. Για κάθε ενότητα προτείνουμε σκέψεις και μετριασμούς που θα συμβάλουν στη μείωση του συνολικού κινδύνου για μια οργανωτική διαδικασία ανάπτυξης συστήματος AI.

1. **Ασφαλής σχεδιασμός**
   Αυτή η ενότητα περιέχει οδηγίες που ισχύουν για το στάδιο σχεδιασμού του κύκλου ζωής ανάπτυξης του συστήματος AI. Καλύπτει την κατανόηση των κινδύνων και τη μοντελοποίηση απειλών, καθώς και συγκεκριμένα θέματα και συμβιβασμούς που πρέπει να ληφθούν υπόψη σχετικά με τον σχεδιασμό του συστήματος και του μοντέλου.

2. **Ασφαλής ανάπτυξη**
   Αυτή η ενότητα περιέχει οδηγίες που ισχύουν για το στάδιο ανάπτυξης του κύκλου ζωής της ανάπτυξης του συστήματος AI, συμπεριλαμβανομένης της ασφάλειας της εφοδιαστικής αλυσίδας, της τεκμηρίωσης και της διαχείρισης περιουσιακών στοιχείων και τεχνικού χρέους.

3. **Ασφαλής εγκατάσταση**
   Αυτή η ενότητα περιέχει οδηγίες που ισχύουν για το στάδιο εγκατάστασης του κύκλου ζωής ανάπτυξης συστήματος AI, συμπεριλαμβανομένης της προστασίας της υποδομής και των μοντέλων από παραβίαση, απειλή ή απώλεια, ανάπτυξη διαδικασιών διαχείρισης συμβάντων και υπεύθυνη γνωστοποίηση.

4. **Ασφαλής λειτουργία και συντήρηση**
   Αυτή η ενότητα περιέχει οδηγίες που ισχύουν για το στάδιο ασφαλούς λειτουργίας και συντήρησης του κύκλου ζωής ανάπτυξης συστήματος AI. Παρέχει οδηγίες σχετικά με ενέργειες που είναι ιδιαίτερα σχετικές μετά την εγκατάσταση ενός συστήματος, συμπεριλαμβανομένης της καταγραφής και της παρακολούθησης, της διαχείρισης ενημερώσεων και της κοινής χρήσης πληροφοριών.

Οι οδηγίες ακολουθούν μια «ασφαλή από προεπιλογή» προσέγγιση και ευθυγραμμίζονται στενά με τις πρακτικές που ορίζονται στο Secure development and deployment guidance (Καθοδήγηση για ασφαλή ανάπτυξη και εγκατάσταση) του NCSC, στο Secure Software Development Framework (Πλαίσιο Ασφαλούς Ανάπτυξης Λογισμικού) του NIST, και «secure by design principles (αρχές ασφάλειας από τον σχεδιασμό)» που δημοσιεύτηκε από την CISA, το NCSC και διεθνείς οργανισμούς στον κυβερνοχώρο. Δίνουν προτεραιότητα:

- στην ανάληψη ευθύνης των αποτελεσμάτων ασφάλειας για τους πελάτες
- στο να υιοθετήσουμε τη ριζική διαφάνεια και υπευθυνότητα
- στην οικοδόμηση οργανωτικής δομής και ηγεσίας ώστε η ασφαλής από τον σχεδιασμό προσέγγιση να αποτελεί κορυφαία επιχειρηματική προτεραιότητα

Εισαγωγή

Τα συστήματα τεχνητής νοημοσύνης (AI) έχουν τη δυνατότητα να αποφέρουν πολλά οφέλη στην κοινωνία. Ωστόσο, για να αξιοποιηθούν πλήρως οι ευκαιρίες της, πρέπει να αναπτυχθεί, να εγκατασταθεί και να λειτουργήσει με ασφαλή και υπεύθυνο τρόπο. Η ασφάλεια στον κυβερνοχώρο είναι απαραίτητη προϋπόθεση για την ασφάλεια, την ανθεκτικότητα, το απόρρητο, τη δικαιοσύνη, την αποτελεσματικότητα και την αξιοπιστία των συστημάτων AI.

Ωστόσο, τα συστήματα AI υπόκεινται σε νέα τρωτά σημεία ασφαλείας που πρέπει να ληφθούν υπόψη παράλληλα με τις τυπικές απειλές για την ασφάλεια στον κυβερνοχώρο. Όταν ο ρυθμός ανάπτυξης είναι υψηλός – όπως συμβαίνει με την AI – η ασφάλεια μπορεί συχνά να αποτελεί δευτερεύουσα σημασία. Η ασφάλεια πρέπει να είναι βασική απαίτηση, όχι μόνο στη φάση ανάπτυξης, αλλά σε όλο τον κύκλο ζωής του συστήματος.

Αυτό το έγγραφο συνιστά οδηγίες για παρόχους[1] οποιωνδήποτε συστημάτων που χρησιμοποιούν AI, είτε αυτά τα συστήματα έχουν δημιουργηθεί από το μηδέν είτε έχουν κατασκευαστεί πάνω από εργαλεία και υπηρεσίες που παρέχονται από άλλους. Η εφαρμογή αυτών των οδηγιών θα βοηθήσει τους παρόχους να δημιουργήσουν συστήματα AI που λειτουργούν όπως προβλέπεται, είναι διαθέσιμα όταν χρειάζεται και λειτουργούν χωρίς να αποκαλύπτουν ευαίσθητα δεδομένα σε μη εξουσιοδοτημένα μέρη.

Αυτές οι κατευθυντήριες γραμμές θα πρέπει να λαμβάνονται υπόψη σε συνδυασμό με την καθιερωμένη ασφάλεια στον κυβερνοχώρο, τη διαχείριση κινδύνων και τις βέλτιστες πρακτικές αντιμετώπισης συμβάντων. Ειδικότερα, παροτρύνουμε τους παρόχους να ακολουθούν τις αρχές «secure by design»[2] που αναπτύχθηκαν από την Υπηρεσία Κυβερνοασφάλειας και Ασφάλειας Υποδομών των ΗΠΑ (CISA), το Εθνικό Κέντρο Κυβερνοασφάλειας του Ηνωμένου Βασιλείου (NCSC) και όλους τους διεθνείς εταίρους μας. Οι αρχές δίνουν προτεραιότητα:

- στην ανάληψη ευθύνης των αποτελεσμάτων ασφάλειας για τους πελάτες
- στο να υιοθετήσουμε τη ριζική διαφάνεια και υπευθυνότητα
- στην οικοδόμηση οργανωτικής δομής και ηγεσίας ώστε η ασφαλής από τον σχεδιασμό προσέγγιση να αποτελεί κορυφαία επιχειρηματική προτεραιότητα.

Η τήρηση των αρχών «secure by design» απαιτεί σημαντικούς πόρους σε όλο τον κύκλο ζωής ενός συστήματος. Σημαίνει ότι οι προγραμματιστές πρέπει να επενδύσουν στην ιεράρχηση **χαρακτηριστικών**, **μηχανισμών** και **υλοποίησης** εργαλείων που προστατεύουν τους πελάτες σε κάθε επίπεδο του σχεδιασμού του συστήματος και σε όλα τα στάδια του κύκλου ζωής της ανάπτυξης. Κάνοντας αυτό θα αποτρέψουμε τους δαπανηρούς επανασχεδιασμούς αργότερα, καθώς και την προστασία των πελατών και των δεδομένων τους βραχυπρόθεσμα.

Γιατί η ασφάλεια AI είναι διαφορετική;

Σε αυτό το έγγραφο χρησιμοποιούμε το «AI» για να αναφερθούμε συγκεκριμένα σε εφαρμογές μηχανικής εκμάθησης (ML)[3]. Όλοι οι τύποι ML είναι εντός πεδίου εφαρμογής. Ορίζουμε τις εφαρμογές ML ως εφαρμογές που:

- περιλαμβάνουν στοιχεία λογισμικού (μοντέλα) που επιτρέπουν στους υπολογιστές να αναγνωρίζουν και να φέρνουν το πλαίσιο σε μοτίβα δεδομένων χωρίς οι κανόνες να πρέπει να προγραμματίζονται ρητά από άνθρωπο
- δημιουργούν προβλέψεις, συστάσεις ή αποφάσεις που βασίζονται σε στατιστική λογική

Εκτός από τις υπάρχουσες απειλές για την ασφάλεια στον κυβερνοχώρο, τα συστήματα AI υπόκεινται σε νέους τύπους τρωτών σημείων. Ο όρος «adversarial machine learning» (αντίπαλη τεχνητή μάθηση) (AML), χρησιμοποιείται για να περιγράψει την εκμετάλλευση θεμελιωδών τρωτών σημείων σε στοιχεία ML, συμπεριλαμβανομένου υλικού, λογισμικού, ροών εργασίας και αλυσίδων εφοδιασμού. Η AML δίνει τη δυνατότητα στους εισβολείς να προκαλούν ανεπιθύμητες συμπεριφορές σε συστήματα ML που μπορεί μεταξύ άλλων να:

- επηρεάζουν την απόδοση ταξινόμησης ή παλινδρόμησης του μοντέλου
- επιτρέπουν στους χρήστες να εκτελούν μη εξουσιοδοτημένες ενέργειες
- εξάγουν ευαίσθητες πληροφορίες μοντέλου

Υπάρχουν πολλοί τρόποι για αυτά τα αποτελέσματα, όπως οι επιθέσεις άμεσης έγχυσης στον τομέα του μοντέλου μεγάλων γλωσσών (LLM) ή η σκόπιμη καταστροφή των δεδομένων εκπαίδευσης ή των σχολίων των χρηστών (γνωστή ως «δηλητηρίαση δεδομένων»).

