“Τεχνητή νοημοσύνη για την κυβερνοασφάλεια” & “Κυβερνοασφάλεια για την τεχνητή νοημοσύνη”

Κίνητρο

Η χρήση συστημάτων βασισμένων στην τεχνητή νοημοσύνη δημιουργεί νέα επιφάνεια επίθεσης για τα συστήματα που κάνουν χρήση τους αλγορίθμους τεχνητή νοημοσύνη / μηχανικής μάθησης. Τα συστατικά που βασίζονται στην τεχνητή νοημοσύνη είναι συχνά «μαύρα κουτιά», γεγονός που τα καθιστά πρωταρχικούς στόχους για τους επιτιθέμενους που έχουν αναπτύξει τεχνικές για να βλάψουν την ανθεκτικότητα των συστημάτων με τεχνητή νοημοσύνη με επιτιθέμενη τεχνητή νοημοσύνη, να παραβιάσουν την ακεραιότητα των μοντέλων τεχνητής νοημοσύνης και να παρακάμψουν ή να απενεργοποιήσουν τα μοντέλα με κακόβουλα ερωτήματα σε αυτά.

Περιγραφή

Στόχος του έργου AIAS είναι ο σχεδιασμός και η ανάπτυξη μιας καινοτόμου πλατφόρμας ασφαλείας για την προστασία των συστημάτων τεχνητής νοημοσύνης με τη χρήση συστημάτων άμυνας για την επιτιθέμενη τεχνητή νοημοσύνη, τεχνικές εξαπάτησης και επεξηγήσιμων μοντέλων τεχνητής νοημοσύνης, ενισχύοντας την ανθεκτικότητα έναντι επιθέσεων στον κυβερνοχώρο.

Δράση

Στόχος του έργου AIAS είναι η διεξαγωγή έρευνας σχετικά με την επιτιθέμενη τεχνητή νοημοσύνη για το σχεδιασμό και την ανάπτυξη μιας καινοτόμου πλατφόρμας ασφαλείας βασισμένης στην τεχνητή νοημοσύνη για την προστασία της ανθεκτικότητας των συστημάτων τεχνητής νοημοσύνης και των βασισμένων στην τεχνητή νοημοσύνη λειτουργιών των οργανισμών, στηριζόμενη σε μεθόδους άμυνας της επιτιθέμενης τεχνητής νοημοσύνης, σε μηχανισμούς εξαπάτησης καθώς και σε επεξηγήσιμες λύσεις τεχνητής νοημοσύνης.

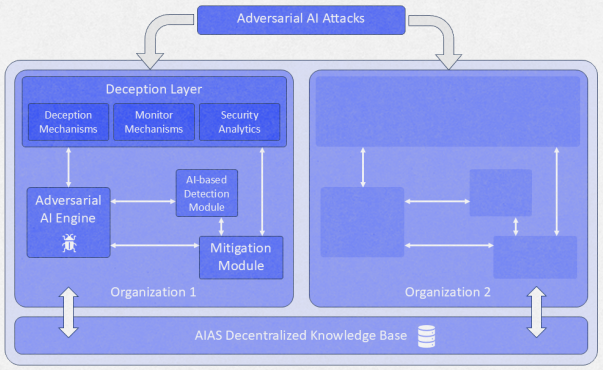
Δεκέμβριος 2024

AIAS

AI-ASSISTED CYBERSECURITY PLATFORM

HORIZON-MSCA-2022-SE-01-01;
HORIZON.1.2 - Marie Skłodowska-Curie Actions (MSCA)

Ιστοσελίδα έργου: https://www.aias-project.eu/
Έναρξη έργου: 1η Ιανουαρίου 2024
Διάρκεια: 48 μήνες
GRANT AGREEMENT: 101131292
EU CONTRIBUTION: EUR 1 564 000
Συντονιστής έργου: Κέντρο Ερευνών Πανεπιστημίου Πειραιώς (ΕΛΛΑΔΑ)



Σημασία της Τεχνητής Νοημοσύνης στο AIAS

Μηχανή επιτιθέμενης τεχνητής νοημοσύνης

Δημιουργία σεναρίων επίθεσης προσαρμοσμένων στα χαρακτηριστικά του οργανισμού.

Generative Adversarial Networks (GANs)

Ανίχνευση επιθέσεων επιτιθέμενης τεχνητής νοημοσύνης.

Επεξηγήσιμη τεχνητή νοημοσύνη

Μηχανή συστάσεων για περαιτέρω μετριασμό των επιθέσεων και κατανόηση των αποφάσεων που λαμβάνει η τεχνητή νοημοσύνη.

Στόχοι

Ολιστική προστασία

Σύλληψη και ανάπτυξη μιας αρχιτεκτονικής υπηρεσιών που θα ενσωματώνει εφαρμογές με τεχνητή νοημοσύνη, μηχανισμούς εξαπάτησης και τεχνικές μετριασμού για την ολιστική προστασία των οργανισμών από κυβερνοεπιθέσεις και την επιτιθέμενη τεχνητή νοημοσύνη.

Σενάρια επίθεσης

Σχεδιασμός και ανάπτυξη μιας νέας μηχανής τεχνητής νοημοσύνης για τη δημιουργία σεναρίων επίθεσης προσαρμοσμένων στα χαρακτηριστικά της υποδομής υλικού και λογισμικού των οργανισμών-στόχων.

Νέες ευφυείς μέθοδοι εξαπάτησης

Σχεδιασμός και εφαρμογή νέων ευφυών μεθόδων εξαπάτησης με βάση τα honeypots υψηλής αλληλεπίδρασης, τα ψηφιακά δίδυμα και τα εικονικά πρόσωπα.

Μέθοδοι προστασίας με βάση την τεχνητή νοημοσύνη

Σχεδιασμός, ανάπτυξη και αξιολόγηση μεθόδων βασισμένων στην τεχνητή νοημοσύνη για την ανίχνευση και το μετριασμό των επιθέσεων στον κυβερνοχώρο, συμπεριλαμβανομένων εχθρικών επιθέσεων τεχνητής νοημοσύνης, καθώς και σύλληψη και εφαρμογή μεθόδων συλλογής και συγχώνευσης δεδομένων.

Μηχανή συστάσεων με βάση την επεξηγήσιμη τεχνητή νοημοσύνη

Ανάπτυξη και επαλήθευση μηχανής συστάσεων βασισμένη στην επεξηγήσιμη τεχνητή νοημοσύνη, η οποία η οποία επιτρέπει στον άνθρωπο να αποφασίζει προληπτικά για να μετριάσει πλήρως τις επιθέσεις της τεχνητής νοημοσύνης.

Χρήση στην πραγματική ζωή

Αξιολόγηση της λειτουργικότητας, της αποτελεσματικότητας και της αποδοτικότητας του AIAS σε πραγματικά σενάρια.

Μεθοδολογία

Φάση 1: Απαιτήσεις συστήματος και προσδιορισμός των κύριων στοιχείων της πλατφόρμας.

- Προσδιορισμός και ορισμός των απαιτήσεων ασφάλειας, ιδιωτικότητας, λειτουργικότητας και δεοντολογίας.
- Ανασκόπηση της βιβλιογραφίας στους βασικούς τομείς του έργου, τις μεθόδους εξαπάτησης, τις μεθόδους ανίχνευσης και μετριασμού με βάση την τεχνητή νοημοσύνη.
- Καθορισμός των εργαλείων και των εφαρμογών που θα χρησιμοποιηθούν για την υλοποίηση των συστημάτων του AIAS.

Φάση 2: Εφαρμογή και επικύρωση των κύριων στοιχείων της πλατφόρμας. Κάθε στοιχείο θα σχεδιαστεί σύμφωνα με γνωστές τεχνικές, εφαρμόζοντας σχετικές μεθόδους και κατευθυντήριες γραμμές που παρέχονται από την ΕΕ, καθώς και με τη χρήση τεχνολογιών αιχμής.

Φάση 3: Ενσωμάτωση, μελέτη απόδειξης της λειτουργίας και αξιολόγηση σε πραγματικές συνθήκες. Ο κύριος στόχος αυτής της φάσης είναι να παραδοθεί η πλατφόρμα AIAS και οι ενότητες που την απαρτίζουν να είναι λειτουργικές και να λειτουργούν απρόσκοπτα. Μόλις ολοκληρωθεί η ολοκλήρωση της πλατφόρμας, οι συμμετέχοντες φορείς θα αξιολογήσουν την πλατφόρμα μέσω προσεκτικά επιλεγμένων πιλοτικών περιπτώσεων χρήσης σε πραγματικές συνθήκες. Επίσης μπορεί να περιλαμβάνει τροποποιήσεις της πλατφόρμας με βάση την ανατροφοδότηση που θα αποκτηθεί κατά τη διάρκεια των πειραμάτων.