Πανεπιστήμιο Πειραιώς
Σχολή Τεχνολογιών Πληροφορικής και Τηλεπικοινωνίων
Τμήμα Ψηφιακών Συστημάτων

Επίπεδο: Μεταπτυχιακό Πρόγραμμα Σπουδών

Διπλωματική Εργασία με θέμα

«Διαχείριση κινδύνων Ασφαλείας και Ιδιωτικότητας σε Συστήματα Τεχνητής Νοημοσύνης»

“Security and Privacy Risk Management in Artificial Intelligence Systems”

Επιβλέπον Καθηγητής: Στέφανος Γκρίτζαλης

Αλέστας Βασίλειος MTE2301

Πειραιάς
15/01/2026

Περίληψη

Η ανάπτυξη των συστημάτων Τεχνητής Νοημοσύνης (ΤΝ) σε κρίσιμους τομείς της κοινωνίας και της οικονομίας έχει αναδείξει νέες προκλήσεις που σχετίζονται με την ασφάλεια των πληροφοριών και την προστασία της ιδιωτικότητας. Τα συστήματα, λόγω της εξάρτησής τους από μεγάλο όγκο δεδομένων, της πολυπλοκότητας και της δυναμικής τους συμπεριφοράς, εισάγουν κινδύνους που δεν καλύπτονται επαρκώς από τις παραδοσιακές προσεγγίσεις διαχείρισης κινδύνων. Στο πλαίσιο αυτό, η παρούσα διπλωματική εργασία εξετάζει συστηματικά τη διαχείριση κινδύνων ασφάλειας και ιδιωτικότητας σε συστήματα Τεχνητής Νοημοσύνης. Αρχικά, παρουσιάζονται οι βασικές έννοιες της Τεχνητής Νοημοσύνης και οι τεχνολογίες που την απαρτίζουν, καθώς και η επίδρασή της στην κοινωνία και στο κανονιστικό περιβάλλον. Στη συνέχεια, αναλύονται οι κύριοι κίνδυνοι ασφάλειας και ιδιωτικότητας που προκύπτουν κατά τον κύκλο ζωής των συστημάτων AI. Παράλληλα, παρουσιάζεται η διαδικασία διαχείρισης και αξιολόγησης κινδύνων, βασισμένη σε διεθνή πρότυπα και πλαίσια, όπως του ISO και του NIST, αλλά και ο ρόλος του Ρυθμιστικού πλαισίου για την Τεχνητή Νοημοσύνη, όπως ο ευρωπαϊκός κανονισμός AI Act. Τέλος, η εργασία εξετάζει σύγχρονα πλαίσια χαρτογράφησης και ταξινόμησης κινδύνων, όπως το Plot4AI, το MITRE ATLAS, το και το AI Risk Taxonomy, τα οποία συμβάλλουν στην κατανόηση και αποτύπωση των ρίσκων που διέπουν την τεχνολογία αυτή, και καταλήγει σε συμπεράσματα και μελλοντικούς προβληματισμούς.

Λέξεις κλειδιά: Τεχνητή Νοημοσύνη, Διαχείριση Κινδύνων, Αντιμετώπιση Κινδύνων, ISO, NIST, Ρυθμιστικό πλαίσιο, Χαρτογράφηση Κινδύνων.

Abstract

The development of Artificial Intelligence (AI) systems in critical areas of society has brought up new challenges related to information security and privacy protection. Due to their dependence on large volumes of data, their complexity, and their dynamic behavior, these systems introduce risks that are not adequately covered by traditional risk management approaches. In this context, this thesis systematically examines security and privacy risk management in Artificial Intelligence systems. First, the basic concepts of Artificial Intelligence are presented, as well as its impact on society and the regulatory environment. Next, the main security and privacy risks that arise during the life cycle of AI systems are analyzed. At the same time, the risk management and assessment process is presented, based on international standards and frameworks, such as ISO and NIST, as well as the role of Regulatory Frameworks for Artificial Intelligence, such as the European AI Act. Finally, the paper examines contemporary risk mapping and classification frameworks, such as Plot4AI, MITRE ATLAS, and AI Risk Taxonomy, which contribute to the understanding and mapping of the risks governing this technology, and leads to conclusions and future considerations.

Key Words: Artificial Intelligence, Risk Management, Risk Response, ISO, NIST, Regulatory Framework, Risk Mapping.

Πίνακας Περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1. Συστήματα Τεχνητής Νοημοσύνης

1.1 Τι είναι τεχνητή Νοημοσύνη

Ένας από τους πιο συχνούς όρους που ακούμε πλέον στην καθημερινότητά μας είναι ο όρος Τεχνητή Νοημοσύνη. Με ένα πάρα πολύ ευρύ πεδίο εφαρμογής, η Τεχνητή Νοημοσύνη (ΤΝ) εντοπίζεται σε όλες τις πτυχές τις ανθρώπινης δραστηριότητας, από την εκπαίδευση και την Ιατρική, μέχρι τις μεταφορές και το Marketing. Η Εξέλιξη της συγκεκριμένης τεχνολογίας, από τα αργά και σταθερά βήματα που ακολουθούσε στις αρχές το 21^{ου} αιώνα, κατέληξε σε τεράστια άλματα μόνο την τελευταία πενταετία. Πλέον δεν αποτελεί απλώς μια ιδέα ενός μακρινού μέλλοντος, αλλά μια σύγχρονη βιομηχανία, που διαμορφώνει την ανθρώπινη καθημερινότητα και τον τρόπο ζωής.

Τι είναι όμως η Τεχνητή Νοημοσύνη; Τεχνητή Νοημοσύνη είναι η δυνατότητα ενός υπολογιστή να υλοποιήσει ενέργειες, που κανονικά θα συνδέονταν με ευφυή όντα, όπως τους ανθρώπους. Ο όρος εφαρμόζεται συχνά στην ανάπτυξη συστημάτων που διαθέτουν ικανότητες που χαρακτηρίζουν τον άνθρωπο, όπως η ικανότητα να σκέφτεται, να ανακαλύπτει νοήματα, να γενικεύει ή να μαθαίνει από προηγούμενες εμπειρίες. Από την ανάπτυξή τους τη δεκαετία του 1940, τα ψηφιακά συστήματα έχουν αναπτυχθεί τόσο ώστε να εκτελούν πολύ σύνθετες εργασίες -όπως η υλοποίηση μαθηματικών θεωρημάτων ή το σκάκι- με μεγάλη επιδεξιότητα, συγκρίσιμη με την αντίστοιχη ανθρώπινη (Νίκη του AplaGo κατά του παγκόσμιου πρωταθλητή σκακιού το 2016). Παρά τις συνεχείς προόδους στην ισχύ των υπολογιστικών συστημάτων, δεν υπάρχουν ακόμη υλοποιήσεις που να μπορούν να φτάσουν την πλήρη ανθρώπινη ευελιξία σε ευρύτερους τομείς ή σε εργασίες που απαιτούν καθημερινές γνώσεις. Από την άλλη πλευρά, σε ορισμένες περιπτώσεις έχουν επιτευχθεί επίπεδα επιδόσεων των ανθρώπινων ειδικών και επαγγελματιών στην εκτέλεση συγκεκριμένων εργασιών, έτσι ώστε η τεχνητή νοημοσύνη να συναντάται σε εφαρμογές τόσο περίπλοκες όσο η ιατρική διάγνωση, οι μηχανές αναζήτησης, η αναγνώριση φωνής ή γραφής και τα chatbots. (Copeland, 2024). Η τεχνητή νοημοσύνη δεν έχει ως σκοπό να αντιγράψει απόλυτα την ανθρώπινη συμπεριφορά, αλλά να συνεισφέρει, να βελτιώσει και να συντομεύσει ανθρώπινες δραστηριότητες. Για αυτόν τον λόγο, οι κύριες ικανότητες

που αναπτύσσονται αφορούν κυρίως την εκμάθηση, την κατανόηση γραπτού και προφορικού λόγου, την επίλυση προβλημάτων, την αντίληψη και την χρήση της ανθρώπινης γλώσσας για την επικοινωνία.

Για να κατανοήσουμε καλύτερα το πώς η συγκεκριμένη τεχνολογία απέκτησε την επιρροή και την ευρεία αποδοχή, είναι σημαντικό να κατανοήσουμε τις έννοιες που την απαρτίζουν.

1.1.1 Μηχανική Μάθηση (Machine Learning)

Η κύρια προσέγγιση για την κατασκευή συστημάτων ΤΝ είναι η **μηχανική μάθηση (ML)**, όπου οι υπολογιστές μαθαίνουν από μεγάλα σύνολα δεδομένων εντοπίζοντας μοτίβα και σχέσεις μεταξύ των δεδομένων αυτών. Ένας αλγόριθμος μηχανικής μάθησης χρησιμοποιεί στατιστικές τεχνικές για να τον βοηθήσει να «μάθει» πώς να γίνεται προοδευτικά καλύτερος σε μια εργασία, χωρίς απαραίτητα να έχει προγραμματιστεί για τη συγκεκριμένη εργασία. Χρησιμοποιεί ιστορικά δεδομένα ως είσοδο για να προβλέψει νέες τιμές εξόδου. Υπάρχουν πολλές και διαφορετικές μέθοδοι Μηχανικής Μάθησης, με τις πιο διαδεδομένα να είναι :

1) μάθηση με επίβλεψη (Supervised Learning),
2) μάθηση χωρίς επίβλεψη (Unsupervised Learning),
3) μάθηση με ημιεπίβλεψη (Semi-Supervised Learning)
4) ενισχυτική μάθηση (Reinforcement Learning).

Η επιβλεπόμενη μάθηση αποτελείται από την χρήση ορισμένων και επισημασμένων (labelled) συνόλων δεδομένων για την εκπαίδευση ενός αλγόριθμου τεχνητής Νοημοσύνης ώστε να είναι σε θέση να κατηγοριοποιήσει, να ταξινομήσει ή να προβλέψει αποτελέσματα βάσει τα δεδομένα εισόδου που έχει λάβει. Η συγκεκριμένη μαθησιακή μέθοδος αποτελεί την πιο ευρέως αποδεκτή καθώς ο αλγόριθμος μαθαίνει να διακρίνει διαφορές μεταξύ των δεδομένων εισόδου και να προσαρμόζει την απόφασή του καταλλήλως με αποτέλεσμα να γίνεται πιο αποτελεσματικός με την πάροδο του χρόνου (Πχ. Διάκριση μεταξύ εικόνων διαφορετικών ζώων). Η επιβλεπόμενη μάθηση επιτρέπει την επίλυση καθημερινών ζητημάτων που μπορούν να εντοπιστούν σε εταιρικά περιβάλλοντα, όπως η κατηγοριοποίηση και ταξινόμηση διαφορετικών email σε διαφορετικούς φακέλους.

Η μη επιβλεπόμενη μάθηση περιλαμβάνει την ανάλυση δεδομένων χωρίς σήμανση. Κατά την μέθοδο αυτή ο αλγόριθμος καλείται να αναγνωρίσει συγκεκριμένα

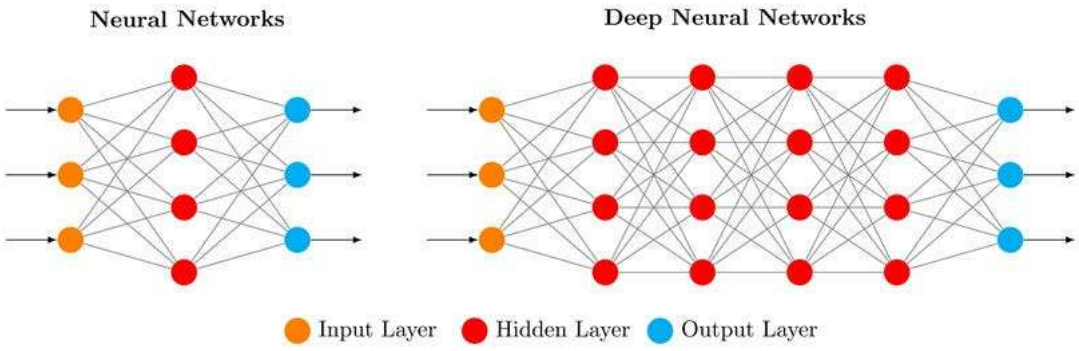
μοτίβα ή σχέσεις μεταξύ των δεδομένων εισόδου χωρίς να γνωρίζει τι είδους δεδομένα είναι ή την σημασία τους και χωρίς ανθρώπινη παρέμβαση. Κατά την υλοποίηση της διαδικασίας αυτής, ο αλγόριθμος καλείται να διερμηνεύσει τα δεδομένα που δέχεται σαν είσοδο και να προσπαθήσει να τα κατηγοριοποιήσει βάσει των μοτίβων που εντοπίζει κατά την επεξεργασία τους. Το προτέρημα της συγκεκριμένης μαθησιακής διαδικασίας, πέραν της δυσκολίας που περιλαμβάνεται σε αυτή, είναι πως μπορούν να βρεθούν σχέσεις σε δεδομένα που με την επιβλεπόμενη μάθηση δεν θα ήταν εύκολο να εντοπιστούν. Συχνά συναντάμε τη μέθοδο αυτή σε αναγνώριση μοτίβων και εικόνων (image & pattern recognition) και την κατηγοριοποίηση πελατών ανάλογα με τις αγοραστικές τους ιδιότητες.

Η ημιεπιβλεπόμενη μάθηση αποτελεί μια χρυσή τομή μεταξύ της επιβλεπόμενης και της μη επιβλεπόμενης μάθησης. Κατά τη διάρκεια της εκπαίδευσης, χρησιμοποιείται ένα μικρότερο σύνολο δεδομένων το οποίο έχει επισημανθεί για να καθοδηγήσει την ταξινόμηση και την εξαγωγή χαρακτηριστικών από ένα μεγαλύτερο σύνολο δεδομένων χωρίς ετικέτες. Η μάθηση με ημιεπίβλεψη μπορεί να λύσει το πρόβλημα της έλλειψης ικανοποιητικού αριθμού επισημασμένων δεδομένων για έναν αλγόριθμο μάθησης με επίβλεψη και παράλληλα συμβάλλει σε περιπτώσεις που είναι πολύ δαπανηρό να επισημανθεί μεγάλο πλήθος δεδομένων. Εφαρμογές της συγκεκριμένης μεθόδου συναντάμε στην ανάλυση προφορικού λόγου, όπου η επισήμανση ακουστικού υλικού είναι ιδιαίτερα χρονοβόρα και κοστοβόρα διαδικασία, αλλά και στην κατηγοριοποίηση περιεχομένου στο διαδίκτυο για αντίστοιχους λόγους.

Τέλος, κατά την **ενισχυτική μάθηση**, ο αλγόριθμος δεν μαθαίνει χρησιμοποιώντας επισημασμένα δεδομένα, αλλά μια αλληλουχία επιτυχών και ανεπιτυχών προσπαθειών ώστε να καταλήξει σε ένα συγκεκριμένο αποτέλεσμα. Στην ουσία ο αλγόριθμος αλληλοεπιδρά με το περιβάλλον λαμβάνοντας ανατροφοδότηση ανάλογα με τις ενέργειές του. Οι συνεχείς επιτυχείς προσπάθειες, υποδεικνύουν στον αλγόριθμο πώς να αναπτύξει την βέλτιστη επίλυση στο πρόβλημα στο οποίο καλείται να λύσει. Συνήθως, η συγκεκριμένη μέθοδος εφαρμόζεται στη ρομποτική, κατά την εκμάθηση των ρομπότ ώστε να υλοποιούν τις βέλτιστες και πιο αποδοτικές κινήσεις για την ολοκλήρωσης μια συγκεκριμένης ενέργειας ή στην εφαρμογή εκπαιδευτικών συστημάτων τα οποία προσαρμόζουν το περιεχόμενό τους βάσει τη μαθησιακή συμπεριφορά του εκάστοτε ατόμου.

1.1.2 Deep Learning

Το Deep Learning αποτελεί ένα υποσύνολο της μηχανικής μάθησης. Χρησιμοποιεί έναν τύπο τεχνητού νευρωνικού δικτύου γνωστού ως βαθύ νευρωνικό δίκτυο (Deep Neural Network), τα οποία περιέχουν έναν αριθμό κρυφών στρωμάτων μέσω των οποίων γίνεται η επεξεργασία των δεδομένων, επιτρέποντας σε μια μηχανή να προχωρήσει σε «βαθιά» μάθηση και να αναγνωρίζει όλο και πιο σύνθετα μοτίβα. Το πρώτο επίπεδο δέχεται τα δεδομένα εισόδου τα οποία υπόκεινται σε επεξεργασία και μεταβαίνουν σε επόμενα επίπεδα χρησιμοποιώντας μη γραμμικές μεθόδους και συναρτήσεις. Εν τέλει, θα καταλήξουν στο τελευταίο επίπεδο (επίπεδο εξόδου) το οποίο παράγει και την πρόβλεψη ή απάντηση του μοντέλου. Λόγω των σύνθετων σχέσεων που δημιουργεί, και της δημιουργίας και αναγνώρισης μοτίβων, το deep Learning είναι ιδιαίτερα αποτελεσματικό σε εργασίες όπως η αναγνώριση εικόνας και ομιλίας και η επεξεργασία φυσικής γλώσσας, γεγονός που την καθιστά κρίσιμη συνιστώσα στην ανάπτυξη και την εξέλιξη των συστημάτων ΤΝ. Όμοια με την μηχανική μάθηση, έτσι και στο Deep Learning, υπάρχουν διαφορετικές μέθοδοι εκπαίδευσης του μοντέλου, όπως η εποπτευόμενη διδασκαλία, η μη εποπτευόμενη διδασκαλία και η ενισχυτική διδασκαλία, η κάθε μία από τις οποίες θεωρείται κατάλληλη για διαφορετικές περιπτώσεις, ανάλογα με την πολυπλοκότητα και τις ανάγκες του έργου προς υλοποίηση.



Εικόνα 1: Νευρωνικά Δίκτυα

Όπως και η μηχανική μάθηση, έτσι και στην περίπτωση του Deep Learning, πέραν από τις μεθόδους εκπαίδευσης του μοντέλου, υφίστανται και διαφορετικά είδη αυτού.

1. **Feedforward Neural Networks (FNN)** : Αυτή η αρχιτεκτονική αντιπροσωπεύει τον απλούστερο τύπο νευρωνικού δικτύου. Σε ένα τέτοιο δίκτυο, η πληροφορία ρέει προς μία μόνο κατεύθυνση - από τους κόμβους

εισόδου, μέσω κρυφών κόμβων και, τέλος, στους κόμβους εξόδου. Αυτή η διαμόρφωση εξασφαλίζει ότι δεν υπάρχουν κύκλοι ή επαναλήψεις μέσα στο δίκτυο. Τα FNN υπερέχουν στην επίλυση προβλημάτων όπου τα δεδομένα πρέπει να εξελίσσονται γραμμικά, όπως οι εργασίες ταξινόμησης ή αναγνώρισης εικόνων. Ένα FNN περιλαμβάνει συνήθως τρία κύρια στρώματα: ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου. Η δομή τους είναι απλή, καθιστώντας το εύκολο στην κατανόηση.

2. **Convolutional Neural Networks (CNNs):** είναι ένας εξειδικευμένος τύπος νευρωνικού δικτύου που έχει σχεδιαστεί για την επεξεργασία δεδομένων με δομή που μοιάζει με πλέγμα, όπως οι εικόνες, ο ήχος και ο λόγος. Είναι ιδιαίτερα αποτελεσματικά σε διάφορες εργασίες υπολογιστικής όρασης (computer vision), όπως η αναγνώριση εικόνων και βίντεο, η ταξινόμηση εικόνων και η ανίχνευση αντικειμένων.
3. **Recurrent Neural Networks (RNNs):** είναι ένα είδος νευρωνικού δικτύου που είναι σε θέση να επεξεργάζονται διαδοχικά δεδομένα, όπως χρονοσειρές και φυσική γλώσσα. Τα RNN, σε αντίθεση με τα FNN, έχουν βρόχους για να διατηρούν πληροφορίες με την πάροδο του χρόνου δημιουργώντας ένα είδος μνήμης, επιτρέποντας εφαρμογές όπως η μοντελοποίηση γλώσσας και η αναγνώριση ομιλίας, όπου απαιτείται να τηρούν προηγούμενες πληροφορίες. Για τον λόγο αυτό δημιουργείται ένα επιπλέον κρυφό επίπεδο που τηρεί τις πληροφορίες αυτές, με αποτέλεσμα να είναι ικανό να διαχειριστεί εργασίες που χρειάζεται να χρησιμοποιήσει ιστορικά δεδομένα.
4. **Generative Adversarial Networks (GANs):** αποτελούνται από δύο δίκτυα - μια γεννήτρια και έναν διαχωριστή (discriminator) – που εκπαιδεύονται ταυτόχρονα και ανταγωνίζονται για τη δημιουργία ρεαλιστικών δεδομένων. Η γεννήτρια αποσκοπεί στην παραγωγή δεδομένων που μιμούνται έναν συγκεκριμένο τύπο δεδομένων εκπαίδευσης, ενώ ο ρόλος του διαχωριστή είναι να διακρίνει μεταξύ των γνήσιων δεδομένων εκπαίδευσης και των δεδομένων που παράγονται από τη γεννήτρια. Τα GAN χρησιμοποιούνται ευρέως για τη δημιουργία εικόνων ή βίντεο και την ενίσχυση δεδομένων.
5. **Transformer Networks:** Τα Transformer networks (μετασχηματιστές) είναι ένας τύπος αρχιτεκτονικής νευρωνικών δικτύων που κυριαρχεί στην επεξεργασία φυσικής γλώσσας λόγω των μηχανισμών αυτοπροσοχής (self-attention mechanisms), η δυνατότητα, δηλαδή, του μοντέλου να αξιολογεί μια

αλληλουχία δεδομένων και να καθορίζει πως το εκάστοτε δεδομένο συσχετίζεται με τα υπόλοιπα. Αυτοί οι μηχανισμοί επιτρέπουν στο δίκτυο να αποδίδει, δυναμικά, σημασία σε συγκεκριμένα σημεία των δεδομένων εισόδου, γεγονός που βελτιώνει την απόδοση σε εργασίες όπως η μετάφραση, η περίληψη κειμένου και η απάντηση ερωτήσεων, δημιουργώντας μοντέλα όπως το GPT.

1.1.3 Natural language processing

Η επεξεργασία φυσικής γλώσσας (Natural language processing - NLP) αναφέρεται στην ικανότητα των υπολογιστών να αναγνωρίζουν και να κατανοούν την ανθρώπινη γλώσσα, η οποία αποτελεί διεπιστημονικό αντικείμενο μεταξύ της επιστήμης των υπολογιστών και της ανθρώπινης γλωσσολογίας. Το μεγαλύτερο ζήτημα της τεχνητής νοημοσύνης είναι ότι η ανθρώπινη σκέψη βασίζεται στη γλώσσα, οπότε η επεξεργασία φυσικής γλώσσας αποτελεί στόχο για την καλύτερη αλληλεπίδραση με τον άνθρωπο.

Το NLP είναι μια τεχνολογία που χρησιμοποιεί φυσική γλώσσα για την επικοινωνία με τους υπολογιστές. Το κλειδί για την επεξεργασία της φυσικής γλώσσας είναι να επιτρέψει στους υπολογιστές να τη "κατανοήσουν", γι' αυτό και ονομάζεται επίσης υπολογιστική γλωσσολογία, η οποία βρίσκεται στη διασταύρωση της επεξεργασίας γλωσσικών πληροφοριών και της τεχνητής νοημοσύνης. Ως μηχανή, αρχικά συλλέγει ηχητικά σήματα. Η τεχνολογία επεξεργασίας φυσικής γλώσσας μετατρέπει τους ήχους σε λέξεις και τις λέξεις σε νοήματα. Μετά την ολοκλήρωση αυτών των δύο διαδικασιών, η μηχανή μπορεί να ακούσει και να κατανοήσει. Η μηχανή που είναι εξοπλισμένη με τεχνολογία αναγνώρισης ομιλίας και σημασιολογικής κατανόησης βελτιστοποιεί την συνεχή μάθηση, έτσι ώστε να μπορεί, όχι μόνο να ακούει, αλλά και να καταλαβαίνει τα ανθρώπινα ερεθίσματα. (Zhang, C. and Lu, Y., 2021).

Η επεξεργασία φυσικής γλώσσας χωρίζεται σε πέντε κατευθύνσεις:

- **Τη γραμματική και σημασιολογική ανάλυση:** η δυνατότητα καθορισμού ρόλων των διαφορετικών λέξεων όπως ρήμα, ουσιαστικό κλπ. Αλλά και η αναγνώριση και κατηγοριοποίηση των λέξεων σε οντότητες, όπως ονόματα ημερομηνίες.

- **Την εξαγωγή πληροφοριών και την εξόρυξη κειμένου:** τη δημιουργία σχέσεων μεταξύ των λέξεων στο κείμενο και την κατηγοριοποίηση αυτών των σχέσεων.
- **Τη παραγωγή γλώσσας:** όπως για παράδειγμα, η μετάφραση της μιας ανθρώπινης γλώσσας σε μια άλλη, η δυνατότητα απλοποίησης ενός μεγάλου κειμένου σε μια περίληψη αλλά και η δυνατότητα να παραχθεί ένα συνεκτικό κείμενο που να αποδίδει νόημα στην ανθρώπινη γλώσσα.
- **Το σύστημα απάντησης και ερωτήσεων:** μέσω του οποίου η TN, έχει την δυνατότητα να αναγνωρίζει και να παρουσιάσει το πιο σχετικό μήνυμα σαν απάντηση σε ένα ερώτημα.
- **Το σύστημα διαλόγου:** δηλαδή η δυνατότητα μιας μηχανής να συμμετέχει σε διάλογο μαζί με τους χρήστες μέσω απαντήσεων σε ερωτήματα και μέσω εργασιών βάσει μιας εισόδου από τον χρήστη.

Η Τεχνολογία αυτή εντοπίζεται σε εφαρμογές που χρησιμοποιούν σαν κύριο μέσο επικοινωνίας την γλώσσα ή που βασίζονται στην γλώσσα όπως εφαρμογές μετάφρασης κειμένου, αναγνώριση φωνής, περίληψη κειμένων και τα Chatbots.

1.1.4 Computer Vision

Η υπολογιστική όραση (Computer Vision) είναι μια ακόμα διαδεδομένη εφαρμογή τεχνικών μηχανικής μάθησης, όπου οι μηχανές επεξεργάζονται ακατέργαστες εικόνες, βίντεο και οπτικά μέσα και εξάγουν χρήσιμες πληροφορίες από αυτά. Το Deep Learning και τα νευρωνικά δίκτυα χρησιμοποιούνται για την ανάλυση των εικόνων σε εικονοστοιχεία και την ανάλογη επισήμανσή τους, γεγονός που βοηθά τους υπολογιστές να διακρίνουν τη διαφορά μεταξύ οπτικών σχημάτων και μοτίβων. Η υπολογιστική όραση χρησιμοποιείται για την αναγνώριση και ταξινόμηση εικόνων, και την ανίχνευση αντικειμένων και ολοκληρώνει εργασίες όπως η αναγνώριση και ο εντοπισμός προσώπων σε αυτοκινούμενα οχήματα και ρομπότ. Η συγκεκριμένη τεχνολογία βασίζεται:

1. Στην Εξαγωγή χαρακτηριστικών (Fearture Extraction) όπου κατά τη διάρκεια αυτής της φάσης, το σύστημα εξετάζει προσεκτικά τα εισερχόμενα οπτικά δεδομένα για να εντοπίσει και να απομονώσει σημαντικά οπτικά στοιχεία και σημαντικές αλλαγές στην ένταση ή το χρώμα, ακμές, σχήματα,

υφές και μοτίβα. Αυτά τα χαρακτηριστικά είναι κρίσιμα καθώς λειτουργούν ως δομικά στοιχεία για τα επόμενα στάδια της ανάλυσης. Για να διευκολυνθεί η επεξεργασία από τον υπολογιστή, αυτά τα αναγνωρισμένα χαρακτηριστικά μεταφράζονται σε αριθμητικές αναπαραστάσεις, μετατρέποντας ουσιαστικά τις οπτικές πληροφορίες σε μορφή που οι μηχανές μπορούν να κατανοήσουν και να χειριστούν πιο αποτελεσματικά.

2. Στην Ανίχνευση Αντικειμένων όπου αφού εξαχθούν τα χαρακτηριστικά και μετατραπούν σε αριθμητικά δεδομένα, οι αλγόριθμοι του συστήματος εντοπίζουν και αναγνωρίζουν συγκεκριμένα αντικείμενα ή οντότητές μέσα στις εικόνες. Αυτό επιτρέπει στους υπολογιστές όχι μόνο να ανιχνεύουν την παρουσία αντικειμένων, αλλά και να κατανοούν τι είναι αυτά τα αντικείμενα, μια ικανότητα που βρίσκει εφαρμογές σε τομείς που κυμαίνονται από αυτόνομα οχήματα που αναγνωρίζουν πεζούς έως συστήματα ασφαλείας που αναγνωρίζουν εισβολείς.
3. Στην κατηγοριοποίηση εικόνων όπου περιλαμβάνει την κατηγοριοποίηση τους σε προκαθορισμένες κλάσεις ή κατηγορίες (πχ. Πρόσωπο, σκύλος κλπ.). Σε αυτό το σημείο έρχονται στο προσκήνιο τα Convolutional Neural Networks (CNN) τα οποία έχουν σχεδιαστεί για εργασίες που σχετίζονται με εικόνες. Διακρίνονται στην εκμάθηση σύνθετων χαρακτηριστικών, γεγονός που τους επιτρέπει να διακρίνουν περίπλοκα μοτίβα και να κάνουν ακριβείς ταξινομήσεις εικόνων. Για παράδειγμα, εταιρείες χρησιμοποιούν την εν λόγω δυνατότητα, για τον εντοπισμό εικόνων χρηστών που θα μπορούσαν να εκθέσουν με οποιονδήποτε τρόπο την οντότητα αυτή.
4. Στην ανίχνευση αντικειμένων που αποτελεί βασική τεχνική στην ανάλυση βίντεο. Περιλαμβάνει τη δυνατότητα παρακολούθησης και εντοπισμού της κίνησης των αντικειμένων καθώς αυτά μετακινούνται κατά την διάρκεια ενός βίντεο. Βασικές εφαρμογές της συγκεκριμένης δυνατότητας συναντάμε, στην επιτήρηση και την αθλητική ανάλυση έως και τη ρομποτική. (OpenCV, 2023)

1.1.5 Generative AI

Η Γενετική Τεχνητή Νοημοσύνη (Generative AI ή GenAI) αποτελεί έναν ραγδαία αναπτυσσόμενο κλάδο της μηχανικής μάθησης, με κύριο στόχο τη δημιουργία νέου,

περιεχομένου, όπως κείμενο, εικόνες, ήχο και βίντεο, με τρόπο που προσομοιάζει την ανθρώπινη δημιουργία. Τα γενετικά μοντέλα δεν βασίζονται στην απλή αναπαραγωγή υπαρχόντων δεδομένων, αλλά σε Deep Learning, μέσω των οποίων αντλούν, επεξεργάζονται και ανασυνθέτουν πληροφορίες από πολύ μεγάλο όγκο δεδομένων. (Goodfellow et al., 2016). Ιδιαίτερα σημαντικό ρόλο παίζουν τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models - LLMs), μέσω των οποίων γίνεται η επεξεργασία φυσικής γλώσσας για την παραγωγή συνεκτικού κειμένου. Κατά την εκπαίδευσή τους, τα μοντέλα αυτά εντοπίζουν μοτίβα και συσχετίσεις μέσα από τα δεδομένα, και εν συνεχεία τα αξιοποιούν για να δημιουργήσουν νέο περιεχόμενο που συνάδει με αυτά πάνω στα οποία έχουν εκπαιδευτεί.

Η εφαρμογή της GenAI είναι πολυδιάστατη και αγγίζει πληθώρα διαφορετικών τομέων. Στην εκπαίδευση, χρησιμοποιείται για την υποστήριξη μαθητών μέσω εξατομικευμένων οδηγών μάθησης, ενώ στον τομέα της υγειονομικής περίθαλψης συμβάλλει στην ανάλυση ιατρικών δεδομένων, στις διαγνώσεις και στη δημιουργία εξατομικευμένων θεραπευτικών οδηγιών. Παράλληλα, στον χώρο των τεχνών όπως η μουσική και η λογοτεχνία, η GenAI επιτρέπει τη δημιουργία νέων έργων γεγονός που εν δυνάμει έχει και θετικές και αρνητικές συνέπειες. Επιπλέον, η ικανότητα της GenAI να παράγει λειτουργικό πηγαίο κώδικα δημιουργεί νέες δυνατότητες στον τομέα της ανάπτυξης λογισμικού, αλλά και προκλήσεις σε ό,τι αφορά την πνευματική ιδιοκτησία, τόσο στο πεδίο της ανάπτυξης αλλά και τη διαχείριση κινδύνων από αυτόματα παραγόμενο λογισμικό (Microsoft, 2023) ή στο πεδίο των τεχνών και της δημιουργίας.

1.2 Επίδραση της Τεχνητής Νοημοσύνης στην κοινωνία

Η Τεχνητή Νοημοσύνη επιφέρει σημαντικές αλλαγές σε κοινωνικό, οικονομικό και τεχνολογικό επίπεδο, με επιδράσεις που εκτείνονται τόσο στην ανθρώπινη δραστηριότητα όσο και στην πρόκληση νέων μορφών κινδύνου. Η ραγδαία πρόοδος της, και ειδικότερα της γενετικής (generative AI), έχει αναδείξει κρίσιμα ζητήματα που αφορούν τη συλλογή, την ανάλυση και τη χρήση προσωπικών δεδομένων, θέτοντας παράλληλα νέες προκλήσεις για την προστασία της ιδιωτικότητας και της ασφάλειας (Floridi & Chiriatti, 2020).

Η Τεχνητή Νοημοσύνη προσφέρει δυνατότητες βελτιστοποίησης αποφάσεων και αυτοματισμού διαδικασιών σε πολλούς τομείς. Στην υγειονομική περίθαλψη, αξιοποιείται για τη διάγνωση ασθενειών μέσω ανάλυσης ιατρικών εικόνων, τη

δημιουργία εξατομικευμένων θεραπευτικών σχεδίων και την πρόβλεψη εκβάσεων, αξιοποιώντας τεχνολογίες όπως το edge computing και τα μεγάλα γλωσσικά μοντέλα (LLMs) (Shariar et al., 2023). Παράλληλα, στον τομέα της κυβερνοασφάλειας, χρησιμοποιείται για την ανίχνευση κακόβουλου λογισμικού, την πρόβλεψη απειλών και την αυτοματοποίηση αποκρίσεων σε περιστατικά. Επιπλέον, ενισχύει τη διαχείριση της εφοδιαστικής αλυσίδας μέσω προγνωστικής συντήρησης και πρόβλεψης ζήτησης, ενώ συμβάλλει και στη βιώσιμη διαχείριση ενεργειακών πόρων (Wang et al., 2020). Επιπρόσθετες εφαρμογές περιλαμβάνουν τη χρήση της στην πρόβλεψη οικονομικών τάσεων και τη βελτίωση της εμπειρίας του χρήστη στο ηλεκτρονικό εμπόριο (Kumar et al, 2024).

Ωστόσο, η εξάπλωση της τεχνολογίας αυτής, συνοδεύεται και από σοβαρές προκλήσεις, ιδίως σε ό,τι αφορά την ιδιωτικότητα και την ασφάλεια των δεδομένων που εμπλέκονται. Η εκτεταμένη συλλογή δεδομένων, συχνά χωρίς συναίνεση των υποκειμένων, ενισχύει τον κίνδυνο παραβίασης της ιδιωτικότητας. Ειδικότερα, η χρήση δεδομένων προσωπικού χαρακτήρα που αντλούνται από δημόσιες πηγές ή ακόμα και από ιατρικά αρχεία, έχει οδηγήσει σε ενσωμάτωσή τους σε υλικό εκπαίδευσης χωρίς να υπάρχει η απαραίτητη διαφάνεια ως προς τον τρόπο επεξεργασίας και το μέγεθος αυτής ή η απαραίτητη συγκατάθεση από τα φυσικά πρόσωπα στα οποία ανήκουν τα δεδομένα. (Lee et al., 2024). Τα LLMs έχουν επιδείξει ευπάθειες σε διαρροές προσωπικών δεδομένων (PII leakage), ακόμα και όταν έχουν εφαρμοστεί τεχνικές μετριασμού των κινδύνων. (Badhan, 2024) Επιπλέον, η τεχνητή νοημοσύνη είναι ευάλωτη σε επιθέσεις συμπερασμάτων, όπως οι επιθέσεις συμπερασμού μελών (membership inference), αντιστροφής μοντέλου (model inversion), και συμπερασμού ιδιοτήτων (property inference), οι οποίες αποσκοπούν στην εξαγωγή πληροφοριών από τα δεδομένα εκπαίδευσης ενός μοντέλου (Vassiliev et al., 2024).

Ένας επιπλέον προβληματισμός αφορά το φαινόμενο του "surveillance capitalism" κατά το οποίο η τεχνολογία αυτή αξιοποιείται για τη συλλογή και ανάλυση δεδομένων με σκοπό την πρόβλεψη και τον επηρεασμό της ανθρώπινης συμπεριφοράς, συχνά χωρίς διαφάνεια ή συγκατάθεση (Kumar et al, 2024). Όμοια, όπως στις περιπτώσεις στις οποίες η τεχνολογία αυτή συμβάλει στην ενίσχυση του επιπέδου ασφάλειας, υπάρχουν αντίστοιχες κατά τις οποίες η ίδια τεχνολογία, λόγω ελλείψεων στους αμυντικούς μηχανισμούς της, είναι ευάλωτη σε κακόβουλες ενέργειες που είτε μπορούν να επηρεάσουν τον τρόπο με τον οποίο διαχειρίζονται δεδομένα είτε να

επηρεάσουν τον μηχανισμό λήψης αποφάσεων, ή να επηρεάσουν το ίδιο το μοντέλο της Τεχνητής Νοημοσύνης. Τα LLMs μπορούν να αξιοποιηθούν για τη δημιουργία κακόβουλου λογισμικού, την ενίσχυση spear phishing επιθέσεων, ή ακόμη και για την παραγωγή κακόβουλου κώδικα (RatGPT) μέσω prompt injection ή jailbreaking τεχνικών. Οι συγκεκριμένες επιθέσεις, έχουν ως κύριο στόχο να παρακάμψουν τους αμυντικούς μηχανισμούς ή τα φίλτρα που έχουν ενσωματωθεί και οδηγούν το μοντέλο την παραγωγή ανεπιθύμητου ή επιβλαβούς περιεχομένου. Adversarial ή backdoor επιθέσεις ενδέχεται να δημιουργήσουν inputs με μικρές αλλά στοχευμένες παρεμβολές, ώστε να προκαλέσουν σφάλματα στην απόδοση του μοντέλου, όπως για παράδειγμα, η παραπλάνηση συστημάτων αναγνώρισης αντικειμένων σε αυτόνομα οχήματα. (Farzad et al., 2024).

Ιδιαίτερη σημασία πρέπει να δοθεί στην γενετική Τεχνητή Νοημοσύνη ή οποία, εν δυνάμει, εντείνει περαιτέρω τους κινδύνους, ειδικά μέσω της παραγωγής deepfakes και παραπληροφόρησης, προκαλώντας σοβαρές ανησυχίες για την Ιδιωτικότητα, τη δημοκρατία και την εθνική ασφάλεια (Chesney & Citron, 2019). Επιπλέον, ενδέχεται να αναπαράγει κοινωνικές προκαταλήψεις, ιδιαίτερα όταν τα δεδομένα εκπαίδευσης ενσωματώνουν μεροληψίες ή ακόμα και στερεοτυπικές αντιλήψεις, οδηγώντας σε αποφάσεις που συμπεριλαμβάνουν διακρίσεις σε τομείς όπως οι προσλήψεις ή η πιστοληπτική αξιολόγηση (Barocas et al., 2019). Το φαινόμενο της αδιαφάνειας των μοντέλων (black-box nature) δυσχεραίνει τον έλεγχο και τη λογοδοσία, εντείνοντας την ανάγκη για διαφανή μοντέλα. Ταυτόχρονα, ένα από τα πλέον κρίσιμα ζητήματα αφορά στη χρήση δεδομένων που καλύπτονται από πνευματικά δικαιώματα για την εκπαίδευση των μοντέλων, γεγονός που έχει ήδη οδηγήσει σε νομικές διενέξεις μεταξύ δημιουργών περιεχομένου και εταιρειών τεχνολογίας (Rassin & Willems, 2023). Ακόμη, η εισαγωγή της GenAI στην ακαδημαϊκή διαδικασία προκαλεί ανησυχίες για την αυθεντικότητα των φοιτητικών εργασιών και για τον κίνδυνο αντικατάστασης της δημιουργικής σκέψης από παραγωγή περιεχομένου μέσω συστημάτων τεχνητής νοημοσύνης (Cotton et al., 2023). Η τεχνολογία αυτή, παρόλο που προσφέρει χρήσιμα εργαλεία, δύναται να υπονομεύσει την ακαδημαϊκή ακεραιότητα εάν δεν υιοθετηθούν οι κατάλληλες πολιτικές ελέγχου και διαβαθμισμένης ή κλιμακούμενης ενσωμάτωσης τους.

1.2.1 Ρυθμιστικό πλαίσιο

Λόγω της εκθετικά εξελισσόμενης τεχνολογίας και της άμεσης ενσωμάτωσης της σε πληθώρα τομέων της ανθρώπινης δραστηριότητας, έχει προκύψει η ανάγκη ελέγχου και ρύθμισης της Τεχνητής Νοημοσύνης. Το ρυθμιστικό πλαίσιο που την διέπει βρίσκεται σε διαρκή εξέλιξη σε παγκόσμιο επίπεδο, με στόχο την εξισορρόπηση των πλεονεκτημάτων της τεχνολογίας με τις πιθανές απειλές για την ασφάλεια και την ιδιωτικότητα των πολιτών ή ακόμα και την προστασία των ελευθεριών και των δικαιωμάτων τους απέναντι της.

Στην Ευρωπαϊκή Ένωση, ο Νόμος για την Τεχνητή Νοημοσύνη (AI Act) αποτελεί την πρώτη ολοκληρωμένη προσπάθεια παγκοσμίως για τη ρύθμιση της τεχνολογίας βάσει του risk based thinking, ενσωματώνοντας έννοιες όπως η αξιολόγηση κινδύνων, και απαιτήσεις διαφάνειας (European Commission, 2021). Βάσει τον πιθανό κίνδυνο στα ανθρώπινα δικαιώματα που μπορεί να εμφανίσει ένα σύστημα AI, λαμβάνει και την ανάλογη διαβάθμιση. Έχουν διακριθεί τρεις βασικές κατηγορίες κινδύνου συστημάτων:

1 **Μη αποδεκτός κίνδυνος**: Ορισμένες χρήσεις του AI απαγορεύονται εντελώς, όπως η μαζική συλλογή δεδομένων προσώπων για τη δημιουργία βάσεων δεδομένων, η κοινωνική βαθμολόγηση ("social scoring"), που συναντάται σε χώρες όπως η Κίνα, και η αναγνώριση συναισθημάτων στον χώρο εργασίας. Επίσης, η αναγνώριση προσώπων σε πραγματικό χρόνο περιορίζεται, με συγκεκριμένες εξαιρέσεις για λόγους εθνικής ασφάλειας και επιβολής του νόμου.
2 **Υψηλός κίνδυνος**: Συστήματα Τεχνητής Νοημοσύνης που ενέχουν σημαντικό κίνδυνο για κρίσιμες υποδομές, ιατρικά συστήματα, εκπαίδευση, την επιβολή του νόμου και τις δημοκρατικές διαδικασίες, υπόκεινται σε αυστηρότερους περιορισμούς και ελέγχους συμπεριλαμβανομένων λεπτομερή τεκμηρίωση, λήψη διορθωτικών ενεργειών για τον περιορισμό των κινδύνων, καταγραφή ενεργειών στο πλαίσιο του συστήματος, και ανθρώπινη επίβλεψη και παρέμβαση.
3 **Ελάχιστος ή μηδαμινός κίνδυνος**: Συστήματα που εμπίπτουν σε αυτή την κατηγορία ενδέχεται να υπόκεινται μόνο σε κανόνες διαφάνειας.

Παράλληλα, ο Γενικός Κανονισμός για την Προστασία Δεδομένων (GDPR) θέτει ισχυρές βάσεις για την προστασία της ιδιωτικότητας, καθορίζοντας έννοιες όπως την διαφάνεια και την λογοδοσία, τη συγκατάθεση και τη νόμιμη επεξεργασία δεδομένων. Μπορεί ο Γενικός κανονισμός να προϋπήρχε της ραγδαίας εξέλιξης της, τα άρθρα του

όμως καλύπτουν σημαντικό εύρος των ζητημάτων που την αφορούν. Συγκεκριμένα, με το άρθρο 22 του Κανονισμού, παρέχεται το δικαίωμα στα φυσικά πρόσωπα να μην υπόκεινται σε αποφάσεις που βασίζονται αποκλειστικά σε αυτοματοποιημένη επεξεργασία ή αναφέρει ακόμα και το δικαίωμα της ανθρώπινης παρέμβασης. (Aspen Institute, 2023). Παρόμοια, οι αρχές του GDPR, όπως η προστασία της Ιδιωτικής Ζωής από τον Σχεδιασμό και εξ ορισμού (Privacy by Design and Default) (Άρθρο 25) και η υλοποίηση Εκτίμησης Αντικτύπου Προστασίας Δεδομένων (Data Protection Impact Assessment – DPIA) (Άρθρο 35), είναι ζωτικής σημασίας για την ανάπτυξη συστημάτων Τεχνητής Νοημοσύνης που επεξεργάζονται προσωπικά δεδομένα και πλαισιώνουν την αυθαίρετη χρήση τους. (Butcher, 2019)

Στις Ηνωμένες Πολιτείες, αν και απουσιάζει ενιαίος ομοσπονδιακός νόμος για την ΤΝ και εντοπίζονται συγκρουσιακές προσεγγίσεις στην διαχείριση του ζητήματος της τεχνολογίας αυτής, εξακολουθούν να υφίστανται σημαντικοί παράγοντες που διευκολύνουν και συμβάλλουν με σημαντικά πλαίσια όπως το AI Risk Management Framework του NIST, το οποίο στοχεύει στην ενίσχυση της υπεύθυνης και ασφαλούς ανάπτυξης της, εστιάζοντας σε παράγοντες όπως η ιδιωτικότητα, και η ασφάλεια (NIST, 2023). Παράλληλα, η Ομοσπονδιακή Επιτροπή Εμπορίου (FTC) έχει εκδώσει προειδοποιήσεις προς επιχειρήσεις που χρησιμοποιούν συστήματα ΤΝ με μεροληπτικές αποφάσεις ή ψευδή παρουσίαση στοιχείων. Παρόλο που οι προειδοποιήσεις αυτές αντιμετωπίζουν σημαντικά εμπόδια καθώς η προεδρία ακολουθεί διαφορετική προσέγγιση του θέματος, εξακολουθούν να υπάρχουν έλεγχοι και κυρώσεις έναντι ενεργειών χρήσης της Τεχνητής Νοημοσύνης για την μη έγκυρη παρουσίαση στοιχείων από τα συστήματα AI.

Παρόλο που δεν έχει εφαρμοστεί μια οριζόντια πολιτική σε όλες τις πολιτείες, καθώς υπάρχουν περιπτώσεις που επικρατεί η νοοτροπία που αναφέρει πως ο περιορισμός της Τεχνητής Νοημοσύνης παρεμποδίζει την καινοτομία και την πρόοδο, υπάρχουν αυτές οι οποίες μεριμνούν για την ασφαλή διαχείριση και χρήση της Τεχνητής Νοημοσύνης με τοπικές πολιτικές. Στο Κολοράντο, έχει ψηφιστεί ο νόμος SB24-205 που προωθεί την προσέγγιση βάσει ρίσκου σε συστήματα Τεχνητής Νοημοσύνης υψηλού κινδύνου και τονίζει την ευθύνη που έχουν και οι developers ενός τέτοιου συστήματος αλλά και τους Deployers, για να προστατεύουν τους καταναλωτές από αλγοριθμική διάκριση. Αντίστοιχα στην Καλιφόρνια, παρόλο που εντοπίζονται ισχυρές αντιστάσεις, εφαρμόζονται νομοθεσίες όπως ο SB 53 ή αλλιώς Transparency in Frontier Artificial Intelligence Act (TFAIA) που απαιτεί την διαφάνεια, την

λογοδοσία και την ασφάλεια στα Frontier AI συστήματα (πιο εξελιγμένα και μεγάλης κλίμακας συστήματα με πολύ ισχυρές δυνατότητες στην αλληλεπίδραση και στην παραγωγή δεδομένων) όπως το Chat GPT, το GROK και το Gemini. Ο νόμος απαιτεί την τεκμηρίωση που αφορά σχέδια διαχείρισης καταστροφικών ρίσκων (εύρεση, ανάλυση και αντιμετώπιση), όπου καταστροφικά θεωρούνται τα ρίσκα που ενδέχεται να προκαλέσουν ανθρώπινη απώλεια ή μεγάλη οικονομική ζημιά. Παράλληλα, απαιτείται η εφαρμογή του πλαισίου διαχείρισης ρίσκων βάσει εθνικών προτύπων και διεθνών και παγκόσμιων κανονισμών. Συνδυαστικά με τον κανονισμό AB 2013 (Training Data Transparency) όπου απαιτεί από τους σχεδιαστές να δημοσιεύουν σε High Level, πληροφορίες για τα datasets που χρησιμοποιούν κατά την εκπαίδευση, δημιουργούν ένα πλαίσιο διαφάνειας για την χρήση των συστημάτων. Οι πληροφορίες που συλλέγονται ενδεικτικά είναι:

1. Οι πηγές ή οι ιδιοκτήτες των συνόλων δεδομένων.
2. Περιγραφή του τρόπου με τον οποίο τα σύνολα δεδομένων προωθούν τον επιδιωκόμενο σκοπό του συστήματος ή της υπηρεσίας.
3. Περιγραφή των τύπων δεδομένων εντός των συνόλων (συμπεριλαμβανομένων των τύπων ετικετών που χρησιμοποιούνται ή των γενικών χαρακτηριστικών).
4. Εάν τα σύνολα δεδομένων περιλαμβάνουν δεδομένα που προστατεύονται από πνευματικά δικαιώματα, trademarks ή πατέντες ή εάν τα σύνολα δεδομένων είναι εξ ολοκλήρου δημόσια.
5. Εάν τα σύνολα δεδομένων αγοράστηκαν ή αδειοδοτήθηκαν από τον προγραμματιστή.
6. Εάν τα σύνολα δεδομένων περιλαμβάνουν προσωπικές πληροφορίες, όπως αυτές αποτυπώνονται στον αστικό κώδικα της Καλιφόρνια.
7. Εάν τα σύνολα δεδομένων περιλαμβάνουν συγκεντρωτικές πληροφορίες καταναλωτών
8. Εάν υπήρξε οποιοσδήποτε καθαρισμός, επεξεργασία ή άλλη τροποποίηση των συνόλων δεδομένων από τον προγραμματιστή, συμπεριλαμβανομένου του σκοπού αυτών των προσπαθειών σε σχέση με το σύστημα ή την υπηρεσία.

9. Το χρονικό διάστημα κατά το οποίο συλλέχθηκαν τα δεδομένα, συμπεριλαμβανομένης μιας ενημέρωσης εάν η συλλογή είναι σε εξέλιξη.
10. Οι ημερομηνίες κατά τις οποίες τα σύνολα δεδομένων χρησιμοποιήθηκαν για πρώτη φορά κατά την ανάπτυξη του συστήματος ή της υπηρεσίας.
11. Εάν το GenAI σύστημα ή η υπηρεσία χρησιμοποίησε ή χρησιμοποιεί επανειλημμένα τη δυνατότητα παραγωγής συνθετικών δεδομένων κατά την ανάπτυξή του.

Τέλος, η πολιτεία του Τέξας, έθεσε σε εφαρμογή τον κανονισμό Texas's Responsible Artificial Intelligence Governance Act (TRAIGA; C.S.H.B. 149), όπου πλαισιώνει τις δοκιμές της Τεχνητής Νοημοσύνης σε ένα περιβάλλον sandbox ενώ προβλέπει αστικές κυρώσεις και επιβολή από τον Γενικό Εισαγγελέα.

Ομοίως στον Καναδά, έχει θεσπιστεί η "Οδηγία για την Αυτοματοποιημένη Λήψη Αποφάσεων" και "Αξιολογήσεις Αντικτύπου Αλγορίθμων" μέσω του Artificial Intelligence and Data Act (AIDA), όμοιο με το AI Act της ΕΕ, που έχει ως σκοπό την προστασία των ανθρωπίνων δικαιωμάτων σε συστήματα Τεχνητής Νοημοσύνης υψηλής κρισιμότητας, την εφαρμογή πολιτικών και την επιβολή τους κατά την εξέλιξη της τεχνολογίας και τέλος την αποτροπή κακόβουλης χρήσης των εν λόγω συστημάτων. (Yeung, 2021)

Η ανάγκη για εφαρμογή ρυθμιστικού πλαισίου και συνολικής διαχείρισης του AI, εντοπίζεται και πέραν του δυτικού κόσμου. Στη Νέα Ζηλανδία, η κυβέρνηση θέσπισε τον «Χάρτη αλγορίθμων για την Αοτεαρόα Νέα Ζηλανδία», τον οποίο οι κυβερνητικές υπηρεσίες μπορούν να επιλέξουν να υπογράψουν. Οι οργανισμοί που ασχολούνται με την εκπαίδευση, το περιβάλλον, την κοινωνική ανάπτυξη και την αστυνομία, μεταξύ άλλων, συνέβαλαν στην δημιουργία της συγκεκριμένης οδηγίας. Βάσει του χάρτη, οι υπογράφοντες δεσμεύονται να προβαίνουν σε αξιολόγηση των αποφάσεων των αλγορίθμων που χρησιμοποιούν χρησιμοποιώντας έναν πίνακα κινδύνου (Risk Matrix), ο οποίος βοηθά στην εκτίμηση της πιθανότητας ακούσιων αρνητικών

Εικόνα 2: Πίνακας Κινδύνων, Νέα Ζηλανδία

αποτελεσμάτων και της κλίμακας και της σοβαρότητας των επιπτώσεων των αποτελεσμάτων αυτών. Βάσει την αξιολόγηση, οι αποφάσεις κατατάσσονται σε τρεις διαφορετικές αξιολογήσεις κινδύνου - χαμηλή, μέτρια ή υψηλή - και οι αξιολογήσεις αυτές καθορίζουν αντίστοιχα αν οι περιορισμοί που προκύπτουν από την εν λόγω χαρτογράφηση μπορούν, ή πρέπει να εφαρμοστούν. Οι δεσμεύσεις περιλαμβάνουν την ανάληψη διορθωτικών ενεργειών για την αντιμετώπιση των κινδύνων, όπως ο εντοπισμός και η διαχείριση συμβάντων μεροληψίας, η τακτική αξιολόγηση αλγορίθμων και η παροχή ενός μέσου άσκησης του δικαιώματος της αμφισβήτησης ή προσφυγής σε αποφάσεις που προκύπτουν από αλγορίθμους.

Προσπάθειες για τον περιορισμό των επιπτώσεων στην ασφάλεια και στην ιδιωτικότητα, έχουν υλοποιηθεί και σε χώρες της Ασίας αλλά και της Αφρικής. Η Κίνα

και η Σιγκαπούρη είναι δυο παραδείγματα χωρών που έχουν θεσπίσει τοπικό πλαίσιο διαχείρισης της Τεχνητής Νοημοσύνης. Η Κίνα υιοθέτησε το «New Generation AI Development Plan» και το ρυθμιστικό μέτρο «Measures for the Management of Generative AI Services», στοχεύοντας στον κρατικό έλεγχο και την κανονιστική συμμόρφωση των παρόχων ΤΝ. Η Σιγκαπούρη, από την άλλη, ανέπτυξε το AI Verify framework, ένα σύνολο εργαλείων που επιτρέπει στις επιχειρήσεις να ελέγχουν την απόδοση των συστημάτων τους σε σχέση με αρχές όπως η δικαιοσύνη και η διαφάνεια, ενώ αντίστοιχο σκοπό έχει και το Project Moonshot που επιθεωρεί και αξιολογεί τα μεγάλα γλωσσικά μοντέλα. Η Επιτροπή Προστασίας Δεδομένων της Σιγκαπούρης (PDPC) έχει εκδώσει συμβουλευτικές οδηγίες για τη χρήση προσωπικών δεδομένων σε συστήματα συστάσεων και αποφάσεων AI.

Στην Αφρική, έχει τεθεί σε εφαρμογή ένα ρυθμιστικό πλαίσιο υπό διαμόρφωση για τη διαχείριση της τεχνητής νοημοσύνης, το οποίο προωθείται κυρίως από την Αφρικανική Ένωση (AU). Βάσει αυτού, έχει υιοθετηθεί μια προσέγγιση βασισμένη σε στόχους (goals-based), με 55 κράτη μέλη να συμμετέχουν στη διαμόρφωση της στρατηγικής της. Μετά από πρωτοβουλία κρατικών σωμάτων και οντοτήτων που σχετίζονται με τον Πληροφορική, η AU συνέστησε μια ομάδα, υπό την προεδρία της Αιγύπτου, με στόχο τον καθορισμό μιας στρατηγικής για την Τεχνητή Νοημοσύνη για ολόκληρη την Ήπειρο. Οι στόχοι που έχουν τεθεί αντικατοπτρίζονται στις αρχές που ορίζονται στη Σύμβαση της Αφρικανικής Ένωσης για την Κυβερνοασφάλεια και την Προστασία των Δεδομένων Προσωπικού Χαρακτήρα (Σύμβαση του Malabo). Η σύμβαση αυτή αποτελεί ένα κεντρικό ρυθμιστικό κείμενο και εστιάζει:

- στην Προστασία της ιδιωτικής ζωής και των δεδομένων (Privacy and Data Protection)
- στο Ασφαλές ηλεκτρονικό εμπόριο (Secure e-Commerce)
- στην Κυβερνοασφάλεια και στο κυβερνοέγκλημα (Cybersecurity and Cybercrime)

Η Σύμβαση του Malabo προβλέπει τη διερεύνηση της χρήσης της ΤΝ, τη λήψη μέτρων για τη διασφάλιση επαρκών πόρων για τα εγχώρια πλαίσια προστασίας δεδομένων και τη δημιουργία περιφερειακών οργάνων για την παρακολούθηση της εφαρμογής της. (ASPEN Institute, 2023)

Η νομική πτυχή της Τεχνητής Νοημοσύνης, και ιδιαίτερα της GenAI, παραμένει θολή, καθώς τα περισσότερα κανονιστικά πλαίσια δεν έχουν ακόμα προσαρμοστεί πλήρως στις ιδιαιτερότητες της τεχνολογίας. Ανακύπτουν ζητήματα όπως η ευθύνη για το παραγόμενο περιεχόμενο, η προστασία της ιδιωτικότητας και η ρύθμιση της χρήσης

της από κυβερνήσεις, επιχειρήσεις και πολίτες. Κοινή πρόκληση για όλα τα κανονιστικά μοντέλα αποτελεί η ισορροπία ανάμεσα στην προστασία της ιδιωτικότητας και τη χρηστικότητα των δεδομένων. Η υπερβολική ρύθμιση ενδέχεται να περιορίσει τις δυνατότητες καινοτομίας, ενώ η ανεπαρκής ρύθμιση μπορεί να οδηγήσει σε εκτεταμένες παραβιάσεις θεμελιωδών δικαιωμάτων. Η εξέλιξη του ρυθμιστικού πλαισίου για την Τεχνητή Νοημοσύνη οφείλει να ακολουθεί μια προσέγγιση «διαχείρισης βάσει ρίσκου» (risk-based governance), ενσωματώνοντας τόσο νομικά όσο και τεχνικά μέσα για την οικοδόμηση εμπιστοσύνης στις τεχνολογίες τεχνητής νοημοσύνης.

1.3 Προκλήσεις διαχείρισης ρίσκου Τεχνητής Νοημοσύνης

Η διαχείριση των κινδύνων της Τεχνητής Νοημοσύνης παρουσιάζει ιδιαίτερες προκλήσεις που την διαφοροποιούν από την διαδικασία που εφαρμόζεται σε πληροφοριακά συστήματα. Οι προκλήσεις αυτές εκτείνονται σε όλο τον κύκλο ζωής ενός τέτοιου συστήματος, από τον σχεδιασμό και την ανάπτυξη έως την εγκατάσταση, τη λειτουργία και τη συντήρηση του, και απαιτούν μια ολιστική προσέγγιση για την αποτελεσματική αντιμετώπισή τους που να δίνει έμφαση στην πρόσληψη και λιγότερο στην αντιμετώπιση.

Μια από τις σημαντικότερες ιδιαιτερότητες είναι η εγγενής πολυπλοκότητα και η αδιαφάνεια των συστημάτων AI («μαύρο κουτί» ή black box problem). Σε πολύπλοκα μοντέλα, όπως τα νευρωνικά δίκτυα βαθιάς μάθησης, ακόμη και οι ίδιοι οι δημιουργοί συχνά δεν είναι σε θέση να εξηγήσουν με σαφήνεια πώς το σύστημα κατέληξε σε ένα συγκεκριμένο αποτέλεσμα. Αυτή η έλλειψη διαφάνειας δημιουργεί σημαντικά εμπόδια: υπονομεύει την εμπιστοσύνη των χρηστών, δυσκολεύει τον εντοπισμό και τη διόρθωση σφαλμάτων, αυξάνει την πιθανότητα εκμετάλλευσης ευπαθειών από κακόβουλους παράγοντες και έρχεται σε σύγκρουση με νομικές απαιτήσεις όπως το «δικαίωμα ενημέρωσης και διαφάνειας» που προβλέπεται από το GDPR. (SNOWFLAKE, 2025)

Εξίσου ιδιαίτερη είναι η δυναμική και αναδυόμενη φύση των κινδύνων. Σε αντίθεση με το παραδοσιακό software που αναπτύσσεται, τα συστήματα ΤΝ, και ιδιαίτερα όσα ενσωματώνουν τη συνεχή μάθηση (continuous learning), μπορούν να αλλάζουν συμπεριφορά μετά την εγκατάστασή τους. Εξαιτίας αυτού, προκύπτουν φαινόμενα

όπως η μετατόπιση μοντέλου και έννοιας (model and concept drift), όπου η απόδοση υποβαθμίζεται με την πάροδο του χρόνου λόγω αλλαγής των δεδομένων ή λόγω εξωτερικής παρέμβασης με αποτέλεσμα οι κακόβουλες οντότητες να μπορούν να επηρεάσουν τις προβλέψεις ενός μοντέλου. Ως εκ τούτου, οι παραδοσιακές, στατικές εκτιμήσεις κινδύνου δεν επαρκούν· η διαχείριση πρέπει να είναι μια συνεχής, επαναλαμβανόμενη διαδικασία, που να περιλαμβάνει παρακολούθηση, αξιολόγηση και επικαιροποίηση του μοντέλου αλλά και των σχετικών κινδύνων.

Επιπλέον, προκύπτουν νέες ευκαιρίες επίθεσης και ειδικές ευπάθειες. Επιθέσεις όπως η δηλητηρίαση δεδομένων (data poisoning) μπορούν να υπονομεύσουν τη λειτουργία του μοντέλου εισάγοντας αλλοιωμένα δεδομένα εκπαίδευσης, ενώ οι adversarial attacks βασίζονται σε μικρές, εσκεμμένες παραποιήσεις των δεδομένων εισόδου για να εξαπατήσουν το σύστημα. Επιπλέον, η κλοπή μοντέλου (model stealing) και οι επιθέσεις εξαγωγής πληροφοριών (inference attacks), όπως η αναστροφή μοντέλου (model inversion) ή η membership inference, αποτελούν εξειδικευμένες μορφές απειλών που αναδεικνύουν την ανάγκη για εξειδικευμένες στρατηγικές άμυνας που, ίσως να μην είχαν εντοπιστεί ή να μην είχαν συμπεριληφθεί σε αρχικό καταγραφή απειλών. (OWASP, 2025)

Η φύση της συγκεκριμένης τεχνολογίας με επίκεντρο τα δεδομένα καθιστά κρίσιμη τη διαχείριση ζητημάτων που σχετίζονται με τη μεροληψία, την ποιότητα και το απόρρητο των δεδομένων. Τα δεδομένα εκπαίδευσης ενδέχεται να ενισχύουν υπάρχουσες κοινωνικές προκαταλήψεις, οδηγώντας σε διακρίσεις και άδικες αποφάσεις. Παράλληλα, η ανάγκη για τεράστιους όγκους δεδομένων αυξάνει τον κίνδυνο παραβίασης του απορρήτου, είτε μέσω διαρροών είτε μέσω επιθέσεων εξαγωγής. Η εξισορρόπηση μεταξύ της χρησιμότητας των δεδομένων και της προστασίας της ιδιωτικότητας παραμένει μια συνεχής πρόκληση που απαιτεί συνδυασμό τεχνικών λύσεων (όπως differential privacy) και θεσμικών μηχανισμών (OECD, 2024).

Σε οργανωτικό επίπεδο, η διαχείριση κινδύνων της τεχνητής Νοημοσύνης συναντά δυσκολίες που σχετίζονται με την ωριμότητα και την κουλτούρα της διαδικασίας διαχείρισης κινδύνου. Πολλοί οργανισμοί δεν διαθέτουν την απαιτούμενη τεχνογνωσία ή τα κατάλληλα εργαλεία ή δεν έχουν προχωρήσει στην απαραίτητη προετοιμασία για να αντιμετωπίσουν τους ιδιαίτερους κινδύνους των συστημάτων αυτών. Επιπλέον, ο καταμερισμός ευθυνών μεταξύ των ρόλων και των διαφορετικών τομέων που εμπλέκονται στην ομαλή λειτουργία ενός μοντέλου, μεταξύ προγραμματιστών, φορέων

υλοποίησης, χρηστών και άλλων εμπλεκομένων — είναι θολός. Στο ίδιο πλαίσιο, η εξασφάλιση ανθρώπινης εποπτείας είναι μια συνεχιζόμενη πρόκληση, καθώς υπάρχει πάντα ο κίνδυνος της «αυτοματοποιημένης μεροληψίας» (automation bias), δηλαδή της υπερβολικής εμπιστοσύνης στις αποφάσεις που λαμβάνει το σύστημα. Τέλος, εντοπίζεται δυσκολία στην υποκειμενικότητα του ζητήματος. Παρόλη την συνολική προσπάθεια δημιουργίας κοινών πλαισίων και πολιτικών ή διαδικασιών διαχείρισης, δεν έχει καθιερωθεί ένας κοινά αποδεκτός τρόπος μέτρησης και απόδοσης της εποπτείας των συστημάτων αυτών. Η ανθρώπινη παρέμβαση μπορεί να πληγεί από την αποκαλούμενη αυτοματοποιημένη μεροληψία, όπου ο επόπτης δείχνει ιδιαίτερη εμπιστοσύνη στο μοντέλο με αποτέλεσμα να μην δίνει την προσοχή που απαιτείται.

Ακόμα και στην ίδια την διαδικασία διαχείρισης κινδύνων, δεν υπάρχει επαρκής ωριμότητα ώστε να είναι εφικτή η αντικειμενική και γενικά αποδεκτή, αποτύπωση των κινδύνων. Το γεγονός ότι το AI Act και αντίστοιχα νομικά πλαίσια βρίσκονται ακόμη σε στάδιο υιοθέτησης και αρχικής εφαρμογής, δείχνει ότι το πλαίσιο παραμένει υπό ανάπτυξη χωρίς να υπάρχει μια διαδεδομένη σταθερά.

Στην παρούσα εργασία θα γίνει αποτύπωση των ρίσκων χρήσης των Συστημάτων Τεχνητής Νοημοσύνης σε ζητήματα που αφορούν την ασφάλεια και την ιδιωτικότητα των δεδομένων προσωπικού χαρακτήρα.

2. Κίνδυνοι Ασφάλειας Συστημάτων Τεχνητής Νοημοσύνης

Αφού έχει αναπτυχθεί ο λόγος που η Τεχνητή Νοημοσύνη παρουσιάζει νέους κινδύνους, είναι εξίσου σημαντικό να αναδειχθούν οι κίνδυνοι αυτοί. Από την στιγμή που μια εξ ολοκλήρου νέα τεχνολογία ενσωματώνεται και εντάσσεται σε κρίσιμες υποδομές, επιχειρησιακές λειτουργίες και εφαρμογές, είναι εμφανές πως θα προκύψουν και νέοι κίνδυνοι οι οποίοι διαφοροποιούνται τελείως από το σύνολο τους ή θα προκύψουν διαφορετικές εκδοχές ήδη υπαρχόντων κινδύνων, οι οποίοι εν τέλει θα πρέπει να αντιμετωπιστούν ως εξ ολοκλήρου νέοι. Ενώ οι «κλασικές» μορφές απειλών – όπως επιθέσεις με κακόβουλο λογισμικό, παραβιάσεις δικτύων και η μη εξουσιοδοτημένη πρόσβαση σε συστήματα – στοχεύουν κυρίως σε πληροφοριακά συστήματα και δεδομένα, τα συστήματα AI εισάγουν ένα πιο περίπλοκο και δυναμικό πεδίο επιθέσεων. Η φύση τους, που βασίζεται σε μεγάλα σύνολα δεδομένων και πολύπλοκους αλγορίθμους μάθησης, τα καθιστά ιδιαίτερα ευάλωτα σε νέους τύπους κινδύνων για την αντιμετώπιση των οποίων απαιτείται η προσαρμογή των υφιστάμενων μέτρων διαχείρισης ή η δημιουργία καινούργιων.

Ένας από τους βασικούς λόγους που η Τεχνητή Νοημοσύνη εισάγει νέους κινδύνους είναι η εξάρτησή της από δεδομένα. Τα μοντέλα μηχανικής μάθησης εκπαιδεύονται με τεράστιους όγκους δεδομένων, τα οποία συχνά περιέχουν προσωπικές ή ευαίσθητες πληροφορίες. Η αλλοίωση αυτών των δεδομένων, μέσω επιθέσεων τύπου data poisoning, μπορεί να οδηγήσει σε σφάλματα ή σε μεροληψία του μοντέλου (Goldblum et al., 2021). Καθώς τα μοντέλα, συλλέγουν δεδομένα από το διαδίκτυο χωρίς να υφίσταται κάποιο φίλτρο στην διαδικασία συλλογής, δημιουργείται μια περίπλοκη εφοδιαστική αλυσίδα AI (AI supply chain), που περιλαμβάνει δεδομένα, προ-εκπαιδευμένα μοντέλα (foundation models), βιβλιοθήκες λογισμικού και APIs. Κάθε στοιχείο αυτής της αλυσίδας αποτελεί πιθανό σημείο ευπάθειας. Ένα "δηλητηριασμένο" μοντέλο από έναν προμηθευτή ή μια ευάλωτη βιβλιοθήκη λογισμικού μπορεί να θέσει σε κίνδυνο ολόκληρο το σύστημα. Αυτό δημιουργεί κινδύνους τόσο για την ασφάλεια όσο και για την ακεραιότητα των αποτελεσμάτων. (CSA Singapore, 2024)

Επιπλέον, η ευπάθεια σε adversarial attacks διαφοροποιεί την συγκεκριμένη τεχνολογία από τα παραδοσιακά πληροφοριακά συστήματα. Ελάχιστες μεταβολές στα δεδομένα εισόδου – οι οποίες είναι ιδιαίτερα δύσκολο να εντοπιστούν - μπορούν να

διαταράξουν ένα σύστημα που βρίσκεται στην διαδικασία εκπαίδευσης ή λήψης αποφάσεων, οδηγώντας το σε λανθασμένες αποφάσεις ή συμπεράσματα (Goodfellow, 2018). Σε εφαρμογές υψηλού ρίσκου – όπως τα αυτόνομα οχήματα ή οι βιομετρικοί μηχανισμοί ασφαλείας – τέτοιες επιθέσεις δύνανται να προκαλέσουν σοβαρές επιπτώσεις στη δημόσια ασφάλεια.

Τέλος, εισάγονται νέες μορφές επιθέσεων που στοχεύουν το ίδιο το μοντέλο και όχι αποκλειστικά τα δεδομένα ή την υποδομή. Επιθέσεις όπως η εξαγωγή μοντέλου (*model extraction*) ή η αντιστροφή μοντέλου (*model inversion*) μπορούν να οδηγήσουν σε παραβιάσεις πνευματικής ιδιοκτησίας, διαρροή ευαίσθητων δεδομένων και απώλεια ανταγωνιστικού πλεονεκτήματος μεταξύ άλλων.

Παρακάτω αναπτύσσονται τα είδη των κινδύνων ασφάλειας που υφίστανται σε συστήματα Τεχνητής Νοημοσύνης.

1.1 Κίνδυνοι κατά τον χρόνο ανάπτυξης (Development Phase)

Κατά την φάση ανάπτυξης του μοντέλου ή της εφαρμογής, οι κίνδυνοι εστιάζουν κυρίως στην χειραγώγηση των δεδομένων και του μοντέλου πριν αυτό τεθεί σε λειτουργία. Αυτό περιλαμβάνει τα περιβάλλοντα ανάπτυξης και εκπαίδευσης και την εφοδιαστική αλυσίδα. Η συλλογή και η τήρηση δεδομένων αποτελούν σημαντικές διεργασίες και είναι αυτές που αναδεικνύουν και τους περισσότερους κινδύνους. Οι ιδιαιτερότητες που χαρακτηρίζουν την φάση της ανάπτυξης αφορούν κυρίως τα δεδομένα, τα οποία αποτελούν πραγματικά δεδομένα που εν δυνάμει είναι προσωπικά ή ακόμα και ευαίσθητα. Σε πολλές περιπτώσεις, η χρήση ψευδών (dummy) δεδομένων δεν μπορεί να εφαρμοστεί, καθώς δεν θα είναι αποτελεσματική η εκπαίδευση του μοντέλου. Παράλληλά εντοπίζονται ιδιαιτερότητες και στην τήρηση των συστατικών ενός συστήματος AI, όπως ο κώδικας, οι παραμετροποιήσεις και τα ίδια τα δεδομένα. Τέλος όπως αναφέρθηκε και παραπάνω, είναι σημαντικό να αναλογιστούμε και τους κινδύνους της εφοδιαστικής αλυσίδας, όπως το μέσο από το οποίο θα συλλεχθούν τα δεδομένα ή το μοντέλο που θα χρησιμοποιηθεί. (OWASP, 2025)

1.1.1 Data Poisoning

Η ακεραιότητα των δεδομένων που θα εκπαιδεύσουν ένα μοντέλο είναι ιδιαίτερα κρίσιμη. Οποιαδήποτε αλλοίωση μπορεί να επηρεάσει δραστικά την συμπεριφορά ενός

συστήματος. Στόχος της Data poisoning Επίθεσης (Δηλητηρίαση Δεδομένων), έχει ως στόχο να «μολύνουν», δηλαδή να παραποιήσουν τα δεδομένα ή να εισάγουν ψευδή και κακόβουλα δεδομένα στο στάδιο εκπαίδευσης ενός μοντέλου, οδηγώντας το σε εσφαλμένες προβλέψεις. (Goldblum, 2021). Μέσω αυτών μπορεί να εισαχθούν επιτηδευμένες προκαταλήψεις (Bias Injection) στον τρόπο που λαμβάνει αποφάσεις το μοντέλο, οδηγώντας σε διακρίσεις. Οι επιθέσεις data poisoning είναι ιδιαίτερα επικίνδυνες καθώς δύσκολα εντοπίζονται. Οι αλλοιώσεις στα δεδομένα μπορεί να είναι μικρές, γεγονός που καθιστά την ανίχνευσή τους περίπλοκη διαδικασία. Επιπλέον, οι αλγόριθμοι μηχανικής μάθησης συχνά τείνουν να «γενικεύουν» τα μοτίβα των δεδομένων εκπαίδευσης, με αποτέλεσμα ένα μικρό πλήθος αλλοιωμένων παραδειγμάτων να μπορεί να επηρεάσει δυσανάλογα το τελικό μοντέλο και την εκπαιδευτική διαδικασία.

1.1.2 Backdoors

Πέρα από το λανθασμένο εκπαιδευτικό μοντέλο, μέσω κακόβουλων δεδομένων, ένας επιτιθέμενος χρήστης έχει την δυνατότητα να δημιουργήσει Backdoors προς το σύστημα. Τα Backdoors αποτελούν διόδους τις οποίες μπορεί να εκμεταλλευτεί μια κακόβουλη οντότητα ώστε να αποκτήσει πρόσβαση και να προχωρήσει σε άλλες επιθέσεις και κακόβουλες ενέργειες σε μεταγενέστερο χρόνο. Τέτοια σημεία μπορούν να παραμείνουν ανενεργά για μεγάλο χρονικό διάστημα, διαφεύγοντας του εντοπισμού ακόμη και μετά από τυπικές διαδικασίες αξιολόγησης ασφάλειας. Η ενεργοποίηση των Backdoors συνήθως συνοδεύεται από κάποιο «μοχλό πυροδότησης» (trigger) όπως κάποιο prompt το οποίο συναντάμε κυρίως στην Φάση εκτέλεσης που θα αναλύσουμε παρακάτω.

1.1.3 Διαρροή δεδομένων

Εξίσου σημαντικός κίνδυνος, που επηρεάζει άμεσα και την ασφάλεια αλλά και την ιδιωτικότητα, είναι η διαρροή δεδομένων. Είτε αναφερόμαστε σε προσωπικά και ευαίσθητα δεδομένα φυσικών προσώπων που μπορεί να χρησιμοποιούνται για την εκπαίδευση και την ανάπτυξη του μοντέλου, είτε αναφερόμαστε στο ίδιο το μοντέλο και τον τρόπο που λαμβάνει τις αποφάσεις και επεξεργάζεται τα δεδομένα, υφίσταται πάντα ο κίνδυνος υποκλοπής και διαρροής. Αποτέλεσμα αυτού, σοβαρές νομικές κυρώσεις, κίνδυνος για την ασφάλεια των φυσικών προσώπων ανάλογα με την φύση

των δεδομένων και ζητήματα που αφορούν την φήμη και την αξιοπιστία. (OWASP, 2025)

1.2 Κίνδυνοι κατά την χρήση (Through Use Phase)

Στο στάδιο αυτό αναφερόμαστε σε κινδύνους που μπορεί να αντιμετωπίσει ένα σύστημα Τεχνητής Νοημοσύνης, το οποίο έχει ήδη εκπαιδευτεί και βρίσκεται στο παραγωγικό περιβάλλον. Σε αντίθεση με τις επιθέσεις κατά τον χρόνο ανάπτυξης (που στοχεύουν στο training set ή την εκπαίδευση), οι through-use επιθέσεις εκμεταλλεύονται πιθανές ευπάθειες στις διεπαφές χρήστη (UIs, APIs) όπως εντοπίζουμε σε πληθώρα διαφορετικών συστημάτων, την ανοχή των μοντέλων σε μεταβολές στα δεδομένα εισόδου, καθώς και την πληροφορία που αποκαλύπτεται από τις απαντήσεις του μοντέλου.

1.2.1 Model Inversion (Αντιστροφή Μοντέλου)

Η επίθεση τύπου **model inversion** συνίσταται στην εκμετάλλευση της πρόσβασης στις εξόδους (π.χ. πιθανότητες, confidence scores) ενός εκπαιδευμένου μοντέλου με στόχο την ανακατασκευή — είτε μερική είτε πλήρης — χαρακτηριστικών ή ακόμη και ολόκληρων δειγμάτων που ανήκουν στα δεδομένα εκπαίδευσης. Ο επιτιθέμενος χρησιμοποιεί τις απαντήσεις του μοντέλου ως «πληροφοριακό κανάλι» για να αναστρέψει τη διαδικασία μάθησης και να αποκτήσει ευαίσθητα δεδομένα (π.χ. ιατρικά χαρακτηριστικά, φωτογραφίες προσώπων, προσωπικά στοιχεία) που θεωρούνταν προστατευμένα. Αυτή η μορφή επίθεσης είναι ιδιαίτερα προβληματική όταν το μοντέλο παρέχει πλούσια πληροφορία στην έξοδό του (Πχ. confidence scores) και όταν το μοντέλο έχει υπερεκπαιδευτεί σε συγκεκριμένα δείγματα, καθώς η «Εξοικείωση» αυξάνει την πληροφορία που αφορά την αποδοτικότητα ως προς τα αρχικά δεδομένα. Οι πρακτικές συνέπειες περιλαμβάνουν παραβιάσεις ιδιωτικότητας, νομικά και ηθικά ζητήματα και απώλεια εμπιστοσύνης σε συστήματα που διαχειρίζονται προσωπικές πληροφορίες. Για να είναι εφικτή η μείωση του κινδύνου θα πρέπει να περιορίζεται η πληροφορία που αποκαλύπτεται από τις απαντήσεις του συστήματος (π.χ. παροχή μόνο κλάσης αντί για πλήρη scores) και χρήση τεχνικών όπως differential privacy, η οποία προστατεύει τα δεδομένα εκπαίδευσης ώστε να μην εφικτή η διαρροή τους από το σύστημα. (Das. B., 2024)

1.2.2 Inference Attacks (Επιθέσεις εξαγωγής συμπερασμάτων)

Οι επιθέσεις τύπου inference αποτελούν ένα ευρύ φάσμα τεχνικών στις οποίες ο επιτιθέμενος προσπαθεί να αντλήσει, από τις παρατηρήσεις που του παρέχει ένα μοντέλο (ή η συμπεριφορά του), ευαίσθητες πληροφορίες σχετικά με τα δεδομένα εκπαίδευσης ή τα χαρακτηριστικά των ατόμων που αυτά αφορούν. Σε αυτή την κατηγορία περιλαμβάνονται τόσο οι model inversion επιθέσεις που στοχεύουν στην ανακατασκευή δειγμάτων ή χαρακτηριστικών (βλ. προηγούμενη ενότητα), όσο και οι attribute inference και membership inference επιθέσεις, οι οποίες αντλούν συμπεράσματα για την παρουσία συγκεκριμένων χαρακτηριστικών ή για το αν ένα δείγμα συμμετείχε στο σύνολο των δεδομένων εκπαίδευσης.

Οι Membership Inference επιθέσεις εστιάζουν στο αν περιλαμβάνεται ένα συγκεκριμένο δείγμα στο training set του μοντέλου στόχου. Η επίθεση αξιοποιεί το γεγονός ότι τα μοντέλα, ιδιαίτερα όταν έχουν εκπαιδευτεί σε μεγάλο βαθμό, εμφανίζουν διαφορετική συμπεριφορά (π.χ. υψηλότερα confidence scores) σε δείγματα που έχουν δει κατά την εκπαίδευση σε σύγκριση με νέα ή πρωτόγνωρα δείγματα. Ο επιτιθέμενος συλλέγει για ένα πλήθος δειγμάτων τις εξόδους του μοντέλου, και στη συνέχεια εκπαιδεύει ένα επιπλέον βοηθητικό μοντέλο το οποίο μαθαίνει το μοτίβο της διαφοράς μεταξύ της συμμετοχής ή μη των δεδομένων με βάση αυτές τις εξόδους. Στην πραγματικότητα, ο επιτιθέμενος χρησιμοποιεί το βοηθητικό μοντέλο για να προβλέψει αν ένα νέο δείγμα συμμετείχε στα δεδομένα εκπαίδευσης. Η επίθεση είναι ιδιαίτερα αποτελεσματική όταν το μοντέλο αποκαλύπτει περισσότερη πληροφορία (π.χ. confidence scores αντί μόνο της τελικής ετικέτας) και όταν το μοντέλο υπερπροσαρμόζεται σε μεμονωμένα δείγματα.(Shokri et al., 2017).

Οι πρακτικές συνέπειες των επιθέσεων αυτών σχετίζονται με την αναγνώριση ότι ένα άτομο ή μια εγγραφή περιλαμβάνεται σε ένα training set, με αποτέλεσμα να μπορεί να εκθέσει ιδιωτικά δεδομένα (π.χ. συμμετοχή σε βάση δεδομένων για μια ασθένεια, χρήση μιας υπηρεσίας), ακόμη και αν το ίδιο το περιεχόμενο των δεδομένων δεν αποκαλύπτεται. Σε τομείς όπως η υγεία ή η νομική πληροφόρηση, τέτοιες αποκαλύψεις έχουν σοβαρές ηθικές και νομικές επιπτώσεις.

1.2.3 Model Extraction / Model Stealing

Η κλοπή ή εξαγωγή μοντέλου αναφέρεται στη διαδικασία κατά την οποία ένας επιτιθέμενος ανακατασκευάζει, προσεγγιστικά ή και σε μεγάλο βαθμό, τη λειτουργική συμπεριφορά ενός μοντέλου μηχανικής μάθησης αποκλειστικά μέσω της αλληλεπίδρασης με τις διεπαφές του (π.χ. prediction APIs). Στην πιο απλή μορφή της επίθεσης, ο επιτιθέμενος εισάγει ερωτήματα (queries) και συλλέγει τις απαντήσεις (labels, confidence scores,κλπ.), χρησιμοποιώντας τους συνδυασμούς εισόδου–εξόδου για να εκπαιδεύσει ένα υποκατάστατο (surrogate) μοντέλο που μιμείται την συμπεριφορά του αρχικού μοντέλου «θύματος». Σε πιο εξελιγμένα σενάρια, ειδικά σχεδιασμένες ακολουθίες ερωτήσεων (adaptive sampling, active learning) επιτρέπουν την αποδοτική ανάκτηση του τρόπου λειτουργίας του μοντέλου, ενώ διαφορετικές επιθέσεις της συγκεκριμένης κατηγορίας επιτρέπουν την εξαγωγή παραμέτρων της αρχιτεκτονικής του ή ακόμη και εκτιμήσεων για τη σημασία που αποδίδεται σε συγκεκριμένα δεδομένα με την παροχή εκτενέστερων πληροφοριών κατά την έξοδο (Papernot et al., 2017).

Η πρακτική αξία μιας τέτοιας εξαγωγής είναι πολλαπλή: ένας επιτιθέμενος αποκτά λειτουργικό αντίγραφο χωρίς να επωμιστεί το κόστος συλλογής δεδομένων ή εκπαίδευσης, μπορεί να παρακάμψει την ανάγκη πληρωμής μέσω ενός εμπορικού API, να χρησιμοποιήσει το κλεμμένο μοντέλο για τη δημιουργία πιο αποτελεσματικών adversarial παραδειγμάτων ενάντια στην αρχική υπηρεσία, ή να αναζητήσει ευπάθειες/αδυναμίες σε περιβάλλον που τώρα έχει πρόσβαση. Επιπλέον, η εξαγωγή μπορεί να οδηγήσει σε διαρροή εμπορικού απορρήτου και πνευματικής ιδιοκτησίας, πλήττοντας την επιχειρηματική αξία και την επενδυτική πρόθεση (Tramèr et al., 2016).

1.3 Κίνδυνοι κατά τον χρόνο εκτέλεσης (Run time Phase)

Οι επιθέσεις κατά το χρόνο εκτέλεσης στοχεύουν στο στάδιο όπου ένα μοντέλο Τεχνητής Νοημοσύνης λειτουργεί σε παραγωγικό περιβάλλον και εξυπηρετεί αιτήματα από χρήστες ή άλλα συστήματα. Σε αυτό το στάδιο οι επιτιθέμενοι εκμεταλλεύονται την ίδια τη διεπαφή χρήστη (APIs, web endpoints), την ευαισθησία των μοντέλων σε παραμορφώσεις εισόδων, την υποδομή που τα φιλοξενεί και τα metadata που παράγονται. Το αποτέλεσμα είναι ότι οι runtime επιθέσεις μπορούν να πλήξουν τη διαθεσιμότητα, την ακεραιότητα και την εμπιστευτικότητα μιας υπηρεσίας AI, συχνά με άμεσες επιχειρησιακές συνέπειες.

1.3.1 Prompt Injection Attacks

Τα Prompt Injection Attacks αποτελούν μια κρίσιμη κατηγορία επιθέσεων σε συστήματα Τεχνητής Νοημοσύνης, ιδιαίτερα σε μεγάλα γλωσσικά μοντέλα (LLMs) και σε agents που εκτελούν ενέργειες μέσω οδηγιών που παρέχονται με κείμενο. Σε αυτές τις επιθέσεις, ο επιτιθέμενος εισάγει κακόβουλες ή παραπλανητικές οδηγίες εντός πλαισίου εφαρμογής και λειτουργίας του μοντέλου, με στόχο να αλλάξει τη συμπεριφορά του και να παρακάμψει κανόνες ασφαλείας ή πολιτικές προστασίας δεδομένων. Η βασική υπόθεση είναι ότι το μοντέλο ακολουθεί οδηγίες που θεωρεί πως ανήκουν στο πλαίσιο λειτουργίας του, όπως system prompts, user input ή περιεχόμενο από εξωτερικές πηγές, και ότι ο επιτιθέμενος μπορεί να επηρεάσει αυτά τα δεδομένα, ώστε το μοντέλο να λειτουργήσει με τον τρόπο που εκείνος επιθυμεί.

Κατά το prompt injection, ο επιτιθέμενος εκμεταλλεύεται την ικανότητα του μοντέλου να ενσωματώνει οδηγίες μέσα στο prompt. Για παράδειγμα, ένα κακόβουλο prompt μπορεί να περιέχει οδηγίες όπως το να αποκαλύψει κρίσιμες πληροφορίες (πχ. API Keys) ή να αγνοήσει φίλτρα και πολιτικές ασφαλείας. Όταν το μοντέλο λαμβάνει το εισερχόμενο περιεχόμενο ως μέρος του context, μπορεί να εκτελέσει ανεπιθύμητες ενέργειες ή να παρέχει ευαίσθητες πληροφορίες τις οποίες κανονικά δεν θα «είχε την άδεια» να αποκαλύψει.

Οι συνέπειες των prompt injections ποικίλλουν από την αποκάλυψη εμπιστευτικών πληροφοριών μέχρι την υπονόμευση της αξιοπιστίας και ασφάλειας του συστήματος. Η δυνατότητα ενός επιτιθέμενου να παραπλανήσει το μοντέλο δημιουργεί κίνδυνο για διαρροές προσωπικών δεδομένων, παραβιάσεις άρθρων του GDPR, απώλεια εμπιστοσύνης των χρηστών και οικονομικές ή νομικές συνέπειες για τους οργανισμούς που διαχειρίζονται το σύστημα.

1.3.2 Adversarial / Evasion Επιθέσεις

Οι εν λόγω επιθέσεις πρόκειται για σκόπιμες και ιδιαίτερα δύσκολα διακριτές τροποποιήσεις των εισερχόμενων δεδομένων στα μοντέλα ή στα δεδομένα εκπαίδευσης που μπορούν να οδηγήσουν σε λανθασμένες ή ανεπιθύμητες προβλέψεις και αποτελέσματα, όσο λειτουργεί το μοντέλο. Οι κακόβουλες είσοδοι ονομάζονται αντιπαραθετικά παραδείγματα (adversarial examples). Η επιτυχία των επιθέσεων αυτών βασίζεται στο ότι τα μοντέλα — και ιδίως τα βαθιά νευρωνικά δίκτυα — δεν αντιλαμβάνονται τα δεδομένα με ανοικτό, ανθρώπινο τρόπο, αλλά αντιδρούν σε

λεπτομέρειες της μαθηματικής τους αναπαράστασης, με αποτέλεσμα μικρές παραμορφώσεις στα δεδομένα που επεξεργάζονται να μπορούν να προκαλέσουν μεγάλες αποκλίσεις στο αποτέλεσμα που θα παράξουν. Για παράδειγμα, μια ανεπαίσθητη αλλαγή σε μια εικόνα, αόρατη στο ανθρώπινο μάτι, μπορεί να οδηγήσει ένα σύστημα αναγνώρισης εικόνων να ταξινομήσει λανθασμένα ένα σήμα "STOP" ως σήμα ορίου ταχύτητας (Das. B., 2024) Μια σημαντική σημείωση είναι η διαφοροποίηση μεταξύ ψηφιακών και φυσικών επιθέσεων. Ψηφιακές επιθέσεις τροποποιούν απευθείας τα ψηφιακά δεδομένα (π.χ. το αρχείο εικόνας), ενώ φυσικές επιθέσεις μεταφράζονται σε παρεμβάσεις στον πραγματικό κόσμο — όπως αυτοκόλλητα σε πινακίδες δρόμου ή εκτυπώσεις που, όταν φωτογραφηθούν, οδηγούν ένα σύστημα όρασης σε λανθασμένη ανάγνωση

Οι Evasion επιθέσεις, αποτελεί μια υποκατηγορία των Adversarial Επιθέσεων, και έχουν ως στόχο το μοντέλο να δώσει λανθασμένη πρόβλεψη ή απόφαση αλλά όχι να αλλάξει το ίδιο το μοντέλο ή τα δεδομένα εκπαίδευσης.

1.3.3 Επιθέσεις άρνησης υπηρεσίας (DoS / resource exhaustion)

Τέτοιου είδους επιθέσεις κατά την φάση εκτέλεσης επιδιώκουν να διακόψουν ή να επιβραδύνουν τη λειτουργία του συστήματος. Αυτό περιλαμβάνει μαζικά ή ειδικά σχεδιασμένα queries που προκαλούν υπερκατανάλωση υπολογιστικών πόρων (CPU/GPU) ή μνήμης, ή εκμεταλλεύεται αδύναμες συνιστώσες της υποδομής (π.χ. ανεπαρκής scaling, αδύναμα timeouts).

3. Κίνδυνοι Ιδιωτικότητας Συστημάτων Τεχνητής Νοημοσύνης

Όπως έχει ήδη αναφερθεί στην παρούσα εργασία, πέραν της ασφάλειας δεδομένων, η ιδιωτικότητα των υποκειμένων είναι από τα πλέον μείζονα ζητήματα που απασχολούν το ρυθμιστικό πλαίσιο αλλά και που επηρεάζει η εφαρμογή και η λειτουργία των συστημάτων τεχνητής νοημοσύνης, δεδομένου ότι η λειτουργία τους βασίζεται σε μεγάλο βαθμό στη συλλογή, επεξεργασία και ανάλυση μεγάλων όγκων δεδομένων. Σε αντίθεση με άλλες τεχνολογίες πληροφορικής, το AI εξαρτάται άμεσα από την ποιότητα και την ποσότητα των δεδομένων, τα οποία συχνά περιλαμβάνουν προσωπικές ή και ευαίσθητες πληροφορίες. Αυτό δημιουργεί έναν νέο χώρο προκλήσεων που σχετίζονται με την προστασία της ιδιωτικής ζωής και την εξασφάλιση της συμμόρφωσης με τα υφιστάμενα νομικά και κανονιστικά πλαίσια. Οι κίνδυνοι ιδιωτικότητας δεν περιορίζονται μόνο στη φάση συλλογής των δεδομένων, αλλά επεκτείνονται σε όλα τα στάδια του κύκλου ζωής ενός τέτοιου συστήματος, από την εκπαίδευση των μοντέλων έως τη χρήση και τη διάθεσή τους σε παραγωγικό περιβάλλον. Για παράδειγμα, φαινόμενα όπως η διαρροή δεδομένων μέσω των ίδιων των μοντέλων ή η δυνατότητα εξαγωγής πληροφοριών μέσω εξειδικευμένων επιθέσεων (όπως το membership inference και το model inversion, όπως είδαμε και στην προηγούμενη ενότητα) δείχνουν ότι η ιδιωτικότητα δεν μπορεί να θεωρείται δεδομένη. Παράλληλα, η ενσωμάτωση της τεχνολογίας σε κρίσιμους τομείς – όπως η υγεία, οι χρηματοοικονομικές υπηρεσίες και η δημόσια διοίκηση – εντείνει τις πιθανές επιπτώσεις που μπορεί να έχει μια παραβίαση της ιδιωτικότητας τόσο για τα άτομα όσο και για τους οργανισμούς.

3.1 Κίνδυνοι από την Εκπαίδευση Μοντέλων

Η εκπαίδευση των συστημάτων Τεχνητής Νοημοσύνης στηρίζεται σε μεγάλα σύνολα δεδομένων, τα οποία συχνά περιλαμβάνουν προσωπικές ή ευαίσθητες πληροφορίες. Αν και η εκπαίδευση αποτελεί αναγκαία διαδικασία για τη βελτίωση της απόδοσης των μοντέλων, εμπεριέχει σημαντικούς κινδύνους, οι οποίοι αφορούν τόσο την ποιότητα και το περιεχόμενο των δεδομένων εκπαίδευσης όσο και τις επιθέσεις που στοχεύουν στην εκμετάλλευση των μοντέλων μετά την ανάπτυξή τους, όπως αναπτύχθηκε στο Κεφάλαιο 2.

Ένα από τα σοβαρότερα ζητήματα είναι η διαρροή προσωπικών δεδομένων μέσω των ίδιων των μοντέλων. Κατά τη διαδικασία εκπαίδευσης, τα μοντέλα μπορεί να

απομνημονεύσουν συγκεκριμένα παραδείγματα από τα δεδομένα εκπαίδευσης, οδηγώντας σε κίνδυνο επαναταυτοποίησης χρηστών ή αποκάλυψης ευαίσθητων πληροφοριών (Carlini et al., 2021). Αυτό είναι ιδιαίτερα κρίσιμο σε εφαρμογές που επεξεργάζονται δεδομένα υγείας ή οικονομικά δεδομένα, όπου η διαρροή ακόμη και μικρών αποσπασμάτων πληροφορίας μπορεί να έχει σοβαρές συνέπειες για τα υποκείμενα των δεδομένων. Επιθέσεις όπως Membership Inference Attacks, Model Inversion, Attacks, Data Extraction Attacks, έχουν άμεση επίδραση τόσο στην ασφάλεια, όσο και στην ιδιωτικότητα των υποκειμένων.

Τέλος, η εκπαίδευση μοντέλων σε περιβάλλοντα με ανεπαρκή ασφάλεια ενέχει τον κίνδυνο παραβίασης των δεδομένων εκπαίδευσης από μη εξουσιοδοτημένους χρήστες. Δεδομένου ότι τα σύνολα εκπαίδευσης συχνά περιλαμβάνουν δεδομένα από πολλαπλές πηγές, η απώλεια ή η κλοπή τους μπορεί να εκθέσει όχι μόνο ιδιωτικές πληροφορίες αλλά και την ίδια τη λειτουργικότητα του συστήματος, αφού τα δεδομένα αυτά αποτελούν στρατηγικό περιουσιακό στοιχείο για τους οργανισμούς.

3.2 Κίνδυνοι από τη Συλλογή και Χρήση Δεδομένων

Ένας από τους θεμελιώδεις κινδύνους προέρχεται από τον τρόπο που τα συστήματα Τεχνητής Νοημοσύνης συλλέγουν και επεξεργάζονται τεράστιες ποσότητες δεδομένων, συχνά προσωπικών και ευαίσθητων για να μπορέσουν να εκπαιδευτούν και να ικανοποιήσουν τον σκοπό τους και να έχουν υψηλή απόδοση. Ωστόσο, η αυξανόμενη εξάρτηση από δεδομένα φέρνει στην επιφάνεια σημαντικούς κινδύνους για την ιδιωτικότητα, που εκτείνονται από την αθέμιτη συλλογή έως την επαναχρησιμοποίησή τους για διαφορετικούς σκοπούς και την παραβίαση βασικών δικαιωμάτων των φυσικών προσώπων που ανήκουν τα δεδομένα.

Ένα από τα βασικότερα προβλήματα είναι η εκτενής συλλογή δεδομένων. Σε πολλές περιπτώσεις, οργανισμοί συγκεντρώνουν πολύ περισσότερα δεδομένα από όσα είναι αναγκαία για τον επιδιωκόμενο σκοπό. Ωστόσο, αυτή η πρακτική αντίκειται στην αρχή της ελαχιστοποίησης δεδομένων που προβλέπεται από τον Γενικό Κανονισμό Προστασίας Δεδομένων. Ο κίνδυνος έγκειται κυρίως στην ακούσια έκθεση ευαίσθητων πληροφοριών που περιέχονται στα σύνολα δεδομένων ενός μοντέλου. που μπορεί να οδηγήσει σε παραβιάσεις του απορρήτου και νομικές ευθύνες. Η συλλογή όχι μόνο αυξάνει τον κίνδυνο παραβιάσεων, αλλά δημιουργεί και ευρύτερες ηθικές

ανησυχίες, καθώς τα υποκείμενα δεδομένων χάνουν τον έλεγχο των προσωπικών τους πληροφοριών. (Snowflake, 2025).

Παρόμοια, είναι συχνές οι περιπτώσεις που μοντέλα εκπαιδεύονται με δεδομένα τα οποία έχουν συλλεχθεί μαζικά από το διαδίκτυο (Scraping) παραβιάζοντας θεμελιώδεις αρχές της ιδιωτικότητας, καθώς η συλλογή πραγματοποιείται χωρίς την κατάλληλη ενημέρωση, χωρίς την απαραίτητη συγκατάθεση των ιδιοκτητών των δεδομένων και χωρίς να είναι ξεκάθαροι οι σκοποί για τους οποίους συλλέγονται εξαρχής αλλά και ο τρόπος με τον οποίον θα χρησιμοποιηθούν ή αν θα διαμοιραστούν με τρίτους πέραν από τους άμεσα εμπλεκόμενους. Συχνά, παρατηρείται πως ακόμα και οι ίδιοι οι προγραμματιστές που αναπτύσσουν τα λογισμικά δεν γνωρίζουν εκ των προτέρων όλες τις πιθανές χρήσεις των δεδομένων. (OVIC, 2018). Ως αποτέλεσμα, πλήττεται η διαφάνεια στις διαδικασίες συλλογής δεδομένων αλλά ενισχύεται και το πρόβλημα της δευτερογενούς χρήσης τους, δηλαδή της χρήσης των δεδομένων για σκοπούς που δεν είχαν προβλεφθεί ή εγκριθεί κατά την αρχική συλλογή τους. .(38. Shahriar, Sakib, et al., 2023).

Η ανεπαρκής προστασία δεδομένων κατά το στάδιο της συλλογής και χρήσης τους ενδέχεται να έχει σημαντικές κοινωνικές και νομικές επιπτώσεις. Η απώλεια ιδιωτικότητας, διαφάνειας και ορθής χρήσης μπορεί να οδηγήσει σε φαινόμενα διάκρισης, στοχοποίησης ή αποκλεισμού χρηστών, ενώ για τους οργανισμούς συνεπάγεται αυξημένο νομικό ρίσκο και πιθανές κυρώσεις για παραβίαση του GDPR ή άλλων σχετικών κανονισμών όπως το AI Act. Σχετικός με αυτό είναι και ο κίνδυνος της έννοιας της τοξικότητας και των ακατάλληλων προτύπων που μπορεί να εισαχθούν στο μοντέλο λόγω ανεπαρκούς φιλτραρίσματος των δεδομένων. Ιδιαίτερα τα μεγάλα γλωσσικά μοντέλα (LLMs) που εκπαιδεύονται σε δεδομένα διαδικτύου είναι επιρρεπή στο να μάθουν και να αναπαράγουν προσβλητικό, παραπλανητικό ή επικίνδυνο περιεχόμενο (Weidinger et al., 2022).

3.3 Κίνδυνοι από την Δημιουργία Συμπερασμάτων και Νέων δεδομένων

Η Τεχνητή Νοημοσύνη δεν επεξεργάζεται απλώς δεδομένα, αλλά παράγει και καινούργια μέσω συμπερασμάτων που μπορεί να προκύψουν από την ανάλυση της πληροφορίας που έχει συλλεχθεί, είτε για λόγους εκπαίδευσης, είτε για λόγους χρήσης. Αν και αυτή η δυνατότητα αποτελεί σημαντικό πλεονέκτημα για εφαρμογές όπως η ιατρική διάγνωση, η πρόβλεψη συμπεριφορών ή η προσωποποιημένη παροχή

υπηρεσιών, εγείρει σοβαρά ζητήματα ιδιωτικότητας. Αλγόριθμοι μηχανικής μάθησης μπορούν να παράξουν νέα δεδομένα που αποκαλύπτουν προσωπικές λεπτομέρειες. Για παράδειγμα, ένα σύστημα μπορεί να συμπεράνει την πολιτική τοποθέτηση, την κατάσταση της υγείας ή τον σεξουαλικό προσανατολισμό ενός ατόμου από φαινομενικά αθώα δεδομένα, όπως οι αγοραστικές του συνήθειες ή συμπεριφορές που παρουσιάζει στο διαδίκτυο. Αυτό παρακάμπτει τους παραδοσιακούς μηχανισμούς προστασίας, καθώς το άτομο δεν έχει δώσει ποτέ άμεσα αυτές τις πληροφορίες αλλά ούτε και την συναίνεσή του. (Solove, 2024)

Επιπρόσθετα, σε χώρες όπου δεν υπάρχει επαρκές ρυθμιστικό πλαίσιο, ή δεν έχει τεθεί ακόμα σε λειτουργία ή όπου η λειτουργία δεν ελέγχεται επαρκώς, ένας ιδιαίτερα σημαντικός κίνδυνος αποτελεί το profiling των υποκειμένων, δηλαδή η δημιουργία ενός συμπεριφορικού προφίλ με στόχο την εμπορική εκμετάλλευση, τη χειραγώγηση ή ακόμα και τη διάκριση εις βάρος συγκεκριμένων ομάδων. Αντίστοιχα, αναγνωρίζεται και ο κίνδυνος της δημιουργίας συμπερασμάτων μέσω της Φυσιογνωμίας ενός υποκειμένου (Εμφάνιση, φωνή κλπ.), δηλαδή η παραγωγή συμπερασμάτων για την προσωπικότητα, τα κοινωνικά και συναισθηματικά χαρακτηριστικά βάσει των φυσικών χαρακτηριστικών του. (24. Lee, Hao-Ping, et al., 2024) Σε ένα τέτοιο πλαίσιο, οι οργανισμοί ή οι κυβερνήσεις μπορούν να κατασκευάσουν λεπτομερή προφίλ που ξεπερνούν κατά πολύ τα δεδομένα που οι ίδιοι οι χρήστες έχουν παραχωρήσει, εντείνοντας τον κίνδυνο κοινωνικού αποκλεισμού και καταπάτησης θεμελιωδών δικαιωμάτων.

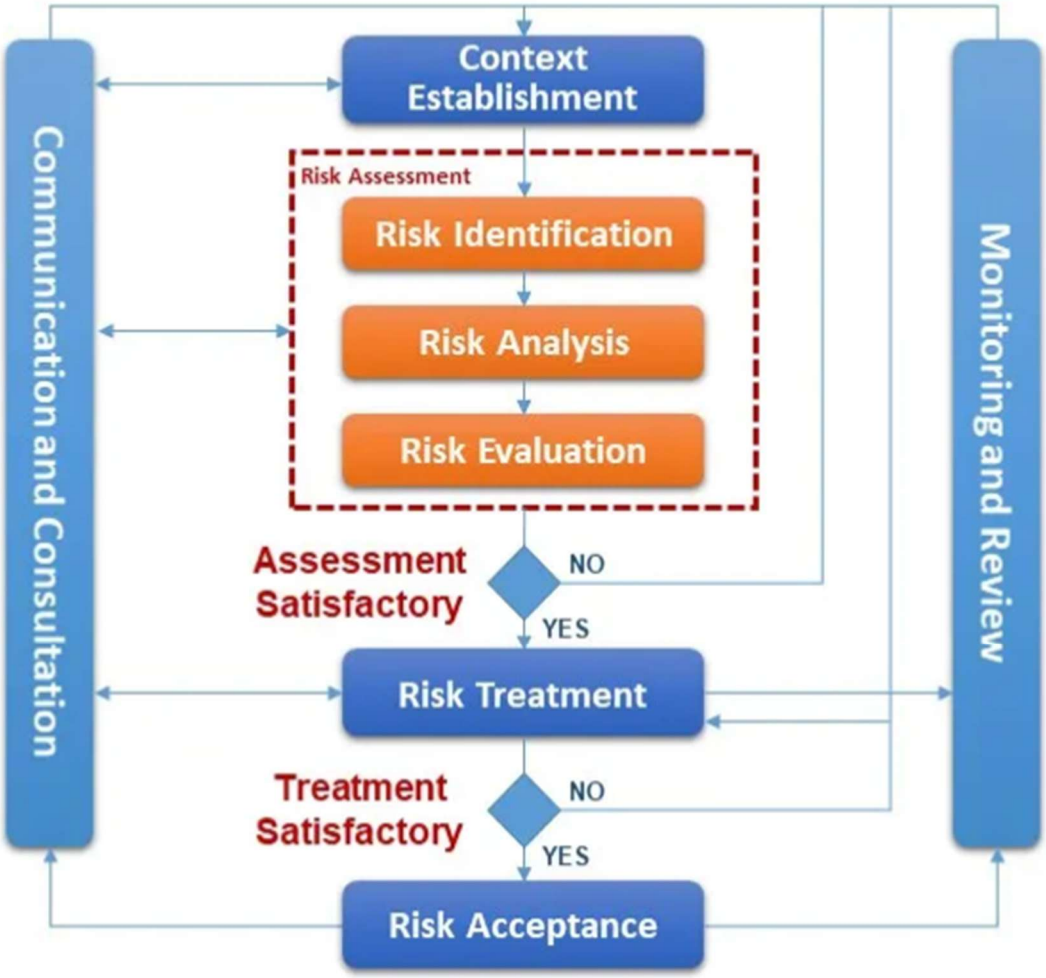
4. Διαχείριση Κινδύνων (Risk Management)

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν οι ιδιαιτερότητες της Τεχνητής Νοημοσύνης και οι νέοι κίνδυνοι που παρουσιάζονται όσον αφορά στην Ιδιωτικότητα και την ασφάλεια πληροφοριών. Για να υπάρξει ορθή διαχείριση των κινδύνων, είναι σημαντικό να αξιοποιηθεί ένα κατάλληλα σχεδιασμένο πλαίσιο, μέσω του οποίου θα μας δοθεί η δυνατότητα αναγνώρισης, ανάλυσης, τεκμηρίωσης και διαχείρισής τους. Το πλαίσιο αυτό μπορεί να το καλύψει μια διαδικασία διαχείρισης κινδύνων (Risk Management Procedure). Η διαδικασία αξιολόγησης κινδύνων (Risk Assessment Procedure) συνιστά τον πυρήνα της διαχείρισης κινδύνων και περιλαμβάνει τη δομημένη προσέγγιση για τον εντοπισμό απειλών, ευπαθειών και δυνητικών επιπτώσεων. Παραδοσιακά, η αξιολόγηση κινδύνων εφαρμόζεται σε τομείς όπως η ασφάλεια πληροφοριών, η επιχειρησιακή συνέχεια και η κανονιστική συμμόρφωση, με έμφαση σε πληροφοριακά συστήματα και υποδομές. Πλαίσια όπως τα ISO 31000 και ISO/IEC 27005 παρέχουν καθοδήγηση για την οργάνωση αυτής της διαδικασίας, υιοθετώντας μια συστηματική και επαναλαμβανόμενη προσέγγιση που βασίζεται στην μείωση και στην διαχείριση ενός κινδύνου.

Πλέον λόγω της αυξημένης πολυπλοκότητας των συστημάτων αλλά και της επίδρασης που έχουν στην καθημερινότητα, η διαχείριση κινδύνων δεν περιορίζεται μόνο σε ζητήματα ασφάλειας πληροφοριών. Επεκτείνεται πλέον σε κινδύνους που αφορούν νομικές, ηθικές και κοινωνικές παραμέτρους. Ζητήματα όπως η προστασία της ιδιωτικότητας, η διαφάνεια των αλγορίθμων και η λογοδοσία των αποφάσεων που λαμβάνονται από συστήματα Τεχνητής Νοημοσύνης αποτελούν πλέον κρίσιμους παράγοντες τους οποίους πρέπει να διαχειριστούν τα υφιστάμενα αλλά και τα καινούργια Risk Management Frameworks. Επέκταση αυτού είναι η ανάγκη επαναπροσδιορισμού της έννοιας του κινδύνου και των μεθόδων αξιολόγησής του. Σε αντίθεση με τα παραδοσιακά πληροφοριακά συστήματα, τα συστήματα AI χαρακτηρίζονται από δυναμική συμπεριφορά, εξάρτηση από δεδομένα εκπαίδευσης και ικανότητα αναπροσαρμογής, γεγονός που καθιστά τους κινδύνους λιγότερο προβλέψιμους και περισσότερο προσαρμοστικούς σε βάθος χρόνου.

Οι κίνδυνοι αυτοί έχουν δημιουργήσει την ανάγκη ανάπτυξης νέων αλλά και εξέλιξης των ήδη υφιστάμενων πλαισίων διαχείρισης κινδύνου όπως το NIST AI Risk Management Frameworks και η risk-based προσέγγιση του Ευρωπαϊκού Κανονισμού για την Τεχνητή Νοημοσύνη (EU AI Act), τα οποία επιχειρούν να συνδυάσουν τεχνικά

και κανονιστικά στοιχεία σε ένα ενιαίο πλαίσιο αξιολόγησης. Η κατανόηση των βασικών αρχών της αξιολόγησης κινδύνων καθίσταται απαραίτητη προκειμένου να αναδειχθούν οι ιδιαιτερότητες της τεχνολογίας και να τεκμηριωθεί η ανάγκη υιοθέτησης προσαρμοσμένων ή συνδυαστικών πλαισίων διαχείρισης κινδύνων, ικανών να ανταποκριθούν στις σύγχρονες τεχνολογικές και κανονιστικές προκλήσεις.



Εικόνα 3: Μεθοδολογία ISO 27005

4.1 Διαδικασία Αξιολόγησης Κινδύνων κατά ISO 27005

Η διαδικασία αξιολόγησης κινδύνων (Risk Assessment) αποτελεί έναν δομημένο και συστηματικό μηχανισμό μέσω του οποίου μπορούμε να αναγνωρίσουμε, να αναλύσουμε και να αξιολογήσουμε τους κινδύνους που ενδέχεται να επηρεάσουν την επίτευξη στόχων. Στο πλαίσιο της Τεχνητής Νοημοσύνης, ο κύριος στόχος είναι η υλοποίηση των δραστηριοτήτων με γνώμονα την ασφάλεια των πληροφοριών και την ιδιωτικότητα των δεδομένων.

Σύμφωνα με το ISO 27005:2022, για να μπορεί να υλοποιηθεί η διεργασία θα πρέπει να αναγνωρίσουμε το πλαίσιο εντός του οποίου θα υλοποιηθεί η αναγνώριση των κινδύνων (Context establishment). Σε αυτό το στάδιο προσδιορίζονται το αντικείμενο της ανάλυσης, τα όρια του συστήματος, τα ενδιαφερόμενα μέρη και οι επιχειρησιακοί, νομικοί και κανονιστικοί στόχοι. Αναγνωρίζοντας το πεδίο εφαρμογής, μπορεί να γίνει πιο κατανοητή η ανοχή στους κινδύνους που επηρεάζουν το σύστημα, τον τρόπο με τον οποίο γίνεται αντιληπτός ο κίνδυνος αλλά και ο τρόπος με τον οποίο θα αξιολογηθεί και που θα αντιμετωπιστεί στην συνέχεια. Στο στάδιο αυτό θα καθοριστούν βασικά κριτήρια αξιολόγησης όπως

- Η κρισιμότητα των συστημάτων που θα εμπλακούν, δηλαδή το πόσο σημαντικά είναι το συστήματα για τις λειτουργίες που εντάσσονται στο πεδίο εφαρμογής
- Οι λειτουργικές ανάγκες και η σημασία τους όσων αφορά στους κεντρικούς πυλώνες της ασφάλειας (εμπιστευτικότητα, διαθεσιμότητα, ακεραιότητα)
- Τις προσδοκίες των ενδιαφερόμενων μερών, όσων αφορά στις αρνητικές επιπτώσεις των κινδύνων.

Αναλογιζόμενοι την κρισιμότητα, προκύπτουν και κριτήρια που αφορούν τις επιπτώσεις των κινδύνων αυτών πάνω στα συστήματα και τους πόρους που έχουν ενταχθεί στο πεδίο εφαρμογής, όπως (όχι περιοριστικά):

- Την διαβάθμιση των πόρων και των συστημάτων αυτών (Πχ. αν εμπλέκονται εμπιστευτικές πληροφορίες όπως είναι τα προσωπικά και τα ευαίσθητα δεδομένα που εν δυνάμει επεξεργάζεται ένα σύστημα)
- Περιστατικά που μπορούν να επηρεάσουν τους πόρους
- Απώλεια της επιχειρησιακής και οικονομικής αξίας

Επιπλέον, προκύπτει ανάγκη αναγνώρισης κριτηρίων αποδοχής των κινδύνων. Πάντα θα εντοπίζονται περιπτώσεις, όπου η επίλυση ή η μείωση ενός κινδύνου δημιουργεί περισσότερα προβλήματα ή δεν είναι βιώσιμη για τον υπεύθυνο διαχείρισης. Τέτοιου είδους κριτήρια μπορεί:

- Να περιλαμβάνουν αναγνώριση ορίων τα οποία επιτρέπουν στον υπεύθυνο υλοποίησης να μην λάβει περαιτέρω ενέργειες για την αντιμετώπισης ενός αναγνωρισμένου κινδύνου.

- Να αποτυπώνονται σε σχέση με το κόστος που απαιτείται για την υλοποίησή μιας διορθωτική ενέργειας
- Να περιλαμβάνουν χρονικά πλαίσια εντός των οποίων ένας υπεύθυνος υλοποίησης δεν θα προχωρήσει σε ενέργειες αντιμετώπισης ώστε να διευθετηθεί το ζήτημα σε μεταγενέστερο χρόνο.

Σε αντίθεση με τον ISO που αποτυπώνει το στάδιο ανάλυσης του επιχειρησιακού περιβάλλοντος ως διακριτό βήμα στην διαδικασία διαχείρισης κινδύνων, ο NIST το αντιμετωπίζει ως μια συνεχόμενη δραστηριότητα που εξελίσσεται καθ' όλη τη διάρκεια του κύκλου ζωής ενός συστήματος. Παρά τις διαφορές αυτές, τα δύο πλαίσια συγκλίνουν στη θεμελιώδη αρχή ότι χωρίς σαφή κατανόηση του οργανωτικού, τεχνικού και κανονιστικού πλαισίου, η αξιολόγηση κινδύνων καθίσταται ελλιπής και αναξιόπιστη.

Αφού αναγνωρισθεί το πλαίσιο και οι υπεύθυνοι υλοποίησης και διαχείρισης ενεργειών, ξεκινάει η διαδικασία του Risk Assessment. Παρότι τα επιμέρους στάδια μπορεί να διαφοροποιούνται ελαφρώς ανάλογα με το εφαρμοζόμενο πλαίσιο ή πρότυπο, η πλειονότητα των διεθνών frameworks συγκλίνει σε τρία βασικά στάδια:

1. τον αναγνώριση κινδύνων (risk identification),
2. την ανάλυση κινδύνων (risk analysis) και
3. την αξιολόγηση κινδύνων (risk evaluation).

4.1.1 Αναγνώριση Κινδύνων (Risk Identification)

Κατά την αναγνώριση των κινδύνων, σκοπός είναι να εντοπίσουμε και να καταγράψουμε τους κινδύνους που επηρεάζουν τον Οργανισμό, το σύστημα και γενικότερα το αντικείμενο της ανάλυσης. Για να μπορέσει να επιτευχθεί αυτό, είναι σημαντικό να έχει καταγραφεί επαρκώς η βάση πάνω στην οποία θα στηριχθεί η συνολική διαδικασία αξιολόγησης των κινδύνων.

Η διαδικασία ξεκινά με την αναγνώριση και κατηγοριοποίηση των περιουσιακών στοιχείων (assets) που σχετίζονται με το υπό εξέταση σύστημα. Στα παραδοσιακά πληροφοριακά συστήματα, αυτά περιλαμβάνουν δεδομένα, λογισμικό, υλικό εξοπλισμό και ανθρώπινους πόρους μεταξύ άλλων. Στην περίπτωση των συστημάτων Τεχνητής Νοημοσύνης, η έννοια των πόρων επεκτείνεται ώστε να περιλαμβάνει τα σύνολα δεδομένων εκπαίδευσης και δοκιμών, τα μοντέλα μηχανικής μάθησης, τις

παραμέτρους τους, τις διεπαφές εισόδου και εξόδου, καθώς και τις παραγόμενες αποφάσεις ή προβλέψεις συνδυαστικά με τους πόρους που εμπλέκονται στα «παραδοσιακά» πληροφοριακά συστήματα. Η σαφής οριοθέτηση των πόρων αποτελεί απαραίτητη προϋπόθεση για την κατανόηση του τι ακριβώς προστατεύεται και ποια στοιχεία είναι κρίσιμα για τη λειτουργία και την αξιοπιστία του συστήματος. Ως πληροφοριακοί πόροι νοούνται όλα τα στοιχεία – τεχνικά, λογισμικά, οργανωτικά και ανθρώπινα – που συμβάλλουν στη λειτουργία και στην ασφάλεια του συστήματος. Δεν περιορίζονται στις υποδομές ή τις εφαρμογές, αλλά επεκτείνονται σε όλα τα δεδομένα, τις διαδικασίες και τους ανθρώπινους πόρους που αποτελούν μέρος του πεδίου εφαρμογής.

Η ταυτοποίηση αυτών των αγαθών αποτελεί τη βάση για την εκτίμηση απειλών, ευπαθειών και επιπτώσεων, εξασφαλίζοντας ότι η ανάλυση κινδύνων καλύπτει το σύνολο του κύκλου ζωής – από το επίπεδο υποδομής έως τα δεδομένα και τους ανθρώπους. Στόχος είναι η ποιοτική αξιολόγηση της σημασίας κάθε αγαθού για το εκάστοτε πληροφοριακό σύστημα και τον φορέα λειτουργίας του, προκειμένου να εκτιμηθούν οι πιθανές επιπτώσεις από τυχόν παραβίαση της εμπιστευτικότητας, της ακεραιότητας ή της διαθεσιμότητας.

Αφού έχει καταγραφεί η υποδομή πάνω στην οποία διεξάγεται η αξιολόγηση, είναι απαραίτητη η αναγνώριση των πιθανών απειλών (threats) και ευπαθειών (Vulnerabilities) που μπορεί να επηρεάσουν το σύστημα και να δημιουργήσουν τους κινδύνους που θα πρέπει να εξετάσουμε.

1. Ευπάθειες είναι οι αδυναμίες ή τα ελλιπή μέτρα ασφάλειας που μπορεί να υπάρχει σε ένα σύστημα. Οι ευπάθειες μπορεί να είναι τεχνικής φύσεως, όπως ανεπαρκή μέτρα ελέγχου πρόσβασης ή ελλιπής προστασία των μοντέλων, αλλά και οργανωτικές, όπως η απουσία πολιτικών διακυβέρνησης AI ή διαδικασιών ανθρώπινης εποπτείας. Οι ευπάθειες αυτούσιες δεν δημιουργούν κινδύνους.
2. Απειλές είναι οι εξωτερικοί ή εσωτερικοί παράγοντες, που θα εκμεταλλευτούν τις ευπάθειες ενός συστήματος ώστε να προκαλέσουν ζημιά. Οι απειλές μπορεί να προέρχονται από κακόβουλους χρήστες, τεχνικές αστοχίες, ανθρώπινα λάθη ή εξωτερικούς παράγοντες, όπως αλλαγές στο κανονιστικό περιβάλλον. Στα συστήματα Τεχνητής Νοημοσύνης, οι απειλές περιλαμβάνουν τόσο κλασικές κυβερνοαπειλές όσο και εξειδικευμένες επιθέσεις, όπως adversarial attacks, model extraction και data poisoning, αλλά και μη σκόπιμες απειλές που

σχετίζονται με τη μεροληψία των δεδομένων ή την ελλιπή αντιπροσώπευση συγκεκριμένων ομάδων

3. Κίνδυνος, είναι η πιθανότητα μια απειλή να εκμεταλλευτεί μια υπάρχουσα ευπάθεια ώστε να προκαλέσει ζημιά ή / και να επηρεάσει έναν πόρο.

Ο συνδυασμός των απειλών και των ευπαθειών θα δημιουργήσει σενάρια κινδύνων (Risk Scenarios) τα οποία θα πρέπει να αναλυθούν, αξιολογηθούν και αντιμετωπιστούν στις επόμενες φάσεις της διαδικασίας.

4.1.2 Ανάλυση Κινδύνων (Risk Analysis)

Η ανάλυση των κινδύνων αποτελεί το στάδιο κατά το οποίο οι κίνδυνοι που έχουν ήδη εντοπιστεί στην προηγούμενη φάση μετατρέπονται από γενικές περιγραφές σε αναλυτικά, αξιολογήσιμα σενάρια. Σκοπός δεν είναι ακόμη η λήψη αποφάσεων, αλλά η κατανόηση της φύσης του κινδύνου, των αιτιών του και των πιθανών συνεπειών του, ώστε να καταστεί δυνατή η τεκμηριωμένη αξιολόγηση στη συνέχεια. Η ανάλυση κινδύνων επικεντρώνεται κυρίως στη μελέτη της πιθανότητας πραγματοποίησης ενός σεναρίου, δηλαδή της πιθανότητας μια απειλή να εκμεταλλευτεί μια ευπάθεια αλλά και της σοβαρότητας των επιπτώσεών που θα έχει η εκμετάλλευση της ευπάθειας, λαμβάνοντας υπόψη τα υφιστάμενα μέτρα ελέγχου.

Η σοβαρότητα του κινδύνου αποτελείται από δυο συνιστώσες όπως αναφέρθηκε παραπάνω. Επομένως, το τελικό Ρίσκο (Risk) υπολογίζεται από ως εξής:

- **Επικινδυνότητα (Risk) = Πιθανότητα × Επίπτωση**

- Η **πιθανότητα** αποτυπώνει το πόσο συχνά ή πόσο εύκολα μπορεί να συμβεί ένα ανεπιθύμητο γεγονός (το πόσο συχνά θα εμφανιστεί μια απειλή), λαμβάνοντας υπόψη το επίπεδο έκθεσης, την ύπαρξη ευπαθειών και την αποτελεσματικότητα των υφιστάμενων ελέγχων.

- Η **επίπτωση** αποτυπώνει τη σοβαρότητα των ευπαθειών εφόσον τις εκμεταλλευτεί μια απειλή, οι οποίες μπορεί να είναι τεχνικές, οικονομικές, νομικές ή κοινωνικές, όπως για παράδειγμα η παραβίαση προσωπικών δεδομένων ή η απώλεια εμπιστοσύνης των χρηστών.

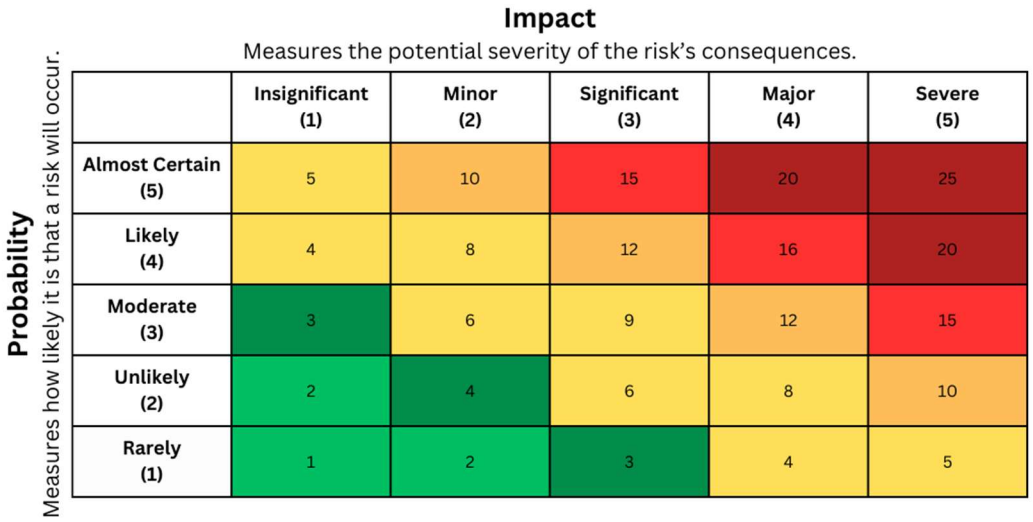
Η πιθανότητα εκδήλωσης ενός κινδύνου μπορεί να εκτιμηθεί ποιοτικά ή ποσοτικά, ανάλογα με τη διαθεσιμότητα δεδομένων και το επίπεδο ωριμότητας του

πλαισίου εντός του οποίου υλοποιείται η διαδικασία. Σε πολλές προσεγγίσεις, χρησιμοποιούνται ποιοτικές κλίμακες (χαμηλή, μεσαία, υψηλή) ή αριθμητικές τιμές (π.χ. 1 έως 5), οι οποίες βασίζονται σε παράγοντες όπως το ιστορικό παρόμοιων περιστατικών, η έκθεση του συστήματος σε εξωτερικούς χρήστες και η πολυπλοκότητα της τεχνολογίας. Αντίστοιχα, οι επιπτώσεις αναλύονται ως προς διαφορετικές διαστάσεις, όπως η εμπιστευτικότητα, η ακεραιότητα και η διαθεσιμότητα, αλλά και ως προς νομικές, οικονομικές και κοινωνικές συνέπειες, ώστε το σενάριο κινδύνου να ανταποκρίνεται όσο το δυνατόν περισσότερο στην πραγματικότητα. Ο συνδυασμός τους οδηγεί σε ένα συνολικό επίπεδο κινδύνου, το οποίο απεικονίζεται συχνά μέσω ενός πίνακα ή μήτρας κινδύνου (risk matrix). Για παράδειγμα, ένας κίνδυνος με υψηλή πιθανότητα και υψηλή επίπτωση κατατάσσεται ως κρίσιμος, ενώ ένας κίνδυνος με χαμηλή πιθανότητα και χαμηλή επίπτωση θεωρείται αποδεκτός ή δευτερεύων, πάλι μέσω κλίμακας που έχει συμφωνηθεί κατά την υλοποίηση. Σε πιο ώριμα περιβάλλοντα, ο υπολογισμός του ρίσκου μπορεί να λάβει ποσοτική μορφή, όπου χρησιμοποιούνται πραγματικά δεδομένα, στατιστικές πιθανότητες και εκτιμήσεις κόστους. Σε αυτή την περίπτωση, η επίπτωση εκφράζεται συχνά σε οικονομικούς όρους ή σε μετρήσιμες απώλειες, επιτρέποντας τον υπολογισμό αναμενόμενης ζημίας. Ωστόσο, στα συστήματα τεχνητής νοημοσύνης και ιδίως στους κινδύνους ασφάλειας και ιδιωτικότητας, η πλήρης ποσοτικοποίηση είναι δύσκολη, καθώς πολλές επιπτώσεις, όπως η κοινωνική βλάβη ή η απώλεια ανωνυμίας, δεν αποτιμώνται εύκολα αριθμητικά.

Έχοντας αποδώσει οπτικά και ποσοτικά τους βαθμούς των κινδύνων από τα σενάρια που έχουν προκύψει, ο υπεύθυνος υλοποίησης μπορεί πλέον να προχωρήσει στο στάδιο της αξιολόγησης των κινδύνων.

4.1.3 Αξιολόγηση Κινδύνων (Risk Assessment)

Κατά την ανάλυση του περιβάλλοντος και του πεδίου εφαρμογής ορίζεται ένα κατώφλι (Threshold), βάσει του οποίου συγκρίνουμε τους βαθμούς των κινδύνων που προέκυψαν κατά τη φάση της ανάλυσης. Πλέον δεν υλοποιείται περαιτέρω ανάλυση των σεναρίων, αλλά λαμβάνονται αποφάσεις για τους κινδύνους που έχουν εντοπιστεί. Έχοντας κατατάξει τους κινδύνους βάσει της σοβαρότητας και έχοντας ορίσει το όριο πάνω από το οποίο οι κίνδυνοι θεωρούνται κρίσιμοι βάσει των πινάκων που έχουν καθοριστεί, προκύπτουν αποδεκτοί και μη αποδεκτοί κίνδυνοι όπου ο υπεύθυνος υλοποίησης καλείται να υλοποιήσει ενέργειες για την διαχείρισή τους.



Εικόνα 4: 5x5 Πίνακας Αξιολόγησης κινδύνου

Ανάλογα με την ιεράρχηση και τα κριτήρια που έχουν οριστεί στην εκάστοτε περίπτωση, τηρείται και ένα σχετικό πλάνο. Αν λάβουμε υπόψη τον πίνακα που έχει παρατεθεί παραπάνω, η αξιολόγηση των κινδύνων θα λάμβανε την εξής μορφή:

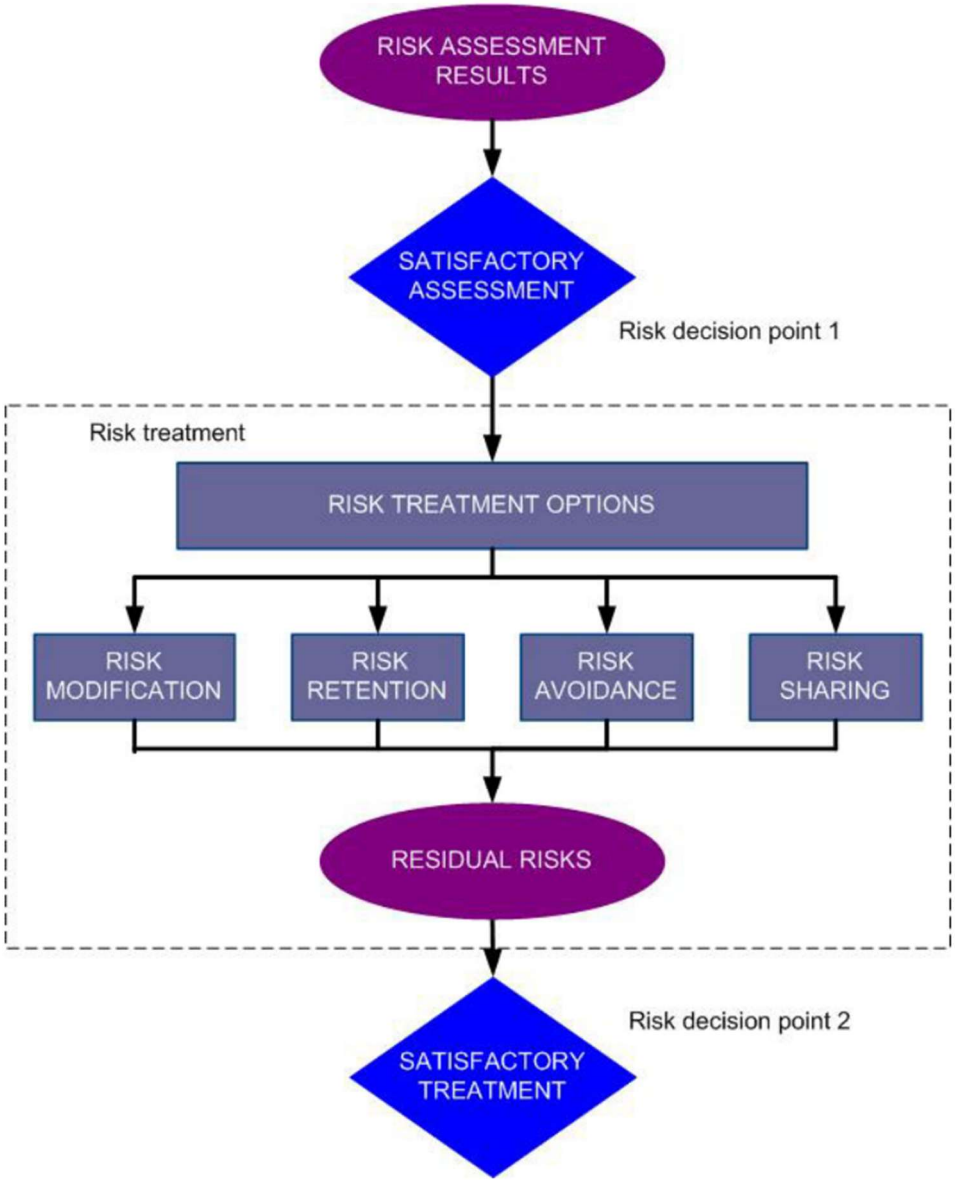
1. Κίνδυνοι με βαθμό >15: Δεν θεωρούνται αποδεκτοί και θα πρέπει να ληφθούν άμεσα ενέργειες
2. Κίνδυνοι 10 – 12: θεωρούνται αποδεικτοί όπου όμως θα πρέπει να παρακολουθούνται και εν δυνάμει να ληφθούν ενέργειες διαχείρισης
3. Κίνδυνοι 4 – 9: αποδεκτοί και υπό παρακολούθηση
4. Κίνδυνοι <4: αποδεκτοί

Για την ιεράρχηση των κινδύνων λαμβάνονται υπόψη παράγοντες όπως η πιθανότητα εμφάνισης πολλαπλών επιπτώσεων, η αβεβαιότητα των εκτιμήσεων και η δυσκολία ανίχνευσης ή αποκατάστασης ενώ ιδιαίτερη βαρύτητα δίνεται και στη συμμόρφωση με το κανονιστικό πλαίσιο. Κίνδυνοι που συνδέονται με πιθανή παραβίαση νομικών και κανονιστικών απαιτήσεων, όπως ο GDPR ή ο υπό διαμόρφωση Κανονισμός ΤΝ της Ευρωπαϊκής Ένωσης, συνήθως χαρακτηρίζονται ως μη αποδεκτοί ανεξαρτήτως της ποσοτικής τους εκτίμησης. Αυτό αντικατοπτρίζει τη μετατόπιση της αξιολόγησης κινδύνων από μια καθαρά τεχνική προσέγγιση σε μια πολυδιάστατη διαδικασία που ενσωματώνει νομικές, ηθικές και κοινωνικές διαστάσεις.

Η φάση της αξιολόγησης ολοκληρώνεται με την τεκμηρίωση των αποτελεσμάτων και τη διαμόρφωση ενός καταλόγου κινδύνων όπου κατατάσσονται σε φθίνουσα σειρά από τους πιο κρίσιμους προς τους αποδεκτούς. Ο κατάλογος αυτός αξιοποιείται και στην Φάση αντιμετώπισης κινδύνων (Risk Treatment).

4.1.4 Αντιμετώπιση Κινδύνων (Risk Treatment)

Πλέον έχουν αξιολογηθεί οι κίνδυνοι και έχουν προκύψει αυτοί που θεωρούνται μη αποδεκτοί και κρίσιμοι και πρέπει να ληφθούν αποφάσεις για την αντιμετώπισή και την διαχείρισή τους. Η αντιμετώπιση κινδύνων δεν στοχεύει στην πλήρη εξάλειψη όλων των κινδύνων, αλλά στη μείωσή τους σε ένα αποδεκτό επίπεδο, λαμβάνοντας υπόψη το επιχειρησιακό πλαίσιο, τους διαθέσιμους πόρους και τις κανονιστικές απαιτήσεις.



Εικόνα 5: Αντιμετώπιση Κινδύνων

Για να υλοποιηθεί η αντιμετώπιση, πρέπει να γίνει η επιλογή της κατάλληλης μεθόδου. Βάσει προτύπου οι διαθέσιμές μέθοδοι αντιμετώπισης είναι οι εξής:

1. **Μείωση κινδύνου (Risk Modification):** αποτελεί τη συχνότερα εφαρμοζόμενη μέθοδο και αφορά την εφαρμογή κατάλληλων μέτρων ελέγχου που περιορίζουν είτε την πιθανότητα εκδήλωσης ενός κινδύνου είτε την επίπτωσή που θα έχει σε περίπτωση που πραγματοποιηθεί. Τα μέτρα ελέγχου μπορεί να είναι τεχνικά, οργανωτικά, διαδικαστικά και συχνά ευθυγραμμίζονται με τα μέτρα που αποτυπώνονται στο πρότυπο Εφαρμογής Συστημάτων Διαχείρισης Ασφάλειας Πληροφοριών του ISO/IEC 27001 και του Παραρτήματος Α (Annex A). Τεχνικά μέτρα μπορεί να περιλαμβάνουν, μεταξύ άλλων, την ενίσχυση της ασφάλειας των δεδομένων εκπαίδευσης, τη χρήση τεχνικών robustness έναντι adversarial επιθέσεων, τον περιορισμό προσβάσεων στο μοντέλο και τη συνεχή παρακολούθηση της συμπεριφοράς του συστήματος. Οργανωτικά μέτρα αφορούν τη σαφή κατανομή ρόλων και ευθυνών, την εκπαίδευση των εμπλεκόμενων, καθώς και την ενσωμάτωση της ασφάλειας και της ιδιωτικότητας στον κύκλο ζωής του συστήματος (security και privacy by design). Παράλληλα, η συμμόρφωση με νομικές απαιτήσεις, όπως ο GDPR και το AI Act της Ευρωπαϊκής Ένωσης, λειτουργεί όχι μόνο ως υποχρέωση αλλά και ως μηχανισμός μετριασμού κινδύνων. Με την εφαρμογή των κατάλληλων controls, ο βαθμός κινδύνου θα μειωθεί έως ότου φτάσει σε έναν επίπεδο που θεωρείται αποδεκτό και πλέον δεν θα απαιτούνται περαιτέρω ενέργειες από τον υπεύθυνο υλοποίησης.
2. **Αποδοχή Κινδύνου (Risk Retention / Acceptance):** επιλέγεται όταν το επίπεδο του υπολειπόμενου κινδύνου θεωρείται αποδεκτό σύμφωνα με τα προκαθορισμένα κριτήρια που έχουν οριστεί στο αρχικό στάδιο υλοποίησης της διαδικασίας. Τα κριτήρια αποδοχής μπορεί να προκύπτουν από
    1. Το επίπεδο του κινδύνου
    2. Τις νομικές και κανονιστικές απαιτήσεις
    3. Τη σχέση κόστους – οφέλους
    4. Την υφιστάμενη κατάσταση
    5. Την Εφαρμογή αλλαγών που θα επηρεάσουν τον κίνδυνο που έχει προκύψει.

Σύμφωνα με το ISO/IEC 27005, η αποδοχή αποτελεί συνειδητή και τεκμηριωμένη απόφαση, η οποία πρέπει να εγκρίνεται από την αρμόδια διοίκηση. Η αποδοχή κινδύνου μπορεί να αφορά κινδύνους χαμηλής επίπτωσης ή χαμηλής πιθανότητας, οι οποίοι δεν δικαιολογούν την εφαρμογή πρόσθετων μέτρων ή που στην παρούσα συνθήκη δεν υφίστανται οι απαιτούμενοι πόροι. Ακόμη και σε αυτή την περίπτωση, επισημαίνεται η ανάγκη για συνεχή παρακολούθηση, καθώς πάντα υπάρχει η δυνατότητα μεταβολής του επιπέδου κινδύνου με την πάροδο του χρόνου ή με την αλλαγή του περιβάλλοντος που επιδρά στο πεδίο εφαρμογής.

3. **Αποφυγή Κινδύνου (Risk Avoidance):** αποτελεί την διαδικασία λήψης αποφάσεων που εξαλείφουν πλήρως τη δραστηριότητα ή την αιτία που δημιουργεί τον κίνδυνο εξαρχής. Συνήθως, η μέθοδος αυτή επιλέγεται όταν το επίπεδο κινδύνου κρίνεται μη αποδεκτό και δεν υπάρχουν ρεαλιστικά ή οικονομικά βιώσιμα μέτρα για τη μείωσή του. Για παράδειγμα, η συγκεκριμένη μέθοδος θα μπορούσε να εφαρμοστεί με την αποφυγή χρήσης ενός μοντέλου Τεχνητής Νοημοσύνης υψηλού ρίσκου που δεν καλύπτει τις νομικές απαιτήσεις που έχουν τεθεί.
4. **Μεταφορά κινδύνου (Risk Sharing):** αφορά τον διαμοιρασμό της επίπτωσης ενός κινδύνου με τρίτους, χωρίς ωστόσο να αίρεται πλήρως η ευθύνη του Αρχικού υπεύθυνου Διαχείρισης. Η μεταφορά κινδύνου υλοποιείται συνήθως μέσω συμβάσεων, SLAs (Service Level Agreements) ή ασφαλιστικών μηχανισμών με ασφαλιστικές Εταιρείες. Αυτό μπορεί να περιλαμβάνει τη χρήση υπηρεσιών AI τρίτων παρόχων, με σαφείς ρήτρες για την ασφάλεια, την προστασία δεδομένων και τη διαχείριση περιστατικών. Ωστόσο, είναι σημαντικό να τονιστεί πως η μεταφορά κινδύνου δεν αντικαθιστά την ανάγκη για αξιολόγηση και παρακολούθηση, ιδίως όταν ο οργανισμός παραμένει υπεύθυνος έναντι χρηστών ή ρυθμιστικών αρχών. Εν αντιθέσει, στις περισσότερες περιπτώσεις που πρέπει να υλοποιηθεί συμφωνία πάνω στην οποία βασίζεται η μεταφορά ή ο διαμοιρασμός του κινδύνου, πρέπει να εξακριβωθεί μέσω ελέγχων, πως έχουν εφαρμοστεί τεχνικά και οργανωτικά μέτρα για την διαχείριση του κινδύνου.

Με την έγκριση και εφαρμογή των μεθόδων και των ενεργειών που έχουν επιλεγεί, ο υπεύθυνος υλοποίησης θα χρειαστεί να υπολογίσει τον εναπομείναντα

ή υπολειπόμενο κίνδυνο (Residual Risk). Θα πρέπει να γίνει επαναξιολόγηση των επιπτώσεων αλλά και της πιθανότητας εμφάνισης ώστε να διαπιστωθεί εάν το επίπεδο κινδύνου πλέον βρίσκεται σε αποδεκτά επίπεδα και δεν απαιτούνται περαιτέρω ενέργειες αντιμετώπισης. Σε περίπτωση που διαπιστωθεί ότι κάποιος κίνδυνος εξακολουθεί να είναι εκτός των ορισμένων ορίων, η διαδικασία επαναλαμβάνεται.

Το στάδιο της αντιμετώπισης κινδύνων ολοκληρώνεται με την τεκμηρίωση των αποφάσεων και των μέτρων που επιλέχθηκαν, συνήθως μέσω ενός σχεδίου αντιμετώπισης κινδύνων (risk treatment plan). Το σχέδιο αυτό συνδέει κάθε κίνδυνο με συγκεκριμένες ενέργειες αντιμετώπισης, υπευθύνους, χρονοδιαγράμματα και αξιολογήσεις των υλοποιημένων ενεργειών και αποτελεί την κύρια τεκμηρίωση και απεικόνιση της αντιμετώπισης κινδύνων.

4.2 ISO 42001 & ISO 23894

Με την καθιέρωση της Τεχνητής Νοημοσύνης, ήταν αυστηρή απαίτηση να προτυποποιηθεί η λειτουργία της και να ενταχθεί σε ένα πλαίσιο διαχείρισης, ώστε να είναι εύκολο στον επιχειρηματικό κόσμο όχι μόνο να την εντάξει στην λειτουργία του αλλά να το κάνει με έναν τρόπο που δεν θα δημιουργούσε ζητήματα και επιπλέον κινδύνους από τους ήδη υπάρχοντες. Απάντηση στο πρόβλημα αυτό αποτέλεσε η δημιουργία ενός προτύπου που περιγράφει τον τρόπο με τον οποίο ένας Οργανισμός μπορεί να αξιοποιήσει ένα Σύστημα Διαχείρισης Τεχνητής Νοημοσύνης (Information Technology - Artificial Intelligence – Management System -AIMS).

Όπως τα περισσότερα συστήματα Διαχείρισης του ISO, έτσι και το AIMS βασίζεται στον κύκλο PDCA (Plan – Do – Check – Act) και βασίζεται στην αναγνώριση και διαχείριση κινδύνων. Σκοπός του Συστήματος δεν είναι να αναγνωρίσει μόνο τους τεχνολογικούς κινδύνους και τους κινδύνους ιδιωτικότητας, αλλά να εντοπίσει και τις κοινωνικές επιπτώσεις που εν δυνάμει θα προκαλέσει ένα σύστημα Τεχνητής Νοημοσύνης. Ακολουθεί την ίδια δομή με άλλα ISO Συστήματα ξεκινώντας με τον καθορισμό του πλαισίου και του πεδίου εφαρμογής, την ενημέρωση των ενδιαφερόμενων μερών, την συνεχή παρακολούθηση μέσω στόχων και επιθεωρήσεων, την διαχείριση των κινδύνων και των επιπτώσεων και την συνεχή βελτίωση. Αντίστοιχα με το ISO 27001, έτσι και

το ISO 42001 παρέχει το παράρτημα (Annex A), εντός του οποίου παρατίθενται μέτρα ελέγχου με σκοπό να ενσωματωθούν στην κουλτούρα και την καθημερινότητα ενός Οργανισμού για την καλύτερη διαχείριση του AI.

Όσον αφορά στην διαχείριση των κινδύνων, η διαδικασία δεν διαφοροποιείται ιδιαίτερα από την διαδικασία που αναλύεται στο ISO 27005. Η ίδια λογική μπορεί να εφαρμοστεί και στα συστήματα τεχνητής Νοημοσύνης, καθορίζοντας το πεδίο εφαρμογής, αναγνωρίζοντας ευπάθειες και απειλές, καταγράφοντας τις πιθανές επιπτώσεις και καταλήγοντας σε ένα σχέδιο αντιμετώπισης κινδύνων για τα ρίσκα που υπερβαίνουν το όριο ανοχής κινδύνων που έχει τεθεί.

Εκεί που το πρότυπο διαφοροποιείται από ένα πρότυπο διαχείρισης ασφάλειας και ιδιωτικότητας, είναι η απαίτηση υλοποίησης Ανάλυσης Επιπτώσεων (Impact Assessment) του συστήματος AI. Ο οργανισμός καλείται να αξιολογήσει όχι μόνο το αν ένα σύστημα λειτουργεί σωστά από τεχνική άποψη, αλλά και το πώς οι αποφάσεις ή τα αποτελέσματά που προκύπτουν από αυτό επηρεάζουν τα ενδιαφερόμενα μέρη (stakeholders), ιδίως όταν αυτά περιλαμβάνουν φυσικά πρόσωπα ή/και την ίδια την κοινωνία. Για κάθε ενδιαφερόμενο μέρος εξετάζεται ο τρόπος με τον οποίο οι αποφάσεις του συστήματος μπορεί να επηρεάσουν δικαιώματα, ελευθερίες, ασφάλεια ή άλλα συμφέροντα. Στη συνέχεια, αναλύονται οι τύποι επιπτώσεων, οι οποίοι στο ISO 42001 εκτείνονται πέρα από τους πυλώνες της ασφάλειας (εμπιστευτικότητα–ακεραιότητα–διαθεσιμότητα) και εξετάζονται επιπτώσεις όπως:

- η μεροληψία και οι διακρίσεις,
- η έλλειψη διαφάνειας και επεξηγηματικότητας,
- η απώλεια ανθρώπινης εποπτείας,
- η παραβίαση ιδιωτικότητας,
- καθώς και οι κοινωνικές ή περιβαλλοντικές συνέπειες.

Η διαδικασία αυτή, δεν αποτελεί ένα απλό βήμα στην διαδικασία υλοποίησης ενός συστήματος Τεχνητής Νοημοσύνης. Αντιθέτως εφαρμόζεται σε όλα τα στάδια της υλοποίησης και επικαιροποιείται βάσει των αλλαγών του συστήματος αλλά και των μεταβολών του πλαισίου εντός του οποίου υλοποιείται. (ISO, 2023)

Σημαντική συμβολή στην υλοποίηση της διαδικασίας διαχείρισης κινδύνων στο πλαίσιο ενός συστήματος AI, έχει το πρότυπο ISO 23894 Information technology — Artificial intelligence — Guidance on risk management (Οδηγίες για την

Διαχείριση Κινδύνων). Το πρότυπο αυτό λειτουργεί συνεργατικά με το ISO 42001 με βασικό στόχο να παρέχει οδηγίες και κατευθυντήριες γραμμές για τη διαχείριση κινδύνων, εστιάζοντας σε αυτούς που προκύπτουν από τη φύση των αλγορίθμων, των δεδομένων και της αυτοματοποιημένης λήψης αποφάσεων. Σε αντίθεση με το ISO 42001, το οποίο ορίζει τις απαιτήσεις ενός Συστήματος Διαχείρισης, το ISO 23894 λειτουργεί περισσότερο ως πλαίσιο υλοποίησης. Οι κύριοι τομείς στους οποίους εμβαθύνει το συγκεκριμένο πρότυπο είναι:

1. **Μεροληψία στα μοντέλα τεχνητής νοημοσύνης**: Εφαρμογή τεχνικών για τον εντοπισμό και τον μετριασμό της μεροληψίας στα μοντέλα τεχνητής νοημοσύνης και στις διαδικασίες λήψης αποφάσεων.
2. **Προστασία δεδομένων στα συστήματα τεχνητής νοημοσύνης**: Διασφάλιση της συμμόρφωσης με τους κανονισμούς προστασίας δεδομένων και εφαρμογή τεχνικών τεχνητής νοημοσύνης που προστατεύουν την ιδιωτικότητα.
3. **Ηθικές πτυχές της τεχνητής νοημοσύνης**: Ανάπτυξη και τήρηση κατευθυντήριων γραμμών για την ηθική ανάπτυξη και εφαρμογή της τεχνητής νοημοσύνης.
4. **Ασφάλεια των συστημάτων τεχνητής νοημοσύνης:** Εφαρμογή ισχυρών μέτρων ασφαλείας για την προστασία των συστημάτων από επιθέσεις και μη εξουσιοδοτημένη πρόσβαση.
5. **Διαφάνεια στη λήψη αποφάσεων με τεχνητή νοημοσύνη:** Προσπάθεια για την ανάπτυξη μοντέλων οι αποφάσεις των οποίων μπορούν να εξηγηθούν.

4.3 NIST Risk Management Framework (RMF) & Artificial Intelligence (AIRMF)

Το NIST Risk Management Framework (RMF) αναπτύχθηκε για την ενσωμάτωση της ασφάλειας και της προστασίας της ιδιωτικότητας κατά τον κύκλο ζωής των πληροφοριακών συστημάτων. Κεντρική φιλοσοφία του RMF είναι ότι η διαχείριση κινδύνων δεν αποτελεί μεμονωμένη δραστηριότητα, αλλά μια συνεχής και επαναλαμβανόμενη διαδικασία, άμεσα συνδεδεμένη με τους επιχειρησιακούς στόχους και την στρατηγική του οργανισμού. (NIST, 2025) Υιοθετεί μια πιο λειτουργική προσέγγιση, ενσωματώνοντας τη διαχείριση κινδύνων σε όλα τα στάδια ανάπτυξης και λειτουργίας ενός συστήματος. Σε αντίθεση με το ISO/IEC

27005, το RMF δεν περιορίζεται στην αξιολόγηση κινδύνων ως αυτόνομη φάση, αλλά τη συνδέει άμεσα με την επιλογή, υλοποίηση και συνεχή παρακολούθηση των μέτρων ασφάλειας. Η έμφαση στη συνεχή αξιολόγηση και προσαρμογή καθιστά το NIST RMF ιδιαίτερα κατάλληλο για περιβάλλοντα υψηλής μεταβλητότητας, όπως τα συστήματα Τεχνητής Νοημοσύνης.

Στον πυρήνα του, το NIST RMF λειτουργεί ως ένας κύκλος επτά διακριτών αλλά αλληλεξαρτώμενων φάσεων, οι οποίες δεν εκτελούνται γραμμικά μία μόνο φορά, αλλά επαναλαμβάνονται καθ' όλη τη διάρκεια της ζωής του συστήματος, υποστηρίζοντας τη συνεχή βελτίωση και την προσαρμογή σε νέες απειλές. Παράλληλα, το RMF εισάγει τη συνεχή παρακολούθηση (continuous monitoring) ως βασικό μηχανισμό προσαρμογής, αναγνωρίζοντας ότι ο κίνδυνος ενός συστήματος μεταβάλλεται δυναμικά με την πάροδο του χρόνου και δεν μένει ποτέ σταθερός. Οι φάσεις του RMF διακρίνονται στις εξής:

1. **Prepare (Προετοιμασία)**: στην παρούσα φάση, ο Οργανισμός καθορίζει το επιχειρησιακό και κανονιστικό πλαίσιο και γενικότερα το πεδίο εφαρμογής μαζί με Ρόλους, ευθύνες και το επίπεδο του ανεκτού κινδύνου.
2. **Categorize (Κατηγοριοποίηση):** το σύστημα κατηγοριοποιείται βάσει του δυνητικού αντικτύπου που θα είχε μια παραβίαση εμπιστευτικότητας, ακεραιότητας ή διαθεσιμότητας, συνήθως σε low, Moderate και High. Η κατηγοριοποίηση επηρεάζει άμεσα την αυστηρότητα των μέτρων ασφαλείας που θα εφαρμοστούν. Σαν αποτέλεσμα της παρούσας φάσης, θα προκύψουν τα χαρακτηριστικά του συστήματος και η κατηγοριοποίησης που έχει προκύψει.
3. **Select (Επιλογή):** Βάση της κατηγοριοποίησης του συστήματος όπως προέκυψε από την δεύτερη φάση, γίνεται η επιλογή των κατάλληλων μέτρων ελέγχου όπως παρουσιάζονται στο πρότυπο NIST 800-53 λαμβάνοντας υπόψη το επίπεδο κινδύνου, το επιχειρησιακό περιβάλλον και τυχόν κανονιστικές απαιτήσεις. Τα μέτρα που έχουν επιλεγεί, προσαρμόζονται στην παρούσα κατάσταση και στις απαιτήσεις που έχει το εκάστοτε σύστημα.
4. **Implement (Εφαρμογή):** στην φάση implement εφαρμόζονται τα τεχνικά (περιορισμοί πρόσβασης, Hardening υποδομής κλπ.) και

οργανωτικά μέτρα (Εφαρμογή πολιτικών και διαδικασιών κλπ.) που επέλεξε ο Οργανισμός στην προηγούμενη φάση.

5. **Assess (Αξιολόγηση):** Με την εφαρμογή του μέτρου ασφάλειας ή ιδιωτικότητας αξιολογείται η αποτελεσματικότητα του ως προς την μείωση του κινδύνου και την κάλυψη των απαιτήσεων του Οργανισμού. Σε περίπτωση που κάποιο μέτρο δεν καλύπτει όπως είχε προβλεφθεί, διεξάγεται εκ νέου έρευνα για το τι περαιτέρω μέτρα πρέπει να ληφθούν.
6. **Authorize (Έγκριση):** ένα ανώτατο αρμόδιο στέλεχος (Authorizing Official) όπως έχει αποφασιστεί από την Διοίκηση και κατά την φάση Prepare, αξιολογεί εάν εγκρίνεται η συνολική διαχείριση του συστήματος, των μέτρων που έχουν εφαρμοστεί και οι υπολειπόμενοι κίνδυνοι που προκύπτουν σχετικά με την ασφάλεια και την ιδιωτικότητα.
7. **Monitor (Παρακολούθηση):** διασφαλίζει τη συνεχή παρακολούθηση του συστήματος. Νέες απειλές, αλλαγές στο περιβάλλον ή τεχνικές τροποποιήσεις ανατροφοδοτούν τον κύκλο, οδηγώντας σε αναθεώρηση των προηγούμενων φάσεων. (NIST, 2025)

Από την στιγμή που ήρθε στην επιφάνεια η ευρεία επίδραση της τεχνητής Νοημοσύνης, προέκυψε, όπως έχει αναφερθεί, η ανάγκη να επεκταθούν οι υπάρχουσες μεθοδολογίες και να προσαρμοστούν στις εκτεταμένες απαιτήσεις του AI. Όπως ο ISO με τα πρότυπα 42001 & 23894, έτσι και ο NIST δημοσίευσε το NIST AI Risk Management Framework (AI RMF) το οποίο αναπτύχθηκε ως απάντηση στην ανάγκη για μια εξειδικευμένη προσέγγιση διαχείρισης κινδύνων που να ανταποκρίνεται στις ιδιαιτερότητες της νέας αυτής τεχνολογίας. Σύμφωνα με το NIST, η συγκεκριμένη μεθοδολογία «αποτελεί ένα ζωντανό έντυπο το οποίο προσαρμόζεται διαρκώς ανάλογα με τι νέες απαιτήσεις της Τεχνητής Νοημοσύνης».

Σε αντίθεση με τα παραδοσιακά frameworks, το AI RMF αντιμετωπίζει τον κίνδυνο ως ένα πολυδιάστατο φαινόμενο, το οποίο δεν περιορίζεται σε τεχνικές απειλές, αλλά περιλαμβάνει ζητήματα αξιοπιστίας, δικαιοσύνης, διαφάνειας, λογοδοσίας και προστασίας θεμελιωδών δικαιωμάτων. Η προσέγγιση αυτή αντικατοπτρίζει το γεγονός ότι τα συστήματα τεχνητής νοημοσύνης λειτουργούν σε κοινωνικά πλαίσια και οι επιπτώσεις τους εκτείνονται πέρα από τα όρια του οργανισμού που τα αναπτύσσει ή τα χρησιμοποιεί.

Εικόνα 6: NIST AI RMF

Το πλαίσιο δομείται γύρω από τέσσερις βασικές λειτουργίες διακυβέρνησης και διαχείρισης. Οι τέσσερις λειτουργίες του προτύπου είναι οι:

1. **Govern (Διακυβέρνηση):** μέσω αυτού καθορίζεται το πλαίσιο εντός του οποίου σχεδιάζονται, αναπτύσσονται και χρησιμοποιούνται τα συστήματα τεχνητής νοημοσύνης. Εστιάζει στη καθιέρωση πολιτικών και αρχών υπεύθυνης χρήσης του συστήματος, στον σαφή ορισμό ρόλων, αρμοδιοτήτων και μηχανισμών λογοδοσίας, καθώς και στην ενσωμάτωση ανθρώπινης εποπτείας στις αλγοριθμικές αποφάσεις. Παράλληλα, διασφαλίζει τη συμμόρφωση με το ισχύον κανονιστικό και νομικό πλαίσιο που διέπει τον Οργανισμό και το σύστημα Τεχνητής Νοημοσύνης εντός πεδίου εφαρμογής.
2. **Map (Χαρτογράφηση):** αποτυπώνει το πλαίσιο που λειτουργεί το σύστημα AI, ώστε να εντοπιστούν και να αποσαφηνιστούν οι σχετικοί κίνδυνοι. Περιλαμβάνει τη χαρτογράφηση του σκοπού του συστήματος, των δεδομένων που χρησιμοποιούνται, των χρηστών και των ατόμων ή ομάδων που ενδέχεται να

επηρεαστούν από τη λειτουργία του, καθώς και των πιθανών σεναρίων κινδύνου. Αποτέλεσμα της λειτουργίας αυτής αποτελεί η αναγνώριση όχι μόνο τεχνικών κίνδυνων, αλλά και κοινωνικών, νομικών και ηθικών επιπτώσεων, λαμβάνοντας υπόψη το περιβάλλον χρήσης και τις αλληλεπιδράσεις του συστήματος με άλλες διαδικασίες ή τεχνολογίες

3. **Measure (Μέτρηση):** αφορά την ποιοτική και ποσοτική αξιολόγηση των κινδύνων που έχουν εντοπιστεί κατά τη χαρτογράφηση του συστήματος. Αξιολογείται η απόδοση και η συμπεριφορά του συστήματος μέσω μετρήσεων, όπως η ακρίβεια, η αξιοπιστία, η ανθεκτικότητα σε επιθέσεις, η μεροληψία, η επεξηγησιμότητα και η ποιότητα των δεδομένων που χρησιμοποιούνται.
4. **Manage (Διαχείριση):** αφορά τη λήψη αποφάσεων και την εφαρμογή μέτρων αντιμετώπισης των κινδύνων που έχουν εντοπιστεί και αξιολογηθεί στις προηγούμενες λειτουργίες. Στο στάδιο αυτό, ο Οργανισμός επιλέγει μέτρα αντιμετώπισης, τα οποία μπορεί να περιλαμβάνουν τεχνικές παρεμβάσεις στο μοντέλο, αλλαγές στον τρόπο ή στο πλαίσιο χρήσης του συστήματος, ενίσχυση της ανθρώπινης εποπτείας ή ακόμη και την απόσυρση του συστήματος όταν ο υπολειπόμενος κίνδυνος κρίνεται μη αποδεκτός. Παράλληλα, στην παρούσα φάση περιλαμβάνονται ενέργειες συνεχούς παρακολούθησης της αποτελεσματικότητας των ενεργειών, διαδικασίες διαχείρισης αποκλίσεων αλλά και ενημέρωση των αποτελεσμάτων των προηγούμενων φάσεων ώστε να υλοποιηθούν οι απαραίτητες τροποποιήσεις. (Tabassi, 2023)

Η αξιολόγηση κινδύνων στο AI RMF δεν στοχεύει στην ακριβή ποσοτικοποίηση, αλλά στη κατανόηση και τεκμηρίωση των πιθανών επιπτώσεων, λαμβάνοντας υπόψη την αβεβαιότητα και τη δυναμική φύση των μοντέλων. Ιδιαίτερη έμφαση δίνεται στη συνεχή αξιολόγηση και παρακολούθηση, καθώς τα δεδομένα εισόδου, το περιβάλλον χρήσης και οι κοινωνικές συνθήκες μπορούν να μεταβάλουν σημαντικά το προφίλ κινδύνου ενός τέτοιου συστήματος. Σε σύγκριση με το NIST RMF, το AI RMF μετατοπίζει το στόχο από την ασφάλεια πληροφοριών και την ιδιωτικότητα αποκλειστικά, προς μια ολιστική θεώρηση της υπεύθυνης τεχνητής νοημοσύνης, ενώ παράλληλα, παραμένει συμβατό με υφιστάμενα πρότυπα όπως το ISO 42001 και το ISO 23894, λειτουργώντας ως συμπληρωματικό πλαίσιο και όχι ως υποκατάστατό.

4.4 Κανονισμός Ευρωπαϊκής Ένωσης (AI Act)

Όπως οι Οργανισμοί χρειάστηκε να πλαισιώσουν την ανάπτυξη και την χρήση της Τεχνητής Νοημοσύνης, έτσι και η Ευρωπαική Ένωση αντιμετώπισε την ανάγκη νομικού και θεσμικού πλαισίου για την Διαχείριτη της τεχνολογίας αυτής. Έτσι προέκυψε ο κανονισμός διαχείρισης συστημάτων τεχνητής Νοημοσύνης, γνωστός ως AI Act. Ο κανονισμός αυτός παρουσιάζει μια δομημένη, risk based προσέγγιση των συστημάτων αυτών. Σκοπός του είναι να διασφαλίσει πως τα συστήματα AI που χρησιμοποιούνται και διατίθεται στους πολίτες της Ευρωπαικής Ένωσης είναι ασφαλή και, κυρίως, σέβονται τα δικαιώματα και τις ελευθερίες των ατόμων.

Πρακτικά, μέσω του κανονισμού, τα συστήματα Τεχνητής Νοημοσύνης κατηγοριοποιούνται βάσει την επίπτωση που μπορεί να έχουν στα άτομα είτε σε επίπεδο ασφάλειας, είτε σε επίπεδο θεμελιωδών δικαιωμάτων. Η ταξινόμηση αυτή περιλαμβάνει τέσσερις κύριες κατηγορίες:

1. unacceptable risk συστήματα (μη αποδεκτού κινδύνου), που θεωρούνται απειλή για τον άνθρωπο όπως η κοινωνική βαθμολόγηση από κράτη ή βιομετρική ταυτοποίηση σε πραγματικό χρόνο σε χώρους δημόσιας πρόσβασης. Η χρήση τους απαγορεύεται.
2. τα high-risk συστήματα (υψηλού κινδύνου), που χρησιμοποιούνται σε κρίσιμους τομείς όπως οι υποδομές, η εκπαίδευση ή η επιβολή του νόμου και υπόκεινται στις αυστηρότερες απαιτήσεις αξιολόγησης όπως η ανθρώπινη παρέμβαση και η τήρηση αρχείων ενεργειών. Για τα Συστήματα Υψηλού Κινδύνου απαιτείται η δημιουργία ενός συστήματος διαχείρισης κινδύνων, το οποίο λειτουργεί συνεχώς καθ’ όλη τη διάρκεια ζωής του συστήματος και ενσωματώνει ενέργειες εντοπισμού, ανάλυσης και μετριασμού κινδύνων. Επιπλέον απαιτείται η διαβεβαίωση πως τα δεδομένα εκπαίδευσης είναι υψηλής ποιότητας, αντιπροσωπευτικά και απαλλαγμένα από μεροληψίες αλλά και τη κατάρτιση λεπτομερών αρχείων που αποδεικνύουν τη συμμόρφωση και την ιχνηλασιμότητα των αποφάσεων του συστήματος.
3. τα limited-risk (περιορισμένου κινδύνου) συστήματα, όπου απαιτείται κυρίως διαφάνεια προς τον χρήστη ώστε να γνωστοποιείται η χρήση τους, και
4. τα minimal-risk (χαμηλού κινδύνου) συστήματα, για τα οποία δεν υπάρχουν πρόσθετες υποχρεώσεις. (OECD, 2023)

Ο Κανονισμός υιοθετεί μια διαφορετική, κανονιστική προσέγγιση, βασισμένη στην αρχή της διαβάθμισης κινδύνου. Αντί να καθορίζει μια τυπική διαδικασία αξιολόγησης κινδύνων, ο κανονισμός κατηγοριοποιεί τα συστήματα ανάλογα με το επίπεδο κινδύνου που ενέχουν για τα θεμελιώδη δικαιώματα και την ασφάλεια των ατόμων που τα χρησιμοποιούν ή εμπλέκονται σε αυτά έμμεσα ή άμεσα. Σε αντίθεση με τα πρότυπα ISO και NIST, δεν προσφέρει συγκεκριμένη μεθοδολογία υπολογισμού του κινδύνου, αλλά θέτει νομικά δεσμευτικά όρια, μετατρέποντας την αξιολόγηση κινδύνων σε κανονιστική υποχρέωση, ανάλογα με την διαβάθμιση που έχει δοθεί στο εκάστοτε σύστημα. Εφόσον απαιτηθεί νομικά να υλοποιηθεί η αξιολόγση, τότε υπάρχει η δυνατότητα επιλογής ενός εκ των προαναφερθέντων πλαισίων για την διεξαγωγή της.

Στην πραγματικότητα υφίστανται πολλά και διαφορετικά πλαίσια υλοποίησης αξιολόγησης κινδύνων, που είτε αφοσιώνονται στα συστήματα τεχνητής νοημοσύνης, είτε αποτελούν ένα γενικό εργαλείο για την αποτύπωση κινδύνων ασφάλειας και ιδιωτικότητας. Τα εργαλεία αυτά δεν υπερκαλύπτουν ή «ανταγνωίζονται» το ένα το άλλο. Εν αντιθέσει, αλληλοσυμπληρώνονται και μπορούν να αξιοποιηθούν με διαφορετικούς τρόπους για να μπορέσει ένας Οργανισμός να προσεγγίσει σφαιρικά ένα ζήτημα τόσο σύνθετο και ταχέως εξελισσόμενο όσο η Τεχνητή Νοημοσύνη. Το πλαίσιο του ISO προσφέρει έναν πολύ δομημένο τρόπο με τον οποίο προσεγγίζει το ζήτημα της αξιολόγησης κινδύνων, εστιάζοντας στη δημιουργία ενός Συστήματος Διαχείρισης (AIMS), ενσωματώνοντας τη διαχείριση στις διοικητικές δομές, τις πολιτικές και τις διαδικασίες του οργανισμού. Το ISO/IEC 23894 συμπληρώνει το ISO/IEC 42001 παρέχοντας μια τεχνική προσέγγιση. Εστιάζει στον εντοπισμό πηγών κινδύνου (risk sources), στις επιπτώσεις και στους στόχους ασφάλειας και αξιοπιστίας.

Τα πλαίσια του NIST από την άλλη, υιοθετούν μια πιο ευέλικτη προσέγγιση, οργανώνοντας τη διαχείριση κινδύνων σε διαρκώς επαναλαμβανόμενες φάσεις καθόλη την διάρκεια υλοποίησης ενός συστήματος. Το πλαίσιο αυτό διαφοροποιείται από τα πρότυπα ISO καθώς δεν έχει τη μορφή συστήματος διαχείρισης, αλλά λειτουργεί ως οδηγός πρακτικής εφαρμογής, διευκολύνοντας τη σύνδεση της ανάλυσης κινδύνων με τις πραγματικές διαδικασίες ανάπτυξης και λειτουργίας συστημάτων AI.

Τέλος, ολοκληρώνοντας το γενικότερο πλαίσιο διαχείρισης κινδύνων, ο Ευρωπαικός κανονισμός του AI ACT ενσωματώνει τις ρυθμιστικές και κανονιστικές απαιτήσεις και την δέσμευση προς αυτές. Συμβάλλει στην κατηγοριοποίηση των συστημάτων και στην αναγνώριση των ενεργειών που πρέπει να υλοποιηθούν ανάλογα με την κατηγοριοποίηση που έχει προκύψει. Ωστόσο, ο AI Act δεν παρέχει αναλυτική

μεθοδολογία υλοποίησης της διαδικασίας διαχείρισης κινδύνων σε οργανωσιακό ή τεχνικό επίπεδο, γεγονός που καθιστά αναγκαία τη συμπληρωματική αξιοποίηση προτύπων και πλαισίων όπως το ISO και το NIST.

Στο παράρτημα Α της παρούσας εργασίας κατατίθεται συγκριτικός πίνακας μεταξύ των πλαισίων για την καλύτερη κατανοήση των διαφορών και των τρόπων με τον οποίον καλύπτουν το ένα το άλλο.

5. Χαρτογράφηση Κινδύνων Τεχνητής Νοημοσύνης

Η Τεχνητή Νοημοσύνη, παρόλη την διάδοση της, αποτελεί μια νεοσύστατη τεχνολογία. Όπως αναφέραμε και σε προηγούμενα κεφάλαια, σε παγκόσμιο επίπεδο, έχουν υλοποιηθεί προσπάθειες να πλαισιωθεί και σε ρυθμιστικό και κανονιστικό επίπεδο (AI Act κλπ.) αλλά και σε οργανωτικό και τεχνικό επίπεδο, με πρώτο βήμα την ανάλυση των κινδύνων που την συνοδεύουν. Τα Frameworks που έχουν συνταχθεί και εφαρμοστεί έχουν βοηθήσει σε πολύ μεγάλο βαθμό στην κατανόηση, την ανάπτυξη και την χρήση της Τεχνητής Νοημοσύνης. Παρόλα αυτά, ακριβώς λόγω της ταχύρρυθμής ενσωμάτωσης της στην καθημερινότητα και στην άμεση έκθεση των οργανισμών και των χρηστών στην τεχνολογία αυτή, δεν έχει υπάρξει ο απαραίτητος χρόνος για προσαρμογή και κατανόηση των κινδύνων που συνοδεύουν το AI.

Έτσι, αναδύεται η ανάγκη για πλαίσια και εργαλεία αποτύπωσης και χαρτογράφησης κινδύνων προσαρμοσμένα στα ιδιαίτερα χαρακτηριστικά του AI. Τα πλαίσια αυτά στοχεύουν στη συστηματική αναγνώριση και οπτικοποίηση των κινδύνων, λαμβάνοντας υπόψη όχι μόνο τεχνικές ευπάθειες και λειτουργικούς κινδύνους ασφάλειας και ιδιωτικότητας, αλλά και ζητήματα διακυβέρνησης, κανονιστικής συμμόρφωσης, κοινωνικών επιπτώσεων και ηθικών παραμέτρων. Μέσω της πολυδιάστατης απεικόνισης των κινδύνων, διευκολύνεται η κατανόηση των αλληλεπιδράσεων μεταξύ τους και η ιεράρχησή τους με τρόπο που να υποστηρίζει τεκμηριωμένες αποφάσεις διαχείρισης. Τα εργαλεία αυτά μπορούν να λειτουργήσουν συμπληρωματικά προς τα καθιερωμένα πρότυπα και κανονιστικά πλαίσια διαχείρισης, όπως το ISO/IEC 42001, και το NIST AI Risk Management Framework ή ως είσοδο σε μια διαδικασία αναγνώρισης, αποτύπωσης και ανάλυσης, καθώς συγκεντρώνουν πληθώρα ρίσκων σε μορφή κατανοητή και διαχωρισμένη ανάλογα το είδος κινδύνου που αναζητά ο εκάστοτε Οργανισμός ή υπεύθυνος υλοποίησης. Δεν μπορούν να αντικαταστήσουν ένα σύστημα διαχείρισης κινδύνων αλλά μπορούν να συνεισφέρουν σημαντικά στην αναγνώριση και την ανάλυσή τους. Στο παρόν κεφάλαιο εξετάζονται ορισμένα υφιστάμενα πλαίσια αποτύπωσης κινδύνων.

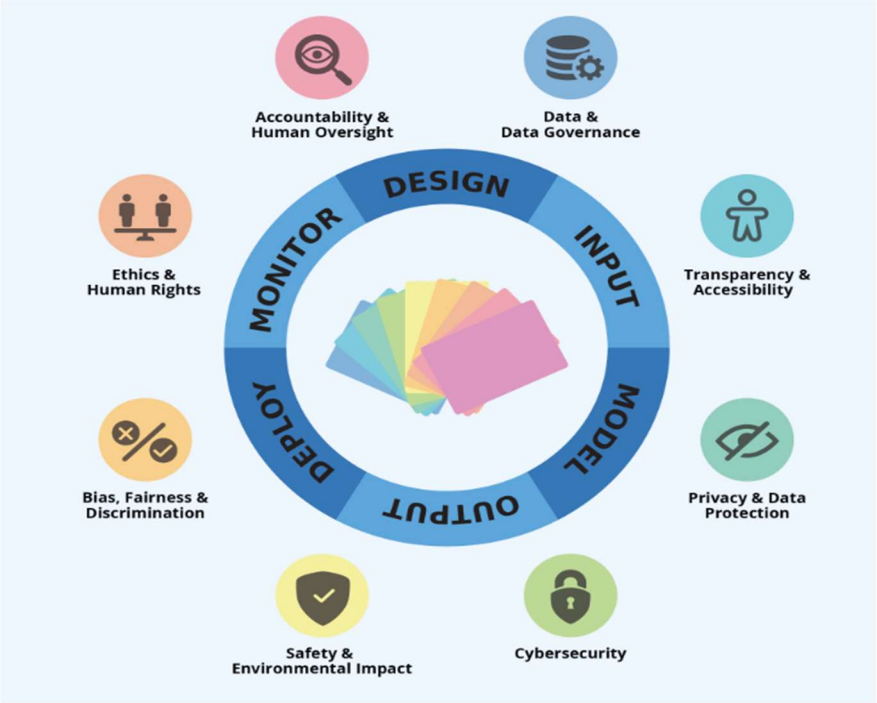
5.1 Plot4AI

5.1.1 Παρουσίαση Εργαλείου

Το PLOT4ai είναι μια ολοκληρωμένη μεθοδολογία μοντελοποίησης απειλών που έχει σχεδιαστεί για προγραμματιστές και υπεύθυνους ανάπτυξης συστημάτων Τεχνητής Νοημοσύνης, με σκοπό την αναγνώριση και τον μετριασμό των κινδύνων που

σχετίζονται με την τεχνητή νοημοσύνη. Περιλαμβάνει μια βιβλιοθήκη με 138 ρίσκα, καλύπτοντας 8 τομείς. Μέσω του Εργαλείου αυτού υπάρχει η δυνατότητα:

1. Μοντελοποίησης απειλών σε συστήματα τεχνητής νοημοσύνης/μηχανικής μάθησης και γενετικής τεχνητής νοημοσύνης, από το σχεδιασμό έως την ανάπτυξη και την παρακολούθηση.
2. Εντοπισμού και μετριασμού των κινδύνων, από παραβιάσεις της ιδιωτικής ζωής έως αστοχίες μοντέλων, συστηματική μεροληψία, κατάχρηση και κακόβουλες επιθέσεις.
3. Εφαρμογή της προστασίας της ιδιωτικότητας κατά τον σχεδιασμό (privacy by design), και την ασφάλεια κατά τον σχεδιασμό (Security by Design), σύμφωνα με τον ΓΚΠΔ, τον AI Act και άλλα ρυθμιστικά πλαίσια.
4. Δημιουργία αξιόπιστης τεχνητής νοημοσύνης, διασφαλίζοντας ότι τα συστήματά δεν είναι μόνο αποτελεσματικά, αλλά και υπεύθυνα, διαφανή και ανθρωποκεντρικά.



Εικόνα 7: Μοντέλο Plot4AI

Η βιβλιοθήκη απειλών του PLOT4AI αποτελεί ένα διαδικτυακό εργαλείο συνοδευόμενο από ένα εργαλείο αξιολόγησης για την εφαρμογή της προτεινόμενης μεθοδολογίας. Είναι δομημένο σε καρτέλες βάσει 8 κατηγοριών οι οποίες είναι οι εξής:

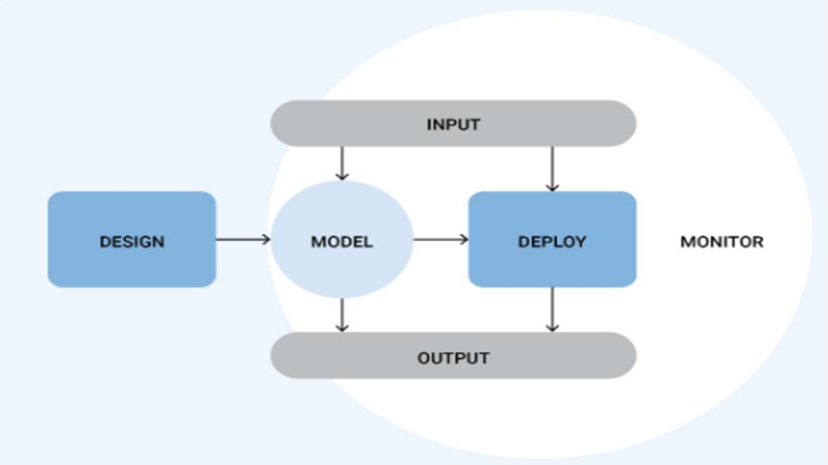
- **Δεδομένα και διακυβέρνηση δεδομένων**: Αφορά την Ανεπαρκή διαχείριση και ποιότητα των δεδομένων που χρησιμοποιούνται σε ένα σύστημα τεχνητής νοημοσύνης, με αποτέλεσμα ανακριβή και επιβλαβή αποτελέσματα.
- **Διαφάνεια και προσβασιμότητα**: Αφορά τις αποφάσεις ή τις αλληλεπιδράσεις της τεχνητής νοημοσύνης που δεν είναι κατανοητές ή προσβάσιμες σε όλους τους χρήστες, γεγονός που περιορίζει τη χρηστικότητα του συστήματος και μειώνει την εμπιστοσύνη προς αυτό.
- **Ιδιωτικότητα και προστασία δεδομένων:** Αφορά την έλλειψη κατάλληλων μέτρων προστασίας των προσωπικών δεδομένων, με αποτέλεσμα την αύξηση του κινδύνου μη εξουσιοδοτημένης πρόσβασης, κατάχρησης ή νομικών παραβιάσεων.
- **Κυβερνοασφάλεια**: Αφορά ανεπαρκή μέτρα ασφαλείας που μπορούν να οδηγήσουν σε παραβιάσεις δεδομένων, κακόβουλες επιθέσεις ή χειραγώγηση του συστήματος.
- **Ασφάλεια και περιβαλλοντικές επιπτώσεις:** Αφορά κίνδυνους και πηγές κινδύνων που ενδέχεται να προκαλέσουν βλάβη σε εργαζόμενους, χρήστες, υποδομές ή το περιβάλλον.
- **Προκατάληψη, δικαιοσύνη και διάκριση:** Αφορά την παρουσία μεροληψίας στα δεδομένα ή στο σχεδιασμό του συστήματος, που οδηγεί σε άδικη μεταχείριση ατόμων ή ομάδων.
- **Ηθική και θεμελιώδη δικαιώματα:** Αφορά την παράβλεψη του ηθικού και κοινωνικού αντίκτυπου του συστήματος, που μπορεί να οδηγήσει σε παραβίαση ενός ή περισσότερων ανθρωπίνων δικαιωμάτων ή σε ακούσια βλάβη.
- **Λογοδοσία και ανθρώπινη εποπτεία:** Αφορά την ασαφή ευθύνη για τις αποφάσεις του συστήματος τεχνητής νοημοσύνης και έλλειψη μηχανισμών ανθρώπινης εποπτείας.

Ο υπεύθυνος υλοποίησης καλείται να απαντήσει στις καρτέλες αυτές και ανάλογα την απάντηση που θα δοθεί, πραγματοποιείται η αξιολόγηση των μέτρων που έχουν ήδη ληφθεί, των μέτρων που δεν έχουν ληφθεί υπόψιν κατά την ανάπτυξη και την λειτουργία του συστήματος και γίνεται αναφορά σε πιθανές διορθωτικές ενέργειες που πρέπει να υλοποιηθούν για την αντιμετώπιση του Κινδύνου (Risk Treatment). Επιπλέον, διαχωρίζει τον κύκλο ανάπτυξης ενός συστήματος σε 6 διακριτές φάσεις για

την περαιτέρω κατηγοριοποίηση και βελτιστοποίηση της εύρεσης των κινδύνων που αφορούν το εκάστοτε σύστημα. Οι φάσεις αυτές, σύμφωνα με το εργαλείο, είναι οι εξής:

1. **Design**: Η φάση σχεδιασμού ενός συστήματος τεχνητής νοημοσύνης περιλαμβάνει τον καθορισμό του σκοπού, της αρχιτεκτονικής και της προσέγγισης του συστήματος, συμπεριλαμβανομένων των αποφάσεων που αφορούν το Hardware, το λογισμικό και τους πόρους που θα χρησιμοποιηθούν.
2. **Input**: Σε αυτή τη φάση, τα δεδομένα είτε συλλέγονται είτε επιλέγονται για χρήση. Αυτό ισχύει και κατά τη διάρκεια της ανάπτυξης του μοντέλου και μετά την εφαρμογή του. Η ποιότητα και η ακεραιότητα των δεδομένων επηρεάζουν άμεσα την απόδοση και την αξιοπιστία του τελικού μοντέλου.
3. **Model:** Κατά τη μοντελοποίηση, τα μοντέλα αναπτύσσονται, εκπαιδεύονται στα σύνολα δεδομένων και βελτιστοποιούνται. Αυτή η φάση επηρεάζει άμεσα την απόδοση και την αποτελεσματικότητα του συστήματος τεχνητής νοημοσύνης.
4. **Output:** Αυτή η φάση επικεντρώνεται στα αποτελέσματα που παράγει το σύστημα μετά την εξαγωγή συμπερασμάτων από το μοντέλο. Περιλαμβάνει την ερμηνεία, την παρουσίαση και την αξιοποίηση των αποτελεσμάτων (output) του συστήματος, διασφαλίζοντας παράλληλα ότι αυτά ανταποκρίνονται στις προσδοκίες και στην προβλεπόμενη χρήση.
5. **Deploy:** Σε αυτή τη φάση, το σύστημα τεχνητής νοημοσύνης μεταβαίνει από την ανάπτυξη στην εγκατάσταση, την κυκλοφορία ή τη διαμόρφωση (configuration) στο παραγωγικό περιβάλλον. Το στάδιο αυτό εισάγει νέες προκλήσεις, όπως παράγοντες που σχετίζονται με το πλαίσιο υλοποίησης και δεν είχαν ληφθεί υπόψη κατά το σχεδιασμό και την ανάπτυξη του.
6. **Monitor:** Η παρούσα φάση καλύπτει ολόκληρο τον κύκλο ζωής ανάπτυξης του συστήματος. Κατά τη διάρκεια της ανάπτυξης, περιλαμβάνονται δοκιμές και επαλήθευση της λειτουργικότητας για τον εντοπισμό προβλημάτων. Μετά την εφαρμογή (Deploy), η συνεχής

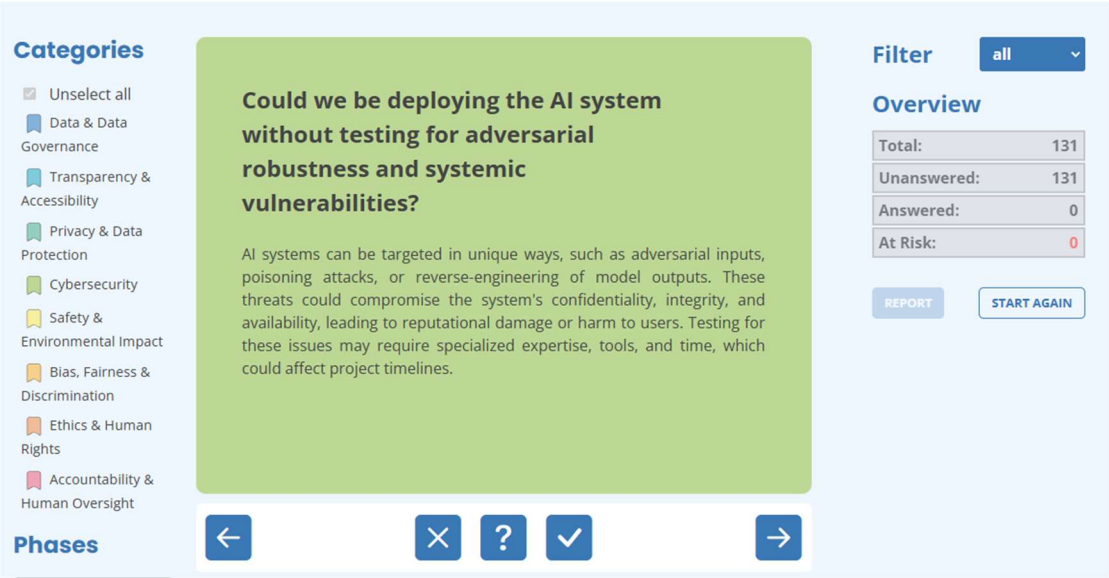
παρακολούθηση εξασφαλίζει την απόδοση του συστήματος, τον εντοπισμό λαθών και κινδύνων και την συνεχή υποστήριξη μέσω αναβαθμίσεων. (Barberá, 2022)



Εικόνα 8: Κύκλος Ανάπτυξης Plot4AI

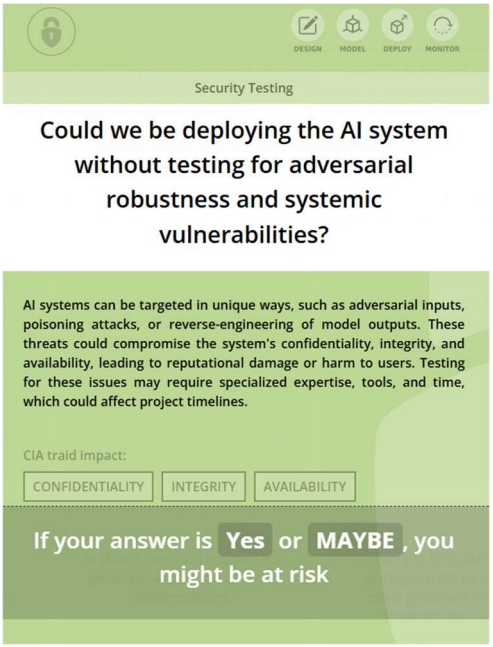
5.1.2 Παρουσίαση Εργαλείου

Το εργαλείο λειτουργεί μέσω ερωτήσεων οι οποίες μπορούν να απαντηθούν με **ΝΑΙ, ΟΧΙ, ΔΕΝ ΕΊΜΑΙ ΣΙΓΟΥΡΟΣ,** λαμβάνοντας υπόψιν το είδος του συστήματος προς ανάπτυξη και τον ρόλο του υπεύθυνου σε αυτό (provider / Deployer). Επιλέγοντας τις κατηγορίες πάνω στις οποίες ο υπεύθυνος επιθυμεί να αξιολογηθεί ξεκινάει η διαδικασία μέχρι την κάλυψη όλων των ερωτήσεων εντός του πεδίου εφαρμογής, σύμφωνα με τις απαντήσεις του χρήστη. Το αποτέλεσμα του θα συντάξει μια σχετική αναφορά με τις απαντήσεις και τις διορθωτικές ενέργειες.

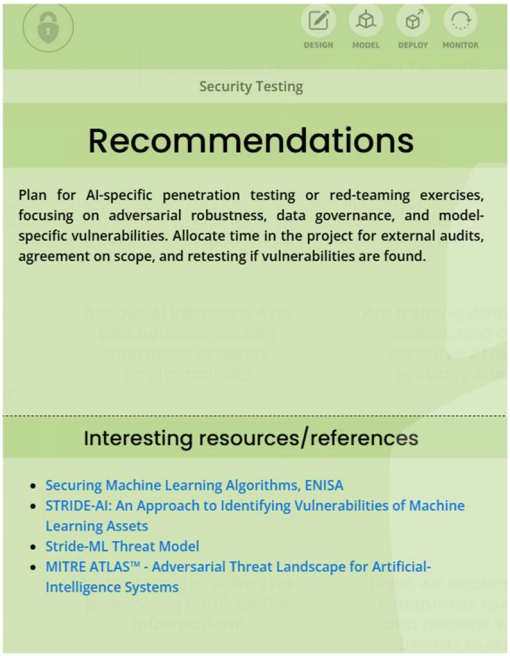


Εικόνα 9: Εργαλείο Αξιολόγησης Plot4AI

Παράλληλα, ο υπεύθυνος, σε περίπτωση που δεν επιθυμεί να αξιολογήσει το σύστημα αλλά να ανατρέξει σε πιθανούς κινδύνους, υπάρχει διαθέσιμη βιβλιοθήκη με τους κινδύνους που έχουν εντοπιστεί αλλά και τις διορθωτικές ενέργειες που τους συνοδεύουν.



Εικόνα 11: Κίνδυνος AI



Εικόνα 10: Αντιμεώπιση Κινδύνου

5.2 MITRE ATLAS

Το MITRE Adversarial Threat Landscape for AI Systems (ATLAS) είναι μια παγκοσμίως προσβάσιμη, ζωντανή βάση δεδομένων με τακτικές και τεχνικές κακόβουλων οντοτήτων, η οποία βασίζεται στην παρατήρηση πραγματικών επιθέσεων από red teams (Ομάδες που δοκιμάζουν την τρωτότητα συστημάτων) και ομάδες ασφάλειας τεχνητής νοημοσύνης. Η Ανάπτυξη του ATLAS αφορά και έχει ως στόχο την ευαισθητοποίηση και την προετοιμασία των Οργανισμών για την αντιμετώπιση των απειλών, των ευπαθειών και γενικότερα των κινδύνων που σχετίζονται με κακόβουλες επιθέσεις σε AI συστήματα. Έχει σχεδιαστεί με βάση το πλαίσιο MITRE ATT&CK το οποίο συνοδεύει και συμπληρώνει για μια πιο ολοκληρωμένη προσέγγιση της ασφάλειας πληροφορίων. Το εργαλείο μπορεί να χρησιμοποιηθεί για:

1. Την ενημέρωση των αναλυτών ασφάλειας και των προγραμματιστών / υλοποιητών AI σχετικά με ρεαλιστικές απειλές.

2. Την αξιολόγηση απειλών και την υλοποίηση red teaming δοκιμών.
3. Την κατανόηση των συμπεριφορών των κακόβουλων οντοτήτων και των τρόπων αντιμετώπισης.
4. Την αναφορά κακόβουλων επιθέσεων στον πραγματικό κόσμο σε συστήματα με ενσωματωμένο AI.

Το εργαλείο εστιάζει στις τεχνικές, τακτικές και διαδικασίες (Tactics, Techniques and Procedures – TTPs) που μπορούν να χρησιμοποιηθούν από κακόβουλους παράγοντες για να υπονομεύσουν την ακεραιότητα, τη διαθεσιμότητα, την εμπιστευτικότητα και την αξιοπιστία συστημάτων Τεχνητής Νοημοσύνης. Η βασική φιλοσοφία του είναι η χαρτογράφηση του κύκλου ζωής μιας επίθεσης σε ένα σύστημα, από τη φάση της αναγνώρισης και της προετοιμασίας (reconnaissance) έως τη φάση της επιβολής επιπτώσεων (Impact).

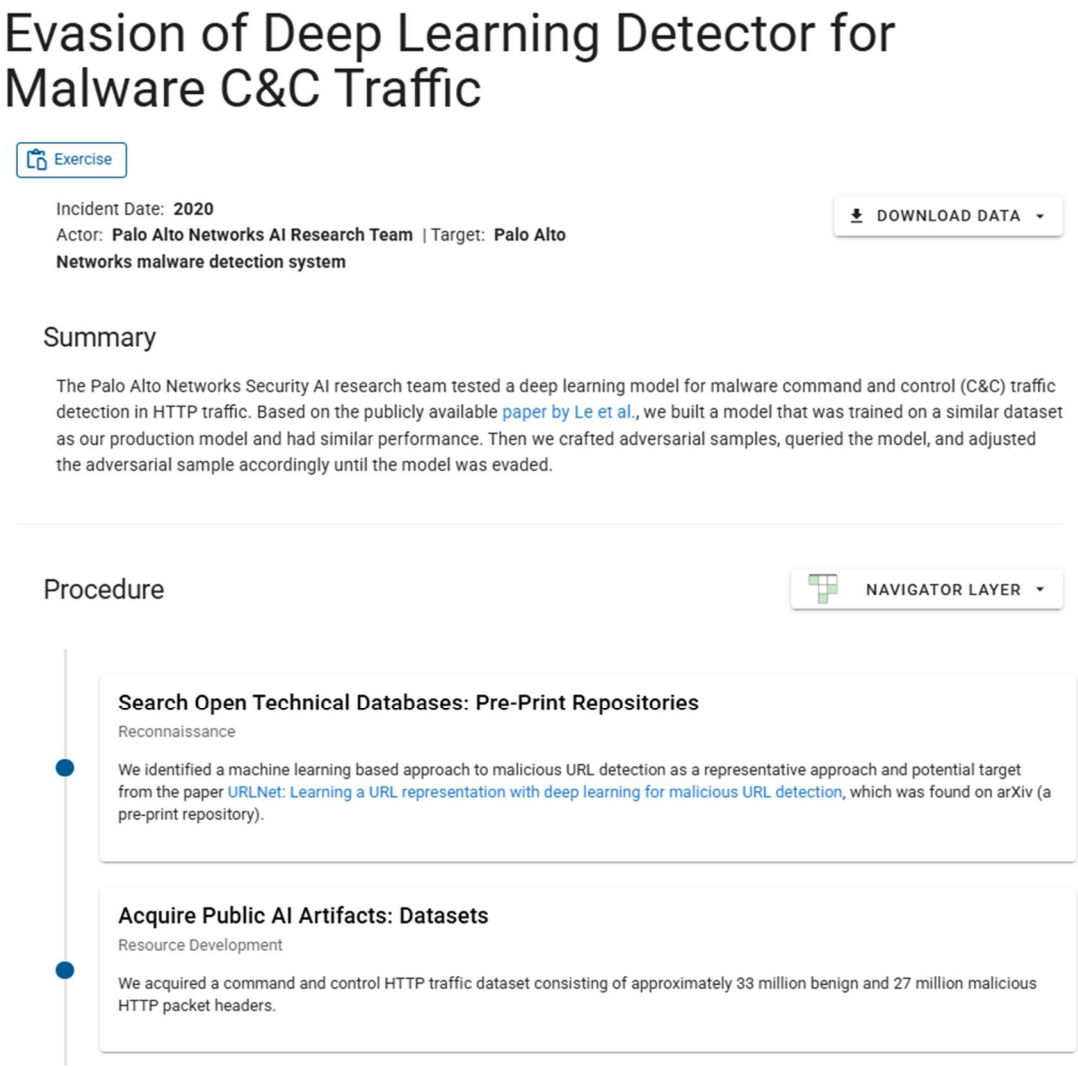
Οι επιθέσεις οργανώνονται σε διακριτές φάσεις που το εργαλείο αποτυπώνει ως τακτικές. Συγκεκριμένα το εργαλείο διακρίνει δέκα έξι τακτικές και τις αποτυπώνει ως εξής:

1. **Reconnaissance**: ο κακόβουλος προσπαθεί να συλλέξει πληροφορίες σχετικά με το σύστημα AI στο οποίο θέλει να επιτεθεί, ώστε να μπορεί να προγραμματίσει τις επόμενες ενέργειες.
2. **Resource Development**: η προσπάθεια συλλογής και δημιουργίας πόρων για την υλοποίηση της επίθεσης. Οι πόροι θα μπορούσαν να αποτελούν υποδομή, λογαριασμοί χρηστών και υποσυστήματα, μεταξύ άλλων.
3. **Initial Access**: η προσπάθεια αρχικής πρόσβασης στα συστήματα.
4. **AI Model Access:** η απόκτηση πρόσβασης στο μοντέλο τεχνητής νοημοσύνης (AI Model), είτε αυτή είναι στο λογισμικό του συστήματος, είτε στο περιβάλλον που τηρούνται τα δεδομένα.
5. **Execution:** η απόπειρα εκτέλεσης κακόβουλου κώδικα.
6. **Persistence:** η απόπειρα διατήρησης της πρόσβασης που έχει αποκτήσει
7. **Privilege Escalation:** η απόπειρα απόκτησης πρόσβασης με περισσότερα δικαιώματα.
8. **Defense Evasion:** η αποφυγή εντοπισμού των ενεργειών του από συστήματα και δικλείδες ασφαλείας.
9. **Credential Access:** απόπειρα κλοπής λογαριασμών, κωδικών χρηστών και διαχειριστών.

10. **Discovery:** Χαρτογράφηση του περιβάλλοντος AI.
11. **Lateral Movement:** η περιήγηση στο περιβάλλον AI.
12. **Collection:** Συλλογή πληροφοριών και υποσυστημάτων.
13. **AI Attack Staging:** Προετοιμασία της επίθεσης βάσει της πληροφορίας που έχει συλλέξει ο επιτιθέμενος.
14. **Command and Control:** Απόπειρα επικοινωνίας και ελέγχου με τα συστήματα.
15. **Exfiltration:** Συλλογή πληροφορίας από τα συστήματα υπό επίθεση.
16. **Impact:** Ο επιτιθέμενος προσπαθεί να χειραγωγήσει, να διακόψει, να υπονομεύσει την εμπιστοσύνη ή να καταστρέψει τα συστήματα τεχνητής νοημοσύνης και τα δεδομένα που εμπεριέχονται σε αυτά.

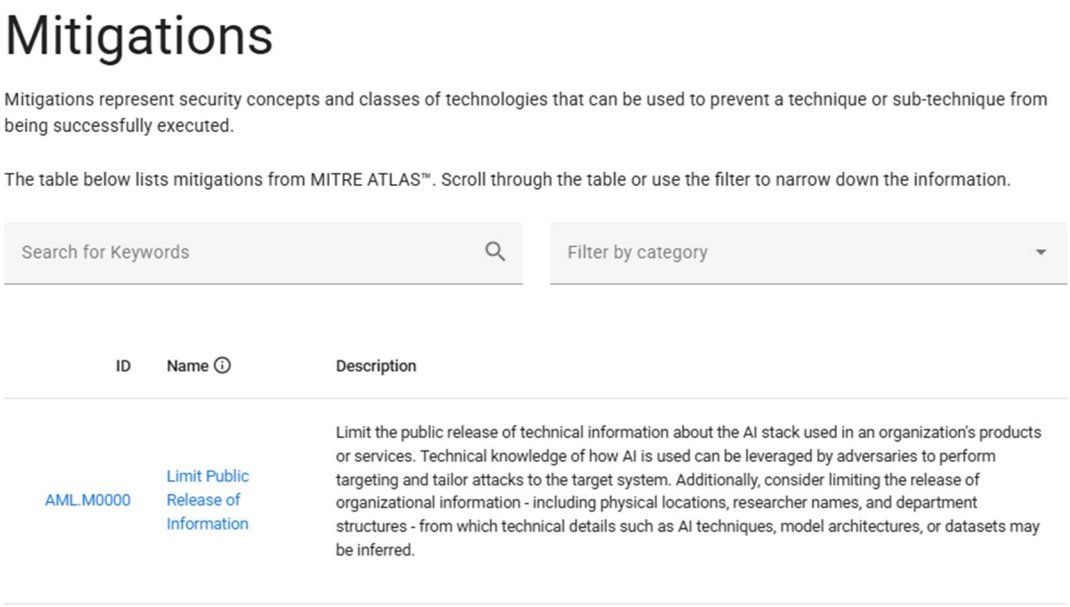
Ανάλογα με την φάση στην οποία βρίσκεται ο επιτιθέμενος, το εργαλείο παρουσιάζει τεχνικές (techniques) που μπορεί να αξιοποιήσει ώστε να πετύχει τον απώτερο σκοπό της κάθε φάσης. Για παράδειγμα κατά την φάση «AI Model Access», αποτυπώνονται τεχνικές όπως **Full AI Model Access** όπου οι επιτιθέμενοι μπορούν να αποκτήσουν πλήρη πρόσβαση «white-box» σε ένα μοντέλο τεχνητής νοημοσύνης. Αυτό σημαίνει ότι έχουν πλήρη γνώση της αρχιτεκτονικής του μοντέλου και των παραμέτρων του. Έτσι μπορούν να εξάγουν το μοντέλο για να δημιουργήσουν κακόβουλα δεδομένα και να επαληθεύσουν την υλοποίηση της επίθεσης σε ένα περιβάλλον εκτός σύνδεσης, όπου είναι δύσκολο να ανιχνευθεί η κίνησή τους.

Τέλος, το MITRE ATLAS επιτρέπει τη σύνδεση συγκεκριμένων απειλών με πιθανά σενάρια επίθεσης (Use Cases) και επιπτώσεις. Παρουσιάζει πραγματικές επιθέσεις που έχουν υλοποιηθεί ανά διαστήματα συμπεριλαμβάνοντας πληροφορίες, όπως τα μέσα και τις τεχνικές που χρησιμοποίησαν οι κακόβουλοι, αλλά και τις επιπτώσεις που προέκυψαν από αυτές. Οι οργανισμοί μπορούν να χρησιμοποιήσουν το εργαλείο για να εντοπίσουν ευπάθειες στα συστήματα τους, να αξιολογήσουν την έκθεσή τους σε συγκεκριμένες κατηγορίες επιθέσεων και να σχεδιάσουν κατάλληλα μέτρα αντιμετώπισης. Με τον τρόπο αυτό, το πλαίσιο υποστηρίζει τόσο την προληπτική ασφάλεια όσο και την ενίσχυση της ανθεκτικότητας των συστημάτων AI.



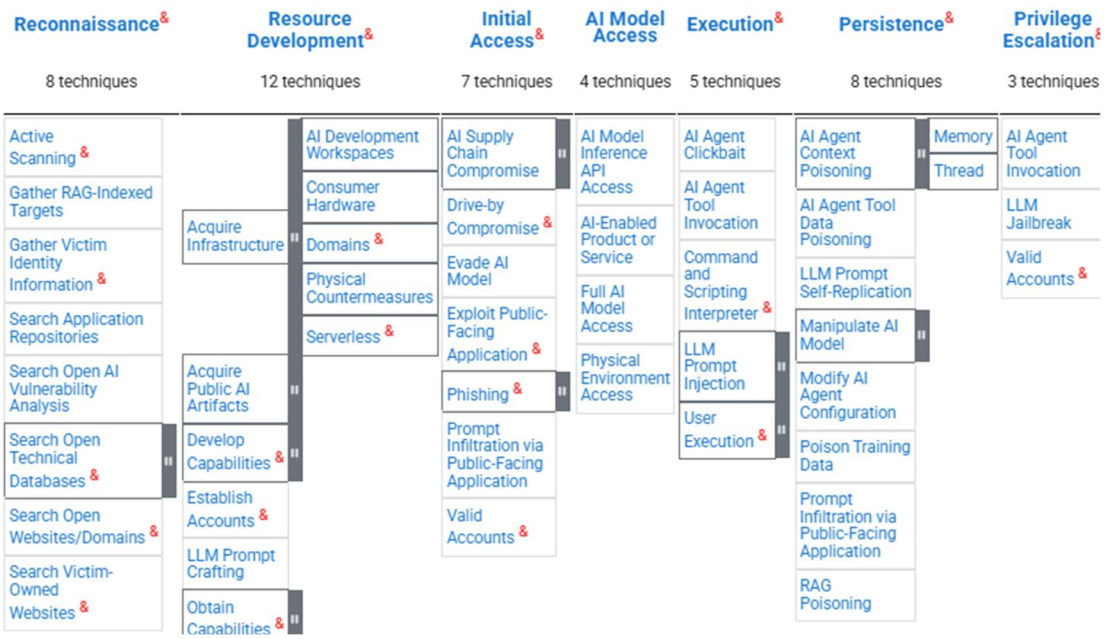
Εικόνα 12: Use Cases

Για κάθε απειλή και πιθανή τεχνική παρουσιάζονται και αντίστοιχες ενέργειες πρόληψης ή αντιμετώπισης. Από συγκεκριμένες τεχνολογίες ή μεθόδους σε βέλτιστες πρακτικές αντιμετώπισης.



Εικόνα 13: Ενέργειες Αντιμετώπισης

Όλη αυτή η πληροφορία αποτυπώνεται σε έναν κεντρικό πίνακα που διαχωρίζεται στις φάσεις που αναφέρθηκαν παραπάνω και στις τεχνικές που συνοδεύουν την εκάστοτε φάση, κάνοντας αναφορές και στον πίνακα ATT&CK όπου υπάρχει άμεση σύνδεση.



Εικόνα 14: Mitre ATLAS Matrix

Το MITRE ATLAS αποτελεί ένα σημαντικό εργαλείο για την κατανόηση και την αντιμετώπιση των σύγχρονων απειλών κατά συστημάτων Τεχνητής Νοημοσύνης, προσφέροντας μια δομημένη, τεχνικά τεκμηριωμένη και εύκολα κατανοητή

προσέγγιση. Η ένταξή του σε μια διαδικασία αξιολόγησης κινδύνων ενισχύει σημαντικά την ικανότητα των οργανισμών να εντοπίζουν, να αξιολογούν και να μετριάζουν εξειδικευμένους κινδύνους που προκύπτουν από τη χρήση τέτοιων συστημάτων.

5.3 AIR Repository (AI Risk)

Το AIR Repository αποτελεί ένα εργαλείο που δημιουργήθηκε από μια ομάδα του MIT με σκοπό να χαρτογραφήσουν τους κινδύνους της Τεχνητής Νοημοσύνης. Αποτελείται από τρία διαφορετικά συστήματα. Στο πρώτο αποτυπώνονται πάνω από 1700 κίνδυνοι που συμβαδίζουν με πρότυπα και πλαίσια που έχουν δημοσιευθεί αλλά και με διαφορετικά είδη της τεχνολογίας. Το δεύτερο αποτελεί την ταξινόμηση των αιτιών που προκύπτουν οι κίνδυνοι στο AI (Causal Taxonomy of AI Risks), ενώ το τρίτο παρουσιάζει την ταξινόμηση των τομέων στους οποίους ανήκουν οι κίνδυνοι (Domain Taxonomy of AI Risks). Η βάση αποτελεί το αποτέλεσμα έρευνας στο πλαίσιο του έργου MIT AI RISK Initiative, το οποίο είχε ως σκοπό την εκπαίδευση και την ενημέρωση του κόσμου και των χρηστών για την χρήση της Τεχνητής Νοημοσύνης και τις βέλτιστες πρακτικές σχετικά με την διαχείριση των κινδύνων της τεχνολογίας αυτής.

Η βάση αποτελεί ένα δωρεάν εργαλείο διαθέσιμο στο κοινό. Οι υπεύθυνοι υλοποίησης ενημερώνουν διαρκώς το περιεχόμενο με νέους κινδύνους που προκύπτουν ή σύμφωνα με την εξέλιξη των πλαισίων και των προτύπων που είναι διαθέσιμα. Λόγω αυτού, όλοι οι κίνδυνοι που ενσωματώνονται στην βάση, τεκμηριώνονται από κάποιο σχετικό άρθρο ή συγκεκριμένα πειστήρια που τεκμηριώνουν την υπόστασή του ως κίνδυνο. Παράλληλα, υπάρχουν αντίγραφα της βάσης διαθέσιμα προς χρήση ώστε το κοινό να μπορεί να την δοκιμάσει και να την κατανοήσει ή ακόμα και να συνεισφέρει στην ενημέρωσή της.

5.3.1 Causal Taxonomy of AI Risks

Η πρώτη ταξινόμηση αφορά τις αιτίες που προκύπτουν οι κίνδυνοι. Συγκεκριμένα, οι αιτίες αναφέρονται:

1. στο μέσο από το οποίο προκύπτουν (Entity). Στην ταξινόμηση παρουσιάζονται τρεις οντότητες:

a. AI: η αιτία προσδίδεται σε σύστημα τεχνητής νοημοσύνης λόγω της απόφασης ή της ενέργειας που πήρε.
b. Human: η αιτία προσδίδεται σε ανθρώπινο παράγοντα λόγω της απόφασης ή της ενέργειας που πήρε.
c. Other: άλλη αιτιότητα.

2. Στο σκοπό που είχε η οντότητα (Intent):
    a. Intentional: εσκεμμένο αποτέλεσμα ενέργειας το οποίο προέκυψε από την επιδίωξη συγκεκριμένου στόχου
    b. Unintentional: ακούσιο αποτέλεσμα ενέργειας το οποίο προέκυψε από την επιδίωξη συγκεκριμένου στόχου
    c. Other: απροσδιόριστου σκοπού
3. Στο χρονικό σημείο στο οποίο προέκυψε (Timing):
    a. Pre-deployment: Πριν από την εγκατάσταση της τεχνητής νοημοσύνης
    b. Post-Deployment: Αφού το μοντέλο τεχνητής νοημοσύνης έχει εκπαιδευτεί και εγκατασταθεί
    c. Other: Χωρίς σαφώς καθορισμένο χρόνο εμφάνισης


Intentional

QuickRef	Risk category	Risk subcategory	Description	Entity	Intent	Timing
Cui2024	Harmful Content		"The LLM-generated content sometimes contains biased, toxic, and private information"	2 - AI	2 - Unintentional	2 - Post-deployment

‹ 7 / 685 ›

Εικόνα 15: Causual Taxonomy

Οι διαφορετικές αιτίες λειτουργούν σαν ετικέτες (Tags) τις οποίες μπορεί να χρησιμοποιήσει ο χρήστης για να εντοπίσει τους κινδύνους που επιθυμεί.

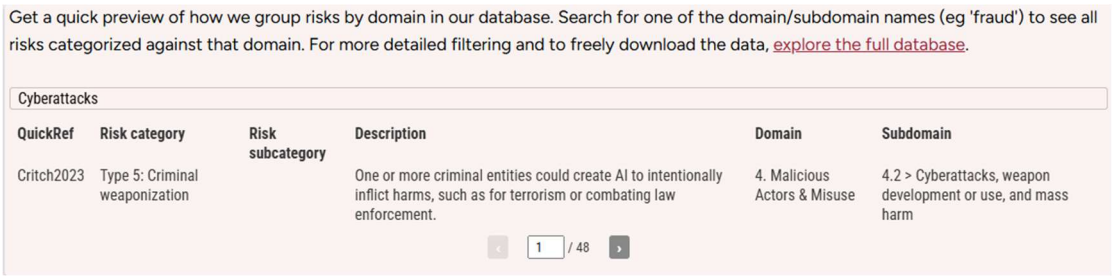
5.3.2 Domain Taxonomy of AI Risks

Αντίστοιχη κατηγοριοποίηση προκύπτει και από τον τομέα όπου εντοπίζεται ο εκάστοτε κίνδυνος. Στο εργαλείο παρουσιάζονται 7 διαφορετικοί κύριοι τομείς και 24 υποκατηγορίες:

- Discrimination & Toxicity
    - Unfair discrimination and misrepresentation
    - Exposure to toxic content

    - Unequal performance across groups
- Privacy & Security
    - Compromise of privacy by obtaining, leaking or correctly inferring sensitive information
    - AI system security vulnerabilities and attacks
- Misinformation
    - False or misleading information
    - Pollution of information ecosystem and loss of consensus reality
- Malicious actors & Misuse
    - Disinformation, surveillance, and influence at scale
    - Cyberattacks, weapon development or use, and mass harm
    - Fraud, scams, and targeted manipulation
- Human-Computer Interaction
    - Overreliance and unsafe use
    - Loss of human agency and autonomy
- Socioeconomic & Environmental Harms
    - Power centralization and unfair distribution of benefits
    - Increased inequality and decline in employment quality
    - Economic and cultural devaluation of human effort
    - Competitive dynamics
    - Governance failure
    - Environmental harm
- AI system safety, failures, and limitations
    - AI pursuing its own goals in conflict with human goals or values
    - AI possessing dangerous capabilities
    - Lack of capability or robustness
    - Lack of transparency or interpretability
    - AI welfare and rights
    - Multi-agent risks

Όπως απεικονίζεται και στις κατηγορίες, το εργαλείο δεν αποτυπώνει αποκλειστικά κινδύνους ασφάλειας και ιδιωτικότητας. Καλύπτει ένα μεγάλο φάσμα κινδύνων σε διάφορους τομείς, όπως φαίνεται και στην παραπάνω λίστα, από κυβερνοασφάλεια μέχρι Κοινωνικοοικονομικούς και περιβαλλοντικούς κινδύνους.


Get a quick preview of how we group risks by domain in our database. Search for one of the domain/subdomain names (eg 'fraud') to see all risks categorized against that domain. For more detailed filtering and to freely download the data, explore the full database.

Cyberattacks

QuickRef	Risk category	Risk subcategory	Description	Domain	Subdomain
Critch2023	Type 5: Criminal weaponization		One or more criminal entities could create AI to intentionally inflict harms, such as for terrorism or combating law enforcement.	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm

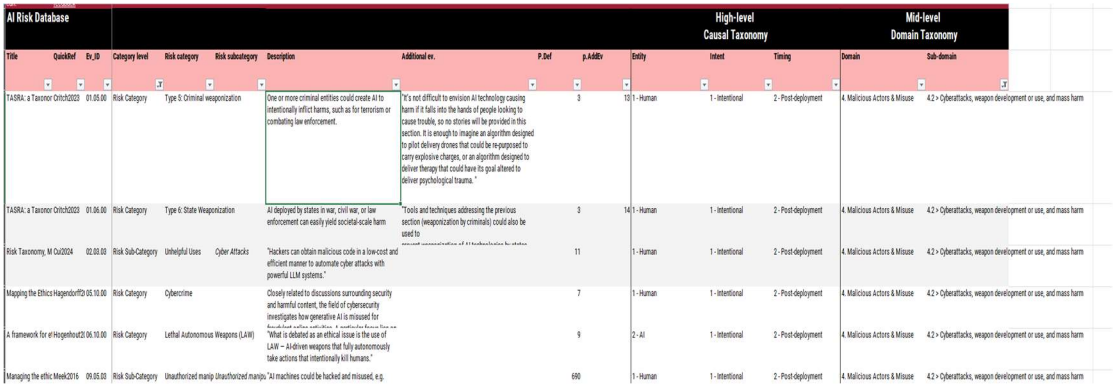
1 / 48

Εικόνα 16: Domain Taxonomy

5.3.3 Κεντρική Βάση Κινδύνων

Η κύρια βάση είναι ή πληροφορία που έχει περιγραφεί συγκεντρωμένη σε ένα αρχείο. Συγκεκριμένα, αποτυπώνονται

1. και οι δυο ταξινομήσεις (Causual, Domain)
2. το άρθρο ή η πηγή από την οποία προέκυψε ο κίνδυνος,
3. περιγραφή του κινδύνου
4. Επιπλέον τεκμηρίωση για την εμφάνιση του κινδύνου
5. Η κατηγοριοποίηση του κινδύνου βάσει των ταξινομήσεων ώστε να του αποδοθούν οι σωστές ετικέτες και να μπορεί να εντοπιστεί.


AI Risk Database

									High-level Causal Taxonomy			Mid-level Domain Taxonomy	
Title	QuickRef	Ev_ID	Category level	Risk category	Risk subcategory	Description	Additional ev.	P.Def	p.AddEv	Entity	Intent	Timing	Domain	Sub-domain
TASRA: a Taxonor	Critch2023	01.05.00	Risk Category	Type 5: Criminal weaponization		One or more criminal entities could create AI to intentionally inflict harms, such as for terrorism or combating law enforcement.	"It's not difficult to envision AI technology causing harm if it falls into the hands of people looking to cause trouble, so no stories will be provided in this section. It is enough to imagine an algorithm designed to pilot delivery drones that could be re-purposed to carry explosive charges, or an algorithm designed to deliver therapy that could have its goal altered to deliver psychological trauma."	8	10	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
TASRA: a Taxonor	Critch2023	01.06.00	Risk Category	Type 6: State Weaponization		AI deployed by states in war, civil war, or law enforcement can easily yield societal-scale harm	"Tools and techniques addressing the previous section (weaponization by criminals) could also be used to	8	14	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
Risk Taxonomy, M	Cui2024	02.00.00	Risk Sub-Category	Unhelpful Uses	Cyber Attacks	"Hackers can obtain malicious code in a low-cost and efficient manner to automate cyber attacks with powerful LLM systems."		11		1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
Mapping the Ethics	Hagendorff20	05.10.00	Risk Category	Cybercrime		Closely related to discussions surrounding security and harmful content, the field of cybersecurity investigates how generative AI is misused for		7		1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
A framework for e	Hogenhout21	06.10.00	Risk Category	Lethal Autonomous Weapons (LAW)		"What is debated as an ethical issue is the use of LAW – AI-driven weapons that fully autonomously take actions that intentionally kill humans."		9		2 - AI	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
Managing the ethic	Meek2016	09.05.00	Risk Sub-Category	Unauthorized manip	Unauthorized manipu	"AI machines could be hacked and misused, e.g.		690		1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm

Εικόνα 17: Αποτύπωση κινδύνων

5.3.4 Διαχείριση κινδύνων

Το εργαλείο πέρα από τους ίδιους τους κινδύνους παρουσιάζει έναν επιπλέον συγκεντρωτικό πίνακα στον οποίο αποτυπώνονται διορθωτικές ενέργειες για την

διαχείριση τους. Μέσω του AI Risk Mitigation Taxonomy, ο χρήστης μπορεί να διερευνήσει πιθανές λύσεις για τους κινδύνους που διέπουν την τεχνητή Νοημοσύνη μέσω τεσσάρων (4) κεντρικών κατηγοριών και εικοσιτριών (23) υποκατηγοριών όπως αποτυπώνονται στον πίνακα.

1. Governance & Oversight Controls	2. Technical & Security Controls	3. Operational Process Controls	4. Transparency & Accountability Controls
1.1 Board Structure & Oversight	2.1 Model & Infrastructure Security	3.1 Testing & Auditing	4.1 System Documentation
1.2 Risk Management	2.2 Model Alignment	3.2 Data Governance	4.2 Risk Disclosure
1.3 Conflict of Interest Protections	2.3 Model Safety Engineering	3.3 Access Management	4.3 Incident Reporting
1.4 Whistleblower Reporting & Protection	2.4 Content Safety Controls	3.4 Staged Deployment	4.4 Governance Disclosure
1.5 Safety Decision Frameworks		3.5 Post-deployment Monitoring	4.5 Third-Party System Access
1.6 Environmental Impact Management		3.6 Incident Response & Recovery	4.6 User Rights & Recourse
1.7 Societal Impact Assessment			

Εικόνα 18: Mitigation Categories

Αντίστοιχα με την μεθοδολογία που παρουσιάστηκε στη ταξινόμηση των κινδύνων, έτσι και στην περίπτωση των διορθωτικών ενεργειών, έχει υλοποιηθεί αντίστοιχη μελέτη άρθρων και πληροφοριών όπως αποτυπώνεται και στην εκάστοτε διορθωτική ενέργεια.

	Action ID	Action Name	Action Definition	Action Source	MitigationCode
559	A1033_Uuk2024	Prohibiting high-stakes applications	Enforcing use policies that prohibit high-stakes applications. Requires Know-Your-Customer procedures.	Uuk2024	3.3 Access Management
	MITIGATION TAXONOMY (SYNCED) 3.4 Staged Deployment 8				
560	A0024_Bengio2025	Responsible Release and Deployment Strategies	There is a spectrum of release and deployment strategies for AI including staged releases, cloud- based or API access, deployment safety controls, and acceptable use policies. There are some eme...	Bengio2025	3.4 Staged Deployment
561	A0211_Schuett2023	Staged deployment	AGI labs should deploy powerful models in stages. They should start with a small number of applications and fewer users, gradually scaling up as confidence in the model's safety increases.	Schuett2023	3.4 Staged Deployment
562	A0216_Schuett2023	Treat updates similarly to new models	AGI labs should treat significant updates to a deployed model (e.g. additional fine-tuning) similarly to its initial development and deployment. In particular, they should repeat the pre-deployment ...	Schuett2023	3.4 Staged Deployment
563	A0219_Schuett2023	Internal deployments = external deployments	AGI labs should treat internal deployments (e.g. using models for writing code) similarly to external deployments. In particular, they should perform a pre-deployment risk assessment.	Schuett2023	3.4 Staged Deployment
564	A0779_UK Government2...	Reversible and Small-Scale Deployment	Deploy models in small-scale or reversible ways before deploying models in large-scale or irreversible ways. This makes it possible for frontier AI organisations to notice and mitigate harm before t...	UK Government2023	3.4 Staged Deployment
565	A0968_Gipiškis2024	Staged release of model weights	When a model is developed by a provider for use in a certain AI system, it may also be useful to release the model itself more widely. Such developers can follow a staged release approach, in whic...	Gipiškis2024	3.4 Staged Deployment
566	A0969_Gipiškis2024	Gradual or incremental monitored release of m...	AI systems can be released for access incrementally, starting with a small and selected deployer base before progressively being released to a wider user base. Initially, usage to a hosted API can b...	Gipiškis2024	3.4 Staged Deployment
567	A1023_Uuk2024	Deploying powerful models in stages	Starting with a small number of applications and fewer users, and gradually scaling up API-access as rigorous monitoring increases confidence in the model's safety. An API-mediated staged releas...	Uuk2024	3.4 Staged Deployment

Εικόνα 19: Mitigation Taxonomy

Οι εμπλεκόμενοι μπορούν να ερευνήσουν τους πίνακες που παρουσιάζονται και να ενημερωθούν για κινδύνους τους οποίους, πιθανώς, να μην είχαν λάβει υπόψιν τους, ενώ ταυτόχρονα, υπάρχει η δυνατότητα σύνδεσής τους με αντίστοιχες ενέργειες μετριασμού για την ορθή διαχείρισή του.

6. Συμπεράσματα

Η παρούσα διπλωματική εργασία είχε ως σκοπό να αναλύσει τη διαχείριση κινδύνων ασφάλειας και ιδιωτικότητας σε συστήματα Τεχνητής Νοημοσύνης και να παρουσιάσει το υφιστάμενο «τοπίο», αναδεικνύοντας τη σύνθετη φύση των απειλών που προκύπτουν από την υιοθέτηση και τη λειτουργία τέτοιων συστημάτων. Μέσα από τη παρουσίαση των κινδύνων και της διαδικασίας αξιολόγησης και αντιμετώπισής τους, κατέστη σαφές ότι η Τεχνητή Νοημοσύνη δεν αποτελεί απλώς μια εξέλιξη των πληροφοριακών συστημάτων, αλλά ένα εξολοκλήρου νέο τεχνολογικό πεδίο που εισάγει νέες μορφές αβεβαιότητας, πολυπλοκότητας και απειλών αλλά και ευκαιρίες για την δημιουργία εργαλείων και νέων πλαισίων ώστε να διευκολυνθεί η διαχείρισή τους και η ενσωμάτωσή τους στην Κοινωνία.

Η ανάλυση των κινδύνων ασφάλειας ανέδειξε ότι τα συστήματα αυτά είναι ευάλωτα σε επιθέσεις καθ' όλο τον κύκλο ζωής τους, από τη φάση του σχεδιασμού και της ανάπτυξης έως και τον χρόνο εκτέλεσης και λειτουργίας τους. Επιθέσεις όπως οι adversarial, inference, model extraction και prompt injection καταδεικνύουν ότι η ασφάλεια του AI δεν περιορίζεται μόνο στην προστασία της υποδομής και στην «παραδοσιακή» διαχείριση της ασφάλειας πληροφοριών, αλλά επεκτείνεται στο ίδιο το μοντέλο, στα δεδομένα που διαχειρίζεται και το τροφοδοτούν και στη λογική λήψης αποφάσεων. Αντίστοιχα, η εξέταση των κινδύνων ιδιωτικότητας ανέδειξε ότι η χρήση της τεχνολογίας μπορεί να οδηγήσει όχι μόνο σε άμεσες παραβιάσεις προσωπικών δεδομένων, αλλά και στη δημιουργία νέων, έμμεσων πληροφοριών μέσω συμπερασμάτων, γεγονός που εντείνει τον κίνδυνο απώλειας ανωνυμίας και ιδιωτικότητας.

Για να μετριαστούν οι κίνδυνοι που προκύπτουν, ιδιαίτερη σημασία αποδίδεται στη διαδικασία διαχείρισης κινδύνων, βασισμένη σε διεθνώς αναγνωρισμένα πρότυπα όπως το ISO/IEC 27005, και το NIST RMF, καθώς και η προσαρμογή της διαδικασίας αυτής στις ιδιαιτερότητες της τεχνολογίας μέσω αντίστοιχων εξειδικευμένων προτύπων όπως το ISO/IEC 42001, το ISO 23894 και το NIST AI RMF . Η ανάλυση αυτών αναφέρει ότι τα παραδοσιακά πλαίσια διαχείρισης κινδύνων παραμένουν θεμελιώδη, ωστόσο απαιτούν εμπλουτισμό με AI-specific πρακτικές, όπως η συνεχής αξιολόγηση, η ενισχυμένη διακυβέρνηση δεδομένων και οι μηχανισμοί ανθρώπινης εποπτείας. Παράλληλα, η εξέταση του κανονιστικού πλαισίου για την Τεχνητή Νοημοσύνη (όπως το AI Act) ανέδειξε τη σημασία της προσέγγισης βάσει κινδύνου

και της ενσωμάτωσης της συμμόρφωσης ήδη από το στάδιο του σχεδιασμού. Η κατηγοριοποίηση των συστημάτων Τεχνητής Νοημοσύνης για την προστασία των δικαιωμάτων των φυσικών προσώπων και η επιβολή εφαρμογής διαχειριστικών μέτρων μέσω ρυθμιστικών πλαισίων αποτελεί ζήτημα υψίστης σημασίας για τον περιορισμό και την ομαλή ενσωμάτωση στην καθημερινότητα.

Σημαντική συμβολή στην διαχείριση κινδύνων αποτελεί η παρουσίαση και η χρήση πλαισίων αποτύπωσης και ταξινόμησης κινδύνων ΤΝ, όπως το Plot4AI, το MITRE ATLAS, και το AI Risk Depository του MIT, μέσω των οποίων οι Οργανισμοί και οι χρήστες έχουν την δυνατότητα να διερευνήσουν συγκεντρωτικά το τοπίο διαχείρισης κινδύνων. Τα πλαίσια αυτά λειτουργούν συμπληρωματικά προς τα κανονιστικά και οργανωτικά πρότυπα, παρέχοντας πρακτικά εργαλεία για την αναγνώριση, την κατηγοριοποίηση και την κατανόηση σύνθετων σεναρίων κινδύνων.

Εν τέλει, προκύπτει πως η αποτελεσματική διαχείριση κινδύνων στην Τεχνητή Νοημοσύνη προϋποθέτει μια ολιστική προσέγγιση που συνδυάζει τεχνική, οργανωτική και κανονιστική διαχείριση. Η αποσπασματική εφαρμογή μέτρων ασφάλειας ή συμμόρφωσης δεν επαρκεί για την αντιμετώπιση των δυναμικών και διαρκώς εξελισσόμενων κινδύνων της τεχνολογίας αυτής. Αντίθετα, απαιτείται η ενσωμάτωση ενός ατέρμονου κύκλου αξιολόγησης, παρακολούθησης και βελτίωσης, ευθυγραμμισμένος με τον κύκλο ζωής του συστήματος ώστε να είναι εφικτός ο έγκαιρος εντοπισμός των ευπαθειών και των απειλών αλλά και ο μετριασμός τους.

Ολοκληρώνοντας, σε επίπεδο μελλοντικών προκλήσεων, η αυξανόμενη υιοθέτηση μοντέλων και συστημάτων τεχνητής νοημοσύνης θα εντείνει τους κινδύνους ασφάλειας αλλά και ιδιωτικότητας, καθιστώντας την προστασία των φυσικών προσώπων, την πρόβλεψη των πιθανών αποκλίσεων στην διαχείριση των δεδομένων και τον περιορισμό των κινδύνων ασφάλειας ιδιαίτερα δύσκολη. Παράλληλα, η ταχεία εξέλιξη της τεχνολογίας ενδέχεται να υπερβεί τον ρυθμό προσαρμογής των κανονιστικών πλαισίων, δημιουργώντας νέα κενά διακυβέρνησης και επιβολής των νομικών απαιτήσεων για την προστασία των υποκειμένων, τον Οργανισμών αλλά και της διακινούμενης πληροφορίας. Στο πλαίσιο αυτό, η εναρμόνιση με διεθνή πρότυπα, η εξέλιξη των εξειδικευμένων μεθοδολογιών διαχείρισης κινδύνων για το AI και η ενίσχυση της διαφάνειας και της λογοδοσίας αποτελούν κρίσιμα αντικείμενα μελλοντικής έρευνας.

Συμπερασματικά, γίνεται αντιληπτό πως η Τεχνητή Νοημοσύνη μπορεί να αξιοποιηθεί με ασφάλεια και σεβασμό στην ιδιωτικότητα μόνο εφόσον ενταχθεί σε ένα πλαίσιο διαχείρισης κινδύνων που να μπορεί να αναγνωρίσει τα ρίσκα που συνοδεύουν την χρήση της τεχνολογίας αλλά και να αναγνωρίσει τις υφιστάμενες λύσεις και τις ενέργειες μετριασμού που έχουν στην διάθεσή τους οι χρήστες. Η συστηματική κατανόηση των κινδύνων, η αξιοποίηση κατάλληλων πλαισίων εφαρμογής και η ενσωμάτωση της διαχείρισης κινδύνων στον πυρήνα της διακυβέρνησης της Τεχνητής Νοημοσύνης συνιστούν βασικές προϋποθέσεις για τη βιώσιμη, υπεύθυνη ανάπτυξή και την ομαλή ένταξη της τεχνολογίας αυτής στην κοινωνία.

7. Βιβλιογραφία

1. **Al-Kharusi, Younis, et al.** *Open-Source Artificial Intelligence Privacy and Security: A Review.* Vol. 13, no. 13, 2024, p. 311, https://doi.org/10.3390/computers13120311. Accessed 13 Apr. 2025.
2. **Arokun, Esther.** "COMPLEXITIES of AI TRENDS: THREATS to DATA PRIVACY LEGAL COMPLIANCE." *SSRN*, 4 Oct. 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4943466. Accessed 13 Apr. 2025.
3. **Aspen Institute. (2023).** *Generative A.I. regulation and cybersecurity: A global view of policymaking*. Aspen Digital. https://www.aspeninstitute.org/. Accessed 03 Jun. 2025
4. **Badhan, Chandra, et al.** "Security and Privacy Challenges of Large Language Models: A Survey." Security and Privacy Challenges of Large Language Models: A Survey, vol. 1, 30 Jan. 2024, arxiv.org/abs/2402.00888, https://doi.org/10.48550/arXiv.2402.00888. Accessed 31 May 2025..
5. **Bengio, Yoshua, et al.** "Managing Extreme AI Risks amid Rapid Progress." *Science*, vol. 384, no. 6698, 24 May 2024, pp. 842–845, https://doi.org/10.1126/science.adn0117. Accessed 13 Apr. 2025.
6. **Butcher, James, and Irakli Beridze.** "What Is the State of Artificial Intelligence Governance Globally?" The RUSI Journal, vol. 164, no. 5-6, 19 Sept. 2019, pp. 88–96, https://doi.org/10.1080/03071847.2019.1694260. Accessed 3 June 2025.
7. **Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., … Raffel, C.** (2021, August). Extracting Training Data from Large Language Models. 30th USENIX Security Symposium (USENIX Security 21), 2633–2650. Retrieved from https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting, Accessed 28 Sept. 2025.
8. **Chesney, R., & Citron, D.** (2019). *Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review*, 107(6), 1753–1820.
9. **Copeland, B. J.** (2024). *Artificial intelligence.* In *Encyclopedia Britannica*. https://www.britannica.com/technology/artificial-intelligence. Accessed 23 Apr. 2025.
10. **Cotton, D. R. E., Cotton, P. A., & Shipway, J. R.** (2023). *Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International*, 1–12. https://doi.org/10.1080/14703297.2023.2190148

11. **CSA Singapore.** "COMPANION GUIDE on SECURING AI SYSTEMS." 15 Oct. 2024.
12. **"Cyber Security Risks to Artificial Intelligence."** *GOV.UK*, 15 May 2024, www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/cyber-security-risks-to-artificial-intelligence#contents. Accessed 13 Apr. 2025.
13. **Das, B. C., Amini, M. H., & Wu, Y.** (30 June 2024). Security and privacy challenges of large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 1(1), 1–34. https://doi.org/10.48550/arXiv.2402.00888**,** Accessed 26 Sept. 2025.
14. **Dataguard.** "The Growing Data Privacy Concerns with AI: What You Need to Know." *Dataguard*, 4 Sept. 2024, www.dataguard.com/blog/growing-data-privacy-concerns-ai/. Accessed 13 Apr. 2025.
15. **Elbaih, Mohamed.** "THE ROLE of PRIVACY REGULATIONS in AI DEVELOPMENT." *SSRN*, 30 Oct. 2023.
16. **European Commission.** (2021). *Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
17. **Farzad, Nourmohammadzadeh**, et al. "Large Language Models in Cybersecurity: State-of-The-Art." 30 Jan. 2024, Accessed 31 May 2025
18. **Floridi, Luciano & Chiriatti, Massimo**. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds and Machines. 30. 1-14. 10.1007/s11023-020-09548-1. Accessed 31 May. 2025
19. **Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Goldstein, T.** (2021). Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2012.10544, Accessed 25 September. 2025
20. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). *Deep Learning.* MIT Press.
21. **Hakeemat, Ijaiya.** "Harnessing AI for Data Privacy: Examining Risks, Opportunities and Strategic Future Directions." *International Journal of Science and Research Archive*, vol. 13, no. 2, 30 Dec. 2024, pp. 2878–2892, https://doi.org/10.30574/ijsra.2024.13.2.2510. Accessed 13 Apr. 2025.
22. **Barberá, Isabel.** "PLOT4AI - About." Plot4.Ai, 2