Ποιος πρέπει να διαβάσει αυτό το έγγραφο;

Αυτό το έγγραφο απευθύνεται κυρίως σε παρόχους συστημάτων AI, είτε βασίζονται σε μοντέλα που φιλοξενούνται από έναν οργανισμό είτε κάνουν χρήση διεπαφών προγραμματισμού εξωτερικών εφαρμογών (API). Ωστόσο, παροτρύνουμε **όλα** τα ενδιαφερόμενα μέρη (συμπεριλαμβανομένων των επιστημόνων δεδομένων, των προγραμματιστών, των διαχειριστών, των υπευθύνων λήψης αποφάσεων και των ιδιοκτητών κινδύνου) να διαβάσουν αυτές τις οδηγίες για να τους βοηθήσουν να λάβουν τεκμηριωμένες αποφάσεις σχετικά με τον **σχεδιασμό**, **εγκατάσταση** και **λειτουργία** των συστημάτων μηχανικής εκμάθησης AI.

Τούτου λεχθέντος, δεν θα ισχύουν άμεσα όλες οι κατευθυντήριες γραμμές σε όλους τους οργανισμούς. Το επίπεδο πολυπλοκότητας και οι μέθοδοι επίθεσης θα ποικίλλουν ανάλογα με τον αντίπαλο που στοχεύει το σύστημα AI, επομένως οι οδηγίες θα πρέπει να λαμβάνονται υπόψη παράλληλα με τις περιπτώσεις χρήσης και το προφίλ απειλών του οργανισμού σας.

Ποιος είναι υπεύθυνος για την ανάπτυξη ασφαλούς AI;

Υπάρχουν συχνά πολλοί παράγοντες στις σύγχρονες αλυσίδες εφοδιασμού με AI. Μια απλή προσέγγιση προϋποθέτει δύο οντότητες:

- ο «πάροχος» που είναι υπεύθυνος για την επιμέλεια δεδομένων, την αλγοριθμική ανάπτυξη, τον σχεδιασμό, την εγκατάσταση και τη συντήρηση
- ο «χρήστης», ο οποίος παρέχει εισόδους και λαμβάνει εξόδους

Ενώ αυτή η προσέγγιση παρόχου-χρήστη χρησιμοποιείται σε πολλές εφαρμογές, γίνεται ολοένα και πιο ασυνήθιστη[4], καθώς οι πάροχοι μπορεί να επιδιώκουν να ενσωματώσουν λογισμικό, δεδομένα, μοντέλα ή/και απομακρυσμένες υπηρεσίες που παρέχονται από τρίτους στα δικά τους συστήματα. Αυτές οι πολύπλοκες αλυσίδες εφοδιασμού καθιστούν πιο δύσκολο για τους τελικούς χρήστες να κατανοήσουν πού βρίσκεται η ευθύνη για την ασφαλή AI.

Οι χρήστες (είτε «τελικοί χρήστες», είτε πάροχοι που ενσωματώνουν εξωτερικό στοιχείο AI[5]) συνήθως δεν έχουν επαρκή ορατότητα ή/και τεχνογνωσία για να κατανοήσουν πλήρως, να αξιολογήσουν ή να αντιμετωπίσουν τους κινδύνους που σχετίζονται με τα συστήματα που χρησιμοποιούν. Ως εκ τούτου, σύμφωνα με τις αρχές «secure by design», **οι πάροχοι στοιχείων AI θα πρέπει να αναλαμβάνουν την ευθύνη για τα αποτελέσματα ασφάλειας των χρηστών που βρίσκονται πιο κάτω από την αλυσίδα εφοδιασμού.**

Οι πάροχοι θα πρέπει να εφαρμόζουν ελέγχους ασφαλείας και μετριασμούς όπου είναι δυνατόν στα μοντέλα, τους αγωγούς και/ή τα συστήματά τους, και όπου χρησιμοποιούνται ρυθμίσεις, να εφαρμόζουν την πιο ασφαλή επιλογή ως προεπιλογή. Όπου οι κίνδυνοι δεν μπορούν να μετριαστούν, ο πάροχος θα πρέπει να είναι υπεύθυνος για:

- ενημέρωση των χρηστών περαιτέρω στην αλυσίδα εφοδιασμού για τους κινδύνους που αποδέχονται οι ίδιοι και (αν ισχύει) οι δικοί τους χρήστες
- παροχή συμβουλών σχετικά με το πώς να χρησιμοποιούν το εξάρτημα με ασφάλεια

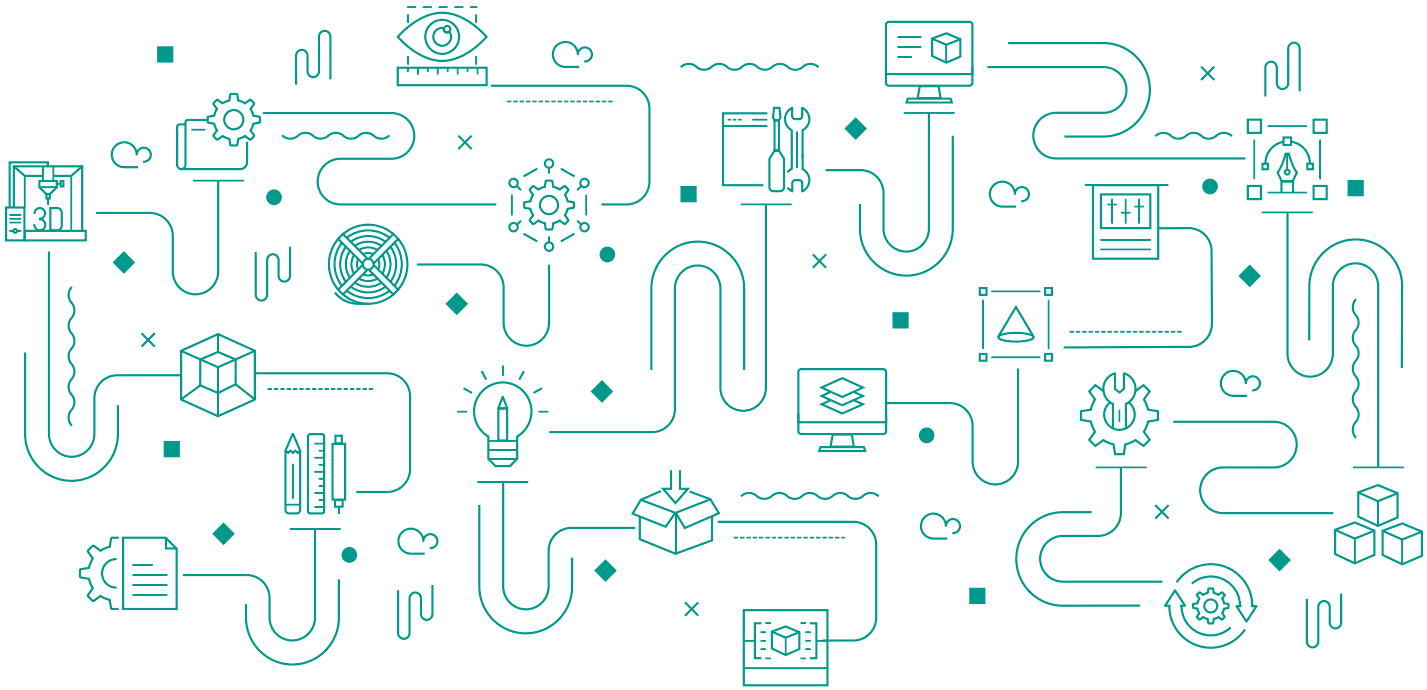
Όπου η παραβίαση του συστήματος θα μπορούσε να οδηγήσει σε απτή ή εκτεταμένη φυσική ζημιά ή προσβολή της φήμης, σημαντική απώλεια επιχειρηματικών λειτουργιών, διαρροή ευαίσθητων ή εμπιστευτικών πληροφοριών ή/και νομικές επιπτώσεις, οι κίνδυνοι για την ασφάλεια στον κυβερνοχώρο της AI θα πρέπει να αντιμετωπίζονται ως **κρίσιμοι.**

Οδηγίες για ασφαλή ανάπτυξη συστήματος AI

Οι οδηγίες αναλύονται σε τέσσερις βασικούς τομείς εντός του κύκλου ζωής της ανάπτυξης του συστήματος AI: **ασφαλής σχεδιασμός**, **ασφαλής ανάπτυξη**, **ασφαλής εγκατάσταση** και **ασφαλής λειτουργία και συντήρηση**. Για κάθε τομέα, προτείνουμε εκτιμήσεις και μετριασμούς που θα συμβάλουν στη μείωση του συνολικού κινδύνου για τη διαδικασία ανάπτυξης οργανωτικού συστήματος AI.

Οι κατευθυντήριες γραμμές που ορίζονται σε αυτό το έγγραφο είναι στενά ευθυγραμμισμένες με τις πρακτικές του κύκλου ζωής ανάπτυξης λογισμικού που ορίζονται στα εξής:

- Secure development and deployment guidance (Καθοδήγηση για ασφαλή ανάπτυξη και εγκατάσταση) του NCSC
- Secure Software Development Framework (Πλαίσιο Ανάπτυξης Ασφαλούς Λογισμικού) (SSDF)[6] του Εθνικού Ινστιτούτου Προτύπων και Τεχνολογίας (NIST)



1. Ασφαλής σχεδιασμός

Αυτή η ενότητα περιέχει οδηγίες που ισχύουν για το στάδιο **σχεδιασμού** του κύκλου ζωής ανάπτυξης του συστήματος AI. Καλύπτει την κατανόηση των κινδύνων και τη μοντελοποίηση απειλών, καθώς και συγκεκριμένα θέματα και συμβιβασμούς που πρέπει να ληφθούν υπόψη σχετικά με τον σχεδιασμό του συστήματος και του μοντέλου.

Ευαισθητοποίηση του προσωπικού σχετικά με τις απειλές και τους κινδύνους

Οι ιδιοκτήτες συστημάτων και τα ανώτερα στελέχη κατανοούν τις απειλές για την ασφάλεια της AI και τους μετριασμούς τους. Οι επιστήμονες και οι προγραμματιστές δεδομένων σας διατηρούν επίγνωση των σχετικών απειλών ασφαλείας και τρόπων αποτυχίας και βοηθούν τους κατόχους κινδύνων να λαμβάνουν τεκμηριωμένες αποφάσεις. Παρέχετε στους χρήστες καθοδήγηση σχετικά με τους μοναδικούς κινδύνους ασφαλείας που αντιμετωπίζουν τα συστήματα AI (για παράδειγμα, ως μέρος της τυπικής εκπαίδευσης InfoSec) και εκπαιδεύετε προγραμματιστές σε τεχνικές ασφαλούς κωδικοποίησης και ασφαλείς και υπεύθυνες πρακτικές AI.

Μοντελοποιήστε τις απειλές για το σύστημά σας

Ως μέρος της διαδικασίας διαχείρισης κινδύνου, εφαρμόζετε μια ολιστική διαδικασία για την αξιολόγηση των απειλών για το σύστημά σας, η οποία περιλαμβάνει την κατανόηση των πιθανών επιπτώσεων στο σύστημα, τους χρήστες, τους οργανισμούς και την ευρύτερη κοινωνία, εάν ένα στοιχείο AI παραβιαστεί ή συμπεριφερθεί απροσδόκητα[7]. Αυτή η διαδικασία περιλαμβάνει την αξιολόγηση του αντίκτυπου των απειλών που σχετίζονται με την AI[8] και την τεκμηρίωση της λήψης των αποφάσεών σας.

Αναγνωρίζετε ότι η ευαισθησία και οι τύποι δεδομένων που χρησιμοποιούνται στο σύστημά σας ενδέχεται να επηρεάσουν την αξία του και να γίνει στόχος για έναν εισβολέα. Η αξιολόγησή σας θα πρέπει να λάβει υπόψη ότι ορισμένες απειλές ενδέχεται να αυξηθούν καθώς τα συστήματα AI γίνονται όλο και περισσότερο αντιληπτά ως στόχοι υψηλής αξίας και καθώς η ίδια η AI επιτρέπει νέους, αυτοματοποιημένους φορείς επιθέσεων.

Σχεδιάστε το σύστημά σας για ασφάλεια, λειτουργικότητα και απόδοση

Είστε βέβαιοι ότι η εργασία που διαθέτετε αντιμετωπίζεται καταλληλότερα χρησιμοποιώντας AI. Έχοντας προσδιορίσει αυτό, αξιολογείτε την καταλληλότητα των επιλογών σχεδίασής σας ειδικά για την AI. Λαμβάνετε υπόψη το μοντέλο απειλής και τους σχετικούς μετριασμούς ασφαλείας μαζί με τη λειτουργικότητα, την εμπειρία χρήστη, το περιβάλλον εγκατάστασης, την απόδοση, τη διασφάλιση, την επίβλεψη, τις ηθικές και νομικές απαιτήσεις, μεταξύ άλλων. Για παράδειγμα:

- εξετάζετε την ασφάλεια της εφοδιαστικής αλυσίδας όταν επιλέγετε να αναπτύξετε εσωτερικά ή να χρησιμοποιήσετε εξωτερικά εξαρτήματα, για παράδειγμα:
    - η επιλογή σας να εκπαιδεύσετε ένα νέο μοντέλο, να χρησιμοποιήσετε ένα υπάρχον μοντέλο (με ή χωρίς βελτίωση) ή να αποκτήσετε πρόσβαση σε ένα μοντέλο μέσω ενός εξωτερικού API είναι κατάλληλη για τις απαιτήσεις σας
    - η επιλογή σας να συνεργαστείτε με έναν εξωτερικό πάροχο μοντέλων περιλαμβάνει μια αξιολόγηση δέουσας επιμέλειας της στάσης ασφαλείας του ίδιου του παρόχου
    - εάν χρησιμοποιείτε εξωτερική βιβλιοθήκη, ολοκληρώνετε μια αξιολόγηση δέουσας επιμέλειας (για παράδειγμα, για να βεβαιωθείτε ότι η βιβλιοθήκη διαθέτει στοιχεία ελέγχου που εμποδίζουν το σύστημα να φορτώνει μη αξιόπιστα μοντέλα χωρίς να εκτίθενται αμέσως σε αυθαίρετη εκτέλεση κώδικα[9])
    - εφαρμόζετε σάρωση και απομόνωση/περιβάλλον δοκιμών κατά την εισαγωγή μοντέλων τρίτων ή σειριακών βαρών, τα οποία θα πρέπει να αντιμετωπίζονται ως μη αξιόπιστος κώδικας τρίτου κατασκευαστή και θα μπορούσαν να ενεργοποιήσουν την απομακρυσμένη εκτέλεση κώδικα

- εάν χρησιμοποιείτε εξωτερικά API, εφαρμόζετε κατάλληλα στοιχεία ελέγχου σε δεδομένα που μπορούν να σταλούν σε υπηρεσίες εκτός του ελέγχου του οργανισμού σας, όπως να απαιτείται από τους χρήστες να συνδεθούν και να επιβεβαιώσουν πριν από την αποστολή δυνητικά ευαίσθητων πληροφοριών
- εφαρμόζετε κατάλληλους ελέγχους και απολύμανση δεδομένων και εισροών· αυτό περιλαμβάνει, κατά την ενσωμάτωση σχολίων χρήστη ή δεδομένων συνεχούς μάθησης στο μοντέλο σας, την αναγνώριση ότι τα δεδομένα εκπαίδευσης καθορίζουν τη συμπεριφορά του συστήματος

- ενσωματώνετε την ανάπτυξη συστημάτων λογισμικού AI στις υπάρχουσες βέλτιστες πρακτικές ασφαλούς ανάπτυξης και λειτουργίας· όλα τα στοιχεία του συστήματος AI είναι γραμμένα σε κατάλληλα περιβάλλοντα χρησιμοποιώντας πρακτικές κωδικοποίησης και γλώσσες που μειώνουν ή εξαλείφουν γνωστές κατηγορίες τρωτών σημείων όπου είναι εύλογο

- εάν τα στοιχεία AI πρέπει να ενεργοποιήσουν ενέργειες, για παράδειγμα τροποποίηση αρχείων ή κατευθύνσεις εξόδου σε εξωτερικά συστήματα, εφαρμόζετε κατάλληλους περιορισμούς στις πιθανές ενέργειες (περιλαμβάνονται εξωτερικά AI και ασφάλειες αστοχίας χωρίς AI εάν είναι απαραίτητο)

- οι αποφάσεις σχετικά με την αλληλεπίδραση των χρηστών ενημερώνονται από κινδύνους που σχετίζονται με την AI, για παράδειγμα:

  - το σύστημά σας παρέχει στους χρήστες αξιοποιήσιμα αποτελέσματα χωρίς να αποκαλύπτει περιττά επίπεδα λεπτομέρειας σε έναν πιθανό εισβολέα
  - εάν είναι απαραίτητο, το σύστημά σας παρέχει αποτελεσματικά προστατευτικά κιγκλιδώματα γύρω από τις εξόδους του μοντέλου
  - εάν προσφέρετε ένα API σε εξωτερικούς πελάτες ή συνεργάτες, εφαρμόζετε κατάλληλα στοιχεία ελέγχου που μετριάζουν τις επιθέσεις στο σύστημα AI μέσω του API
  - ενσωματώνετε τις πιο ασφαλείς ρυθμίσεις στο σύστημα από προεπιλογή
  - εφαρμόζετε τις αρχές των ελάχιστων προνομίων για να περιορίσετε την πρόσβαση στη λειτουργικότητα ενός συστήματος
  - εξηγείτε πιο επικίνδυνες δυνατότητες στους χρήστες και ζητάτε από τους χρήστες να επιλέξουν να τις χρησιμοποιούν· ανακοινώνετε περιπτώσεις απαγορευμένης χρήσης και, όπου είναι δυνατόν, ενημερώνετε τους χρήστες για εναλλακτικές λύσεις

Λάβετε υπόψη τα οφέλη ασφαλείας και τους συμβιβασμούς όταν επιλέγετε το μοντέλο σας AI

Η επιλογή του μοντέλου AI θα περιλαμβάνει την εξισορρόπηση μιας σειράς απαιτήσεων. Αυτό περιλαμβάνει την επιλογή της αρχιτεκτονικής του μοντέλου, της διαμόρφωσης, των δεδομένων εκπαίδευσης, του αλγορίθμου εκπαίδευσης και των υπερπαραμέτρων. Οι αποφάσεις σας ενημερώνονται από το μοντέλο απειλής σας και επανεκτιμώνται τακτικά καθώς εξελίσσεται η έρευνα για την ασφάλεια της AI και η κατανόηση της απειλής.

Όταν επιλέγετε ένα μοντέλο AI, οι σκέψεις σας πιθανότατα θα περιλαμβάνουν, αλλά δεν περιορίζονται:

- στην πολυπλοκότητα του μοντέλου που χρησιμοποιείτε, δηλαδή την επιλεγμένη αρχιτεκτονική και τον αριθμό των παραμέτρων· η επιλεγμένη αρχιτεκτονική του μοντέλου σας και ο αριθμός των παραμέτρων θα επηρεάσουν, μεταξύ άλλων παραγόντων, πόσα δεδομένα εκπαίδευσης απαιτεί και πόσο ανθεκτικό είναι στις αλλαγές στα δεδομένα εισόδου όταν χρησιμοποιείται

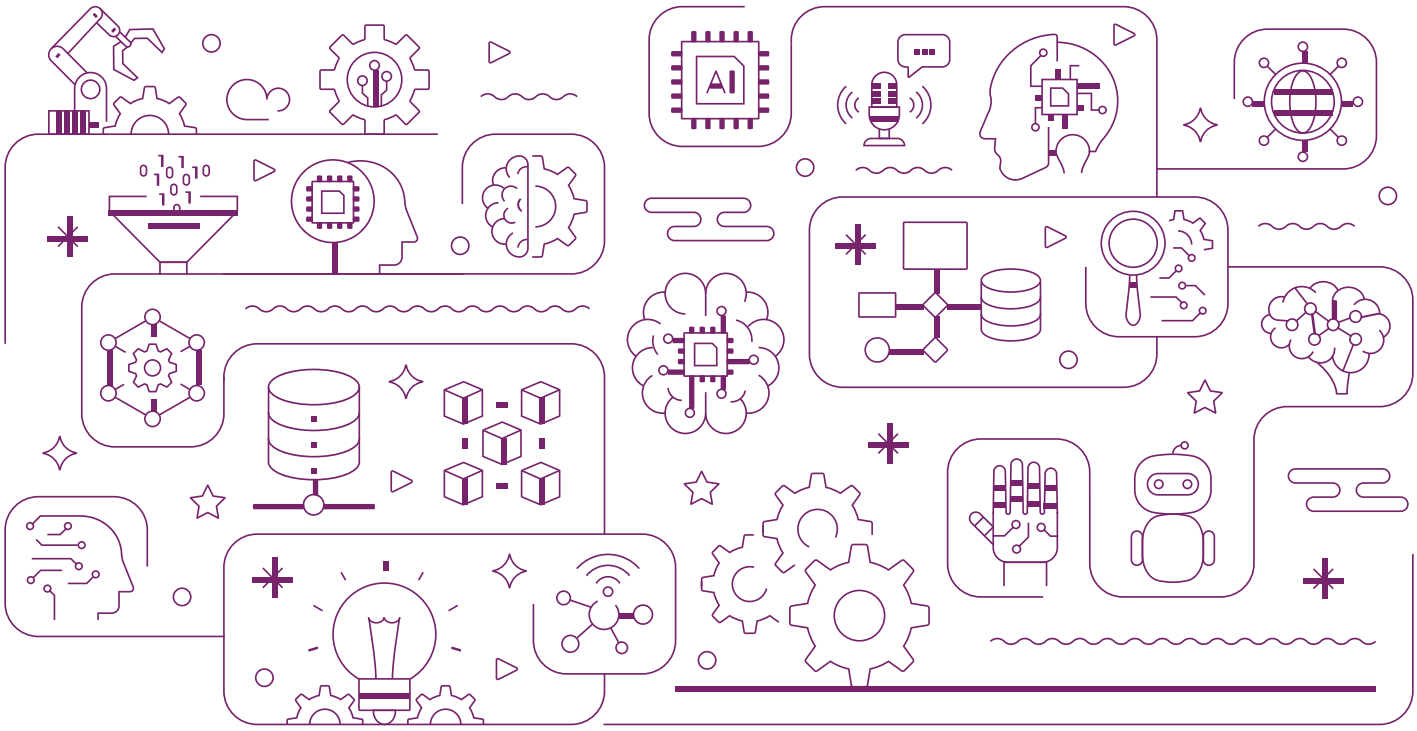
- στην καταλληλότητα του μοντέλου για την περίπτωση χρήσης σας ή/και τη σκοπιμότητα προσαρμογής του στις συγκεκριμένες ανάγκες σας (για παράδειγμα με μικρορύθμιση)

- στη δυνατότητα ευθυγράμμισης, ερμηνείας και επεξήγησης των αποτελεσμάτων του μοντέλου σας (για παράδειγμα για εντοπισμό σφαλμάτων, έλεγχος ή συμμόρφωση με τους κανονισμούς). Μπορεί να υπάρχουν οφέλη από τη χρήση απλούστερων, πιο διαφανών μοντέλων έναντι μεγάλων και πολύπλοκων μοντέλων που είναι πιο δύσκολο να ερμηνευτούν

- στα χαρακτηριστικά των συνόλων δεδομένων εκπαίδευσης, συμπεριλαμβανομένου του μεγέθους, της ακεραιότητας, της ποιότητας, της ευαισθησίας, της ηλικίας, συνάφεια και ποικιλομορφία

- στην αξία της χρήσης τεχνικών σκλήρυνσης μοντέλων (όπως η αντίπαλη εκπαίδευση), τακτοποίησης και/ή τεχνικών ενίσχυσης απορρήτου
- στην προέλευση και τις αλυσίδες εφοδιασμού των εξαρτημάτων, συμπεριλαμβανομένου του μοντέλου ή του μοντέλου θεμελίωσης, των δεδομένων εκπαίδευσης και των σχετικών εργαλείων

Για περισσότερες πληροφορίες σχετικά με το πόσοι από αυτούς τους παράγοντες επηρεάζουν τα αποτελέσματα ασφάλειας, ανατρέξτε στις «Αρχές για την ασφάλεια της μηχανικής μάθησης» του NCSC, ειδικότερα Design for security (model architecture) [Σχεδίαση για ασφάλεια (αρχιτεκτονική μοντέλου)].



2. Ασφαλής ανάπτυξη

Αυτή η ενότητα περιέχει οδηγίες που ισχύουν για το στάδιο **ανάπτυξης** του κύκλου ζωής ανάπτυξης του συστήματος AI, συμπεριλαμβανομένης της ασφάλειας της εφοδιαστικής αλυσίδας, της τεκμηρίωσης και της διαχείρισης περιουσιακών στοιχείων και τεχνικού χρέους.

Ασφαλίστε την αλυσίδα εφοδιασμού σας 

Αξιολογείτε και παρακολουθείτε την ασφάλεια των αλυσίδων εφοδιασμού με AI κατά τη διάρκεια του κύκλου ζωής ενός συστήματος και απαιτείτε από τους προμηθευτές να συμμορφώνονται με τα ίδια πρότυπα που εφαρμόζει ο οργανισμός σας σε άλλο λογισμικό. Εάν οι προμηθευτές δεν μπορούν να συμμορφωθούν με τα πρότυπα του οργανισμού σας, ενεργείτε σύμφωνα με τις υπάρχουσες πολιτικές διαχείρισης κινδύνου.

Όπου δεν παράγονται εσωτερικά, αποκτάτε και διατηρείτε καλά ασφαλισμένα και καλά τεκμηριωμένα στοιχεία υλικού και λογισμικού (για παράδειγμα, μοντέλα, δεδομένα, βιβλιοθήκες λογισμικού, λειτουργικές μονάδες, ενδιάμεσο λογισμικό, πλαίσια και εξωτερικά API) από επαληθευμένα εμπορικά, ανοιχτού κώδικα, και άλλους τρίτους προγραμματιστές για τη διασφάλιση ισχυρής ασφάλειας στα συστήματά σας.

Είστε έτοιμοι να μεταβείτε με την εφεδρική λειτουργία ανακατεύθυνσης σε εναλλακτικές λύσεις για κρίσιμα για την αποστολή συστήματα, εάν δεν πληρούνται τα κριτήρια ασφαλείας. Χρησιμοποιείτε πόρους όπως το Supply Chain Guidance (Καθοδήγηση σχετικά με την Αλυσίδα Εφοδιασμού) του NCSC και πλαίσια όπως τα Supply Chain Levels for Software Artifacts (SLSA)[10] για την παρακολούθηση βεβαιώσεων των κύκλων ζωής της εφοδιαστικής αλυσίδας και της ανάπτυξης λογισμικού.

Προσδιορίστε, παρακολουθήστε και προστατέψτε τα περιουσιακά σας στοιχεία 

Κατανοείτε την αξία για τον οργανισμό σας των στοιχείων που σχετίζονται με την AI, συμπεριλαμβανομένων μοντέλων, δεδομένων (συμπεριλαμβανομένων των σχολίων των χρηστών), προτροπών, λογισμικού, τεκμηρίωσης, αρχείων καταγραφής και αξιολογήσεων (συμπεριλαμβανομένων πληροφοριών σχετικά με δυνητικά μη ασφαλείς δυνατότητες και τρόπους αποτυχίας), αναγνωρίζοντας πότε αντιπροσωπεύουν σημαντικές επενδύσεις και πότε η πρόσβαση σε αυτές επιτρέπει έναν εισβολέα. Αντιμετωπίζετε τα αρχεία καταγραφής ως ευαίσθητα δεδομένα και εφαρμόζετε στοιχεία ελέγχου για την προστασία της εμπιστευτικότητας, της ακεραιότητας και της διαθεσιμότητάς τους.

Γνωρίζετε πού βρίσκονται τα περιουσιακά σας στοιχεία και έχετε αξιολογήσει και αποδεχτεί τυχόν σχετικούς κινδύνους. Έχετε διαδικασίες και εργαλεία για την παρακολούθηση, τον έλεγχο ταυτότητας, τον έλεγχο έκδοσης και την προστασία των στοιχείων σας και μπορείτε να τα επαναφέρετε σε μια γνωστή καλή κατάσταση σε περίπτωση παραβίασης.

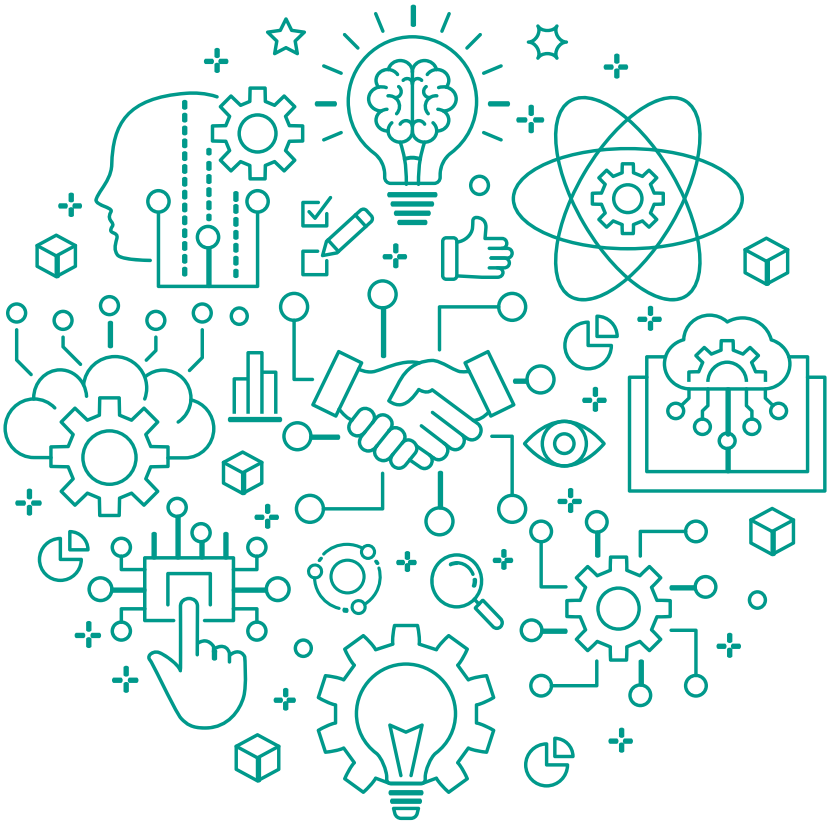
Διαθέτετε διεργασίες και ελέγχους για τη διαχείριση των δεδομένων στα οποία μπορούν να έχουν πρόσβαση τα συστήματα AI και για τη διαχείριση περιεχομένου που δημιουργείται από την AI σύμφωνα με την ευαισθησία του (και την ευαισθησία των εισροών που χρησιμοποιήθηκαν για τη δημιουργία του).

Τεκμηριώστε τα δεδομένα, τα μοντέλα και τις προτροπές σας 

Τεκμηριώνετε τη δημιουργία, τη λειτουργία και τη διαχείριση του κύκλου ζωής οποιωνδήποτε μοντέλων, συνόλων δεδομένων και προτροπών μετα-ή συστήματος. Η τεκμηρίωσή σας περιλαμβάνει πληροφορίες σχετικές με την ασφάλεια, όπως πηγές δεδομένων εκπαίδευσης (συμπεριλαμβανομένης της λεπτομερούς ρύθμισης των δεδομένων και ανθρώπινης ή άλλης επιχειρησιακής ανατροφοδότησης), προβλεπόμενο εύρος και περιορισμούς, προστατευτικά κιγκλιδώματα, κρυπτογραφικές κατακερματίσεις ή υπογραφές, χρόνο διατήρησης, προτεινόμενη συχνότητα ελέγχου και πιθανούς τρόπους αποτυχίας . Χρήσιμες δομές για να γίνει αυτό περιλαμβάνουν κάρτες μοντέλων, κάρτες δεδομένων και λογαριασμούς υλικών λογισμικού (SBOM). Η παραγωγή ολοκληρωμένης τεκμηρίωσης υποστηρίζει τη διαφάνεια και τη λογοδοσία[11].

Διαχειριστείτε το τεχνικό σας χρέος

Όπως με κάθε σύστημα λογισμικού, προσδιορίζετε, παρακολουθείτε και διαχειρίζεστε το «τεχνικό χρέος» σας καθ' όλη τη διάρκεια του κύκλου ζωής ενός συστήματος AI (το τεχνικό χρέος είναι το σημείο όπου λαμβάνονται αποφάσεις μηχανικής που υπολείπονται των βέλτιστων πρακτικών για την επίτευξη βραχυπρόθεσμων αποτελεσμάτων, σε βάρος των μακροπρόθεσμων οφελών). Όπως και το χρηματοοικονομικό χρέος, το τεχνικό χρέος δεν είναι εγγενώς κακό, αλλά θα πρέπει να αντιμετωπίζεται από τα πρώτα στάδια ανάπτυξης[12]. Αναγνωρίζετε ότι κάτι τέτοιο μπορεί να είναι πιο δύσκολο σε ένα πλαίσιο AI από ό,τι για το τυπικό λογισμικό και ότι τα επίπεδα τεχνικού χρέους σας είναι πιθανό να είναι υψηλά λόγω των ταχέων κύκλων ανάπτυξης και της έλλειψης καθιερωμένων πρωτοκόλλων και διεπαφών. Διασφαλίζετε ότι τα σχέδια κύκλου ζωής σας (συμπεριλαμβανομένων των διαδικασιών για τον παροπλισμό συστημάτων AI) αξιολογούν, αναγνωρίζουν και μετριάζουν τους κινδύνους για μελλοντικά παρόμοια συστήματα.



3. Ασφαλής εγκατάσταση

Αυτή η ενότητα περιέχει οδηγίες που ισχύουν για το στάδιο **εγκατάστασης** του κύκλου ζωής ανάπτυξης συστήματος AI, συμπεριλαμβανομένης της προστασίας της υποδομής και των μοντέλων από παραβίαση, απειλή ή απώλεια, ανάπτυξη διαδικασιών διαχείρισης συμβάντων και υπεύθυνη γνωστοποίηση.

Ασφαλίστε την υποδομή σας

Εφαρμόζετε καλές αρχές ασφάλειας υποδομής στην υποδομή που χρησιμοποιείται σε κάθε μέρος του κύκλου ζωής του συστήματός σας. Εφαρμόζετε κατάλληλους ελέγχους πρόσβασης στα API, τα μοντέλα και τα δεδομένα σας, καθώς και στους αγωγούς εκπαίδευσης και επεξεργασίας τους, στην έρευνα και ανάπτυξη καθώς και στην εγκατάσταση. Αυτό περιλαμβάνει κατάλληλο διαχωρισμό περιβαλλόντων που περιέχουν ευαίσθητο κώδικα ή δεδομένα. Αυτό θα βοηθήσει επίσης στον μετριασμό των τυπικών επιθέσεων ασφάλειας στον κυβερνοχώρο που στοχεύουν να κλέψουν ένα μοντέλο ή να βλάψουν την απόδοσή του.

Προστατεύετε το μοντέλο σας συνεχώς

Οι εισβολείς ενδέχεται να είναι σε θέση να ανασυνθέσουν τη λειτουργικότητα ενός μοντέλου[13] ή τα δεδομένα στα οποία εκπαιδεύτηκε[14] αποκτώντας απευθείας πρόσβαση σε ένα μοντέλο (με την απόκτηση βαρών μοντέλου) ή έμμεσα (με ερώτημα στο μοντέλο μέσω εφαρμογής ή υπηρεσίας). Οι εισβολείς μπορεί επίσης να παραβιάσουν μοντέλα, δεδομένα ή προτροπές κατά τη διάρκεια ή μετά την εκπαίδευση, καθιστώντας το αποτέλεσμα αναξιόπιστο.

Προστατεύετε το μοντέλο και τα δεδομένα από άμεση και έμμεση πρόσβαση, αντίστοιχα, με:

- την εφαρμογή τυπικών βέλτιστων πρακτικών ασφάλειας στον κυβερνοχώρο
- εφαρμογή ελέγχων στη διεπαφή ερωτήματος για τον εντοπισμό και την αποτροπή προσπαθειών πρόσβασης, τροποποίησης και διείσδυσης εμπιστευτικών πληροφοριών

Για να διασφαλίσετε ότι τα καταναλωτικά συστήματα μπορούν να επικυρώσουν μοντέλα, υπολογίζετε και μοιράζεστε κρυπτογραφικούς κατακερματισμούς και/ή υπογραφές αρχείων μοντέλων (για παράδειγμα, βάρη μοντέλων) και συνόλων δεδομένων (συμπεριλαμβανομένων των σημείων ελέγχου) αμέσως μόλις το μοντέλο εκπαιδευτεί. Όπως πάντα με την κρυπτογραφία, η καλή διαχείριση κλειδιών είναι απαραίτητη[15].

Η προσέγγισή σας για τον μετριασμό του κινδύνου εμπιστευτικότητας θα εξαρτηθεί σημαντικά από την περίπτωση χρήσης και το μοντέλο απειλής. Ορισμένες εφαρμογές, για παράδειγμα εκείνες που αφορούν πολύ ευαίσθητα δεδομένα, ενδέχεται να απαιτούν θεωρητικές εγγυήσεις που μπορεί να είναι δύσκολο ή δαπανηρό να εφαρμοστούν. Εάν ενδείκνυται, τεχνολογίες ενίσχυσης της ιδιωτικής ζωής (όπως το διαφορικό απόρρητο ή η ομομορφική κρυπτογράφηση) μπορούν να χρησιμοποιηθούν για τη διερεύνηση ή τη διασφάλιση των επιπέδων κινδύνου που σχετίζονται με καταναλωτές, χρήστες και εισβολείς που έχουν πρόσβαση σε μοντέλα και αποτελέσματα.

Ανάπτυξη διαδικασιών διαχείρισης συμβάντων

Το αναπόφευκτο των συμβάντων ασφαλείας που επηρεάζουν τα συστήματα AI σας, αντικατοπτρίζεται στα σχέδια απόκρισης, κλιμάκωσης και αποκατάστασης περιστατικών. Τα σχέδιά σας αντικατοπτρίζουν διαφορετικά σενάρια και επανεκτιμώνται τακτικά καθώς εξελίσσεται το σύστημα και η ευρύτερη έρευνα. Αποθηκεύετε σημαντικούς ψηφιακούς πόρους της εταιρείας σε αντίγραφα ασφαλείας εκτός σύνδεσης. Οι ανταποκριτές έχουν εκπαιδευτεί για να αξιολογούν και να αντιμετωπίζουν περιστατικά που σχετίζονται με την AI. Παρέχετε αρχεία καταγραφής ελέγχου υψηλής ποιότητας και άλλα χαρακτηριστικά ασφαλείας ή πληροφορίες σε πελάτες και χρήστες χωρίς επιπλέον χρέωση, για να ενεργοποιήσετε τις διαδικασίες απόκρισης συμβάντων.

Κυκλοφορήστε την AI υπεύθυνα

Κυκλοφορείτε μοντέλα, εφαρμογές ή συστήματα μόνο αφού τα υποβάλετε σε κατάλληλη και αποτελεσματική αξιολόγηση ασφαλείας, όπως συγκριτική αξιολόγηση και κόκκινη ομαδοποίηση (καθώς και άλλες δοκιμές που δεν εμπίπτουν στο πεδίο εφαρμογής αυτών των κατευθυντήριων γραμμών, όπως ασφάλεια ή δικαιοσύνη) και είστε σαφείς στους χρήστες σας σχετικά με γνωστούς περιορισμούς ή πιθανούς τρόπους αποτυχίας. Λεπτομέρειες για τις βιβλιοθήκες δοκιμών ασφαλείας ανοιχτού κώδικα παρέχονται στην ενότητα περαιτέρω ανάγνωσης στο τέλος αυτού του εγγράφου.

Διευκολύνετε τους χρήστες να κάνουν τα σωστά πράγματα

Αναγνωρίζετε ότι κάθε νέα ρύθμιση ή επιλογή διαμόρφωσης πρέπει να αξιολογείται σε συνδυασμό με το επιχειρηματικό όφελος που αποκομίζει και τυχόν κινδύνους ασφαλείας που εισάγει. Στην ιδανική περίπτωση, η πιο ασφαλής ρύθμιση θα ενσωματωθεί στο σύστημα ως η μόνη επιλογή. Όταν η διαμόρφωση είναι απαραίτητη, η προεπιλεγμένη επιλογή θα πρέπει να είναι γενικά ασφαλής έναντι κοινών απειλών (δηλαδή, ασφαλής από προεπιλογή). Εφαρμόζετε στοιχεία ελέγχου για να αποτρέψετε τη χρήση ή την εγκατάσταση του συστήματός σας με κακόβουλους τρόπους.

Παρέχετε στους χρήστες καθοδήγηση σχετικά με την κατάλληλη χρήση του μοντέλου ή του συστήματός σας, η οποία περιλαμβάνει την επισήμανση περιορισμών και πιθανούς τρόπους αποτυχίας. Δηλώνετε ξεκάθαρα στους χρήστες για ποιες πτυχές ασφάλειας είναι υπεύθυνοι και είστε διαφανείς σχετικά με το πού (και πώς) τα δεδομένα τους μπορούν να χρησιμοποιηθούν, να προσπελαστούν ή να αποθηκευτούν (για παράδειγμα, εάν χρησιμοποιούνται για επανεκπαίδευση μοντέλων ή εξετάζονται από υπαλλήλους ή συνεργάτες).

4. Ασφαλής λειτουργία και συντήρηση

Αυτή η ενότητα περιέχει οδηγίες που ισχύουν για το στάδιο **ασφαλούς λειτουργίας και συντήρησης** του κύκλου ζωής ανάπτυξης συστήματος AI . Παρέχει κατευθυντήριες γραμμές σχετικά με ενέργειες που είναι ιδιαίτερα σημαντικές όταν έχει εγκατασταθεί ένα σύστημα, συμπεριλαμβανομένης της καταγραφής και παρακολούθησης, της διαχείρισης ενημερώσεων και της κοινής χρήσης πληροφοριών.

Παρακολουθήστε τη συμπεριφορά του συστήματός σας

Μετράτε τα αποτελέσματα και την απόδοση του μοντέλου και του συστήματός σας έτσι ώστε να μπορείτε να παρατηρήσετε ξαφνικές και σταδιακές αλλαγές στη συμπεριφορά που επηρεάζουν την ασφάλεια. Μπορείτε να λάβετε υπόψη και να εντοπίσετε πιθανές εισβολές και παραβιάσεις, καθώς και φυσική μετατόπιση δεδομένων.

Παρακολουθήστε τις εισόδους του συστήματός σας

Σύμφωνα με τις απαιτήσεις απορρήτου και προστασίας δεδομένων, παρακολουθείτε και καταγράφετε εισόδους στο σύστημά σας (όπως αιτήματα συμπερασμάτων, ερωτήματα ή προτροπές) για να ενεργοποιήσετε τις υποχρεώσεις συμμόρφωσης, τον έλεγχο, την έρευνα και την αποκατάσταση σε περίπτωση παραβίασης ή κακής χρήσης. Αυτό θα μπορούσε να περιλαμβάνει ρητή ανίχνευση εκτός διανομής ή/και αντίθετες εισόδους, συμπεριλαμβανομένων εκείνων που στοχεύουν στην εκμετάλλευση των βημάτων προετοιμασίας δεδομένων (όπως περικοπή και αλλαγή μεγέθους για εικόνες).

Ακολουθήστε μια secure by design (ασφαλή από τον σχεδιασμό) προσέγγιση στις ενημερώσεις

Περιλαμβάνετε αυτοματοποιημένες ενημερώσεις από προεπιλογή σε κάθε προϊόν και χρησιμοποιείτε ασφαλείς, αρθρωτές διαδικασίες ενημέρωσης για τη διανομή τους. Οι διαδικασίες ενημέρωσης (συμπεριλαμβανομένων των καθεστώτων δοκιμών και αξιολόγησης) αντικατοπτρίζουν το γεγονός ότι οι αλλαγές σε δεδομένα, μοντέλα ή μηνύματα μπορεί να οδηγήσουν σε αλλαγές στη συμπεριφορά του συστήματος (για παράδειγμα, αντιμετωπίζετε τις σημαντικές ενημερώσεις σαν νέες εκδόσεις). Υποστηρίζετε τους χρήστες να αξιολογούν και να ανταποκρίνονται σε αλλαγές μοντέλου (για παράδειγμα παρέχοντας πρόσβαση σε προεπισκόπηση και API με έκδοση).

Συλλέξτε και μοιραστείτε διδάγματα

Συμμετέχετε σε κοινότητες ανταλλαγής πληροφοριών, συνεργαζόμενοι σε ολόκληρο το παγκόσμιο οικοσύστημα της βιομηχανίας, της ακαδημαϊκής κοινότητας και των κυβερνήσεων για να μοιραστείτε τις βέλτιστες πρακτικές όπως αρμόζει. Διατηρείτε ανοιχτές γραμμές επικοινωνίας για σχόλια σχετικά με την ασφάλεια του συστήματος, τόσο εσωτερικά όσο και εξωτερικά στον οργανισμό σας, συμπεριλαμβανομένης της παροχής συναίνεσης σε ερευνητές ασφάλειας για έρευνα και αναφορά τρωτών σημείων. Όταν χρειάζεται, κλιμακώνετε ζητήματα στην ευρύτερη κοινότητα, για παράδειγμα δημοσιεύοντας ενημερωτικά δελτία που ανταποκρίνονται σε αποκαλύψεις ευπάθειας, συμπεριλαμβανομένης της λεπτομερούς και πλήρους απαρίθμησης κοινών τρωτών σημείων. Αναλαμβάνετε μέτρα για τον μετριασμό και την αποκατάσταση προβλημάτων γρήγορα και κατάλληλα.

Περαιτέρω ανάγνωση

Ανάπτυξη AI

Principles for the security of machine learning (Αρχές για την ασφάλεια της μηχανικής μάθησης)
Οι λεπτομερείς οδηγίες του NCSC σχετικά με την ανάπτυξη, την εγκατάσταση ή τη λειτουργία ενός συστήματος με στοιχείο ML.

Secure by Design - Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software (Ασφάλεια από τον σχεδιασμό - Μεταβολή της ισορροπίας του κινδύνου κυβερνοασφάλειας: Αρχές και προσεγγίσεις για ασφαλές λογισμικό από τον σχεδιασμό)
Συγγραφή από την CISA, το NCSC και άλλους φορείς, αυτή η καθοδήγηση περιγράφει τον τρόπο με τον οποίο οι κατασκευαστές συστημάτων λογισμικού, συμπεριλαμβανομένης της AI, πρέπει να λαμβάνουν μέτρα για να συνυπολογίσουν την ασφάλεια στο στάδιο σχεδιασμού της ανάπτυξης προϊόντος και να αποστέλλουν προϊόντα που βγαίνουν με ασφάλεια από το κουτί.

AI Security Concerns in a Nutshell (Ανησυχίες για την ασφάλεια της AI με λίγα λόγια)
Παρήχθη από την Ομοσπονδιακή Υπηρεσία Ασφάλειας Πληροφοριών της Γερμανίας (BSI), αυτό το έγγραφο παρέχει μια εισαγωγή σε πιθανές επιθέσεις σε συστήματα μηχανικής μάθησης και πιθανές άμυνες έναντι αυτών των επιθέσεων.

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems (Διεθνείς κατευθυντήριες αρχές της διαδικασίας της Χιροσίμα για οργανισμούς που αναπτύσσουν προηγμένα συστήματα AI) και Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (Διεθνής κώδικας συμπεριφοράς της διαδικασίας της Χιροσίμα για οργανισμούς που αναπτύσσουν προηγμένα συστήματα AI)
Αυτά τα έγγραφα, που παράχθηκαν ως μέρος της Διαδικασίας AI της Χιροσίμα από την G7, παρέχουν καθοδήγηση σε οργανισμούς που αναπτύσσουν τα πιο προηγμένα συστήματα AI, συμπεριλαμβανομένων των πιο προηγμένων μοντέλων θεμελίωσης και παραγωγικών συστημάτων AI με στόχο την προώθηση ασφαλούς και αξιόπιστης AI παγκοσμίως.

AI Verify (Επαλήθευση AI)
Πλαίσιο δοκιμών διακυβέρνησης AI και εργαλειοθήκη λογισμικού της Σιγκαπούρης που επικυρώνει την απόδοση των συστημάτων AI έναντι ενός συνόλου διεθνώς αναγνωρισμένων αρχών μέσω τυποποιημένων δοκιμών.

Multilayer Framework for Good Cybersecurity Practices for AI (Πολυεπίπεδο πλαίσιο για καλές πρακτικές κυβερνοασφάλειας για AI) - ENISA (europa.eu)
Ένα πλαίσιο για την καθοδήγηση των εθνικών αρμόδιων αρχών και των ενδιαφερόμενων μερών της AI σχετικά με τα βήματα που πρέπει να ακολουθήσουν για να διασφαλίσουν τα συστήματά τους AI, λειτουργίες και διαδικασίες.

ISO 5338: AI system life cycle processes (Under review) [Διαδικασίες κύκλου ζωής συστήματος AI (Υπό εξέταση)]
Ένα σύνολο διαδικασιών και σχετικών εννοιών για την περιγραφή του κύκλου ζωής των συστημάτων AI που βασίζονται στη μηχανική μάθηση και τα ευρετικά συστήματα.

AI Cloud Service Compliance Criteria Catalogue (AIC4) [Κατάλογος κριτηρίων συμμόρφωσης υπηρεσιών AI Cloud (AIC4)]
Ο κατάλογος κριτηρίων συμμόρφωσης υπηρεσίας AI Cloud της BSI παρέχει κριτήρια ειδικά για την AI, τα οποία επιτρέπουν την αξιολόγηση της ασφάλειας μιας υπηρεσίας AI σε όλο τον κύκλο ζωής της.

NIST IR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning [NIST IR 8269 (Προσχέδιο) Μια ταξονομία και ορολογία της Αντίπαλης Τεχνητής Μάθησης]
Ένα σύνολο διαδικασιών και σχετικών εννοιών για την περιγραφή του κύκλου ζωής των συστημάτων AI που βασίζονται στη μηχανική μάθηση και τα ευρετικά συστήματα.

MITRE ATLAS
Μια βάση γνώσεων αντιπάλων τακτικών, τεχνικών και περιπτωσιολογικών μελετών για συστήματα μηχανικής μάθησης (ML), διαμορφωμένη μετά από μοντέλο και συνδεδεμένη με το πλαίσιο MITER ATT&CK.

An Overview of Catastrophic AI Risks (2023) [Μία επισκόπηση των καταστροφικών κινδύνων AI (2023)]
Το παρόν έγγραφο που δημιουργήθηκε από το Center for AI Safety, καθορίζει τομείς κινδύνου που ενέχει η AI.

Large Language Models: Opportunities and Risks for Industry and Authorities (Μοντέλα μεγάλων γλωσσών: Ευκαιρίες και κίνδυνοι για τη βιομηχανία και τις αρχές)
Έγγραφο που παράγεται από την BSI για εταιρείες, αρχές και προγραμματιστές που θέλουν να μάθουν περισσότερα σχετικά με τις ευκαιρίες και τους κινδύνους ανάπτυξης, εγκατάστασης και/ή χρήσης LLM.

Έργα ανοιχτού κώδικα για να βοηθήσουν τους χρήστες να δοκιμάσουν μοντέλα AI για την ασφάλεια περιλαμβάνουν:

- Adversarial Robustness Toolbox (IBM)
- CleverHans (Πανεπιστήμιο του Τορόντο)
- TextAttack (Πανεπιστήμιο της Βιρτζίνια)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft)
- AI Verify (Infocomm Media Development Authority, Σιγκαπούρη)

Κυβερνοασφάλεια

CISA's Cybersecurity Performance Goals (Στόχοι απόδοσης κυβερνοασφάλειας της CISA)
Ένα κοινό σύνολο προστατευτικών μέτρων που θα πρέπει να εφαρμόσουν όλες οι οντότητες υποδομών ζωτικής σημασίας για να μειώσουν ουσιαστικά την πιθανότητα και τον αντίκτυπο γνωστών κινδύνων και τεχνικών αντιπάλου.

NCSC CAF Framework (Πλαίσιο NCSC CAF)
Το Πλαίσιο Αξιολόγησης Κυβερνοασφάλειας (CAF) παρέχει καθοδήγηση σε οργανισμούς που είναι υπεύθυνοι για ζωτικής σημασίας υπηρεσίες και δραστηριότητες.

MITRE's Supply Chain Security Framework (Πλαίσιο ασφάλειας εφοδιαστικής αλυσίδας της MITRE)
Ένα πλαίσιο για την αξιολόγηση προμηθευτών και παρόχων υπηρεσιών εντός της αλυσίδας εφοδιασμού.

Διαχείριση κινδύνου

NIST AI Risk Management Framework (AI RMF) [Πλαίσιο Διαχείρισης Κινδύνων NIST AI (AI RMF)]
Το AI RMF περιγράφει τον τρόπο διαχείρισης κοινωνικοτεχνικών κινδύνων για άτομα, οργανισμούς και κοινωνία που συνδέονται μοναδικά με την AI.

ISO 27001: Information security, cybersecurity and privacy protection (Ασφάλεια πληροφοριών, κυβερνοασφάλεια και προστασία του απορρήτου)
Αυτό το πρότυπο παρέχει στους οργανισμούς καθοδήγηση σχετικά με τη δημιουργία, την εφαρμογή και τη συντήρηση ενός συστήματος διαχείρισης ασφάλειας πληροφοριών.

ISO 31000: Risk management (Διαχείριση κινδύνων)
Ένα διεθνές πρότυπο που παρέχει στους οργανισμούς κατευθυντήριες γραμμές και αρχές για τη διαχείριση κινδύνων εντός οργανισμών.

NCSC Risk Management Guidance (Καθοδήγηση διαχείρισης κινδύνων NCSC)
Αυτή η καθοδήγηση βοηθά τους επαγγελματίες που ασχολούνται με την ασφάλεια στον κυβερνοχώρο να κατανοήσουν καλύτερα και να διαχειριστούν τους κινδύνους για την ασφάλεια στον κυβερνοχώρο που επηρεάζουν τους οργανισμούς τους.

Σημειωματάριο

1. Εδώ ορίζεται ως άτομο, δημόσια αρχή, οργανισμός ή άλλος φορέας που αναπτύσσει ένα σύστημα AI (ή που έχει αναπτύξει ένα σύστημα AI) και διαθέτει αυτό το σύστημα στην αγορά ή το θέτει σε λειτουργία με το δικό του όνομα ή εμπορικό σήμα
2. Για περισσότερες πληροφορίες σχετικά με το secure by design, ανατρέξτε στην ιστοσελίδα και τις οδηγίες Secure by Design (Ασφάλεια από τον σχεδιασμό) της CISA Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software (Μεταβολή της ισορροπίας του κινδύνου κυβερνοασφάλειας: Αρχές και προσεγγίσεις για ασφαλές λογισμικό από τον σχεδιασμό)
3. Σε αντίθεση με προσεγγίσεις AI που δεν είναι ML, όπως συστήματα βασισμένα σε κανόνες
4. Το CEPS περιγράφει επτά διαφορετικούς τύπους αλληλεπίδρασης ανάπτυξης AI στη δημοσίευσή του 'Reconciling the AI Value Chain with the EU's Artificial Intelligence Act' («Συμφωνία της αλυσίδας αξίας της AI με τον νόμο της ΕΕ για την τεχνητή νοημοσύνη»)
5. ISO/IEC 22989:2022(en) το ορίζει ως «ένα λειτουργικό στοιχείο που δημιουργεί ένα σύστημα AI»
6. Το NIST είναι επιφορτισμένο με την παραγωγή κατευθυντήριων γραμμών (και τη λήψη άλλων ενεργειών) για την προώθηση της ασφαλούς και αξιόπιστης ανάπτυξης και χρήσης της τεχνητής νοημοσύνης (AI). Δείτε: NIST's Responsibilities Under the October 30, 2023 Executive Order (Ευθύνες του NIST βάσει του εκτελεστικού διατάγματος της 30ης Οκτωβρίου, 2023)
7. Περισσότερες πληροφορίες σχετικά με τη μοντελοποίηση απειλών είναι διαθέσιμες από το OWASP Foundation
8. Δείτε το MITER ATLAS Adversarial Machine Learning 101 (Αντίπαλη Τεχνητή Μάθηση 101)
9. GitHub: RCE PoC for Tensorflow using a malicious Lambda layer (RCE PoC για Tensorflow με χρήση κακόβουλου επιπέδου Λάμδα)
10. SLSA: 'Safeguarding artifact integrity across any software supply chain' («Διασφάλιση της ακεραιότητας των τεχνουργημάτων σε οποιαδήποτε αλυσίδα εφοδιασμού λογισμικού»)
11. METI (Ιαπωνικό Υπουργείο Οικονομίας, Εμπορίου και Βιομηχανίας, 2023), 'Guide of Introduction of Software Bill of Materials (SBOM) for Software Management' («Εισαγωγικός Οδηγός Καταλόγου Υλικών Λογισμικού (SBOM) για Διαχείριση Λογισμικού»)
12. Έρευνα Google: Machine Learning: The High Interest Credit Card of Technical Debt (Μηχανική εκμάθηση: Η Πιστωτική Κάρτα Τεχνικού Χρέους Υψηλού Επιτοκίου)
13. Tramèr et al 2016, Stealing Machine Learning Models via Prediction APIs (Κλοπή μοντέλων μηχανικής μάθησης μέσω API πρόβλεψης)
14. Boenisch, 2020, Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework [Επιθέσεις κατά του απορρήτου της μηχανικής μάθησης (Μέρος 1ο): Επιθέσεις αναστροφής μοντέλων με το Πλαίσιο IBM-ART]
15. Εθνικό Κέντρο Ασφάλειας στον Κυβερνοχώρο, 2020, Design and build a privately hosted Public Key Infrastructure (Σχεδίαση και κατασκευή μιας ιδιωτικά φιλοξενούμενης υποδομής δημόσιου κλειδιού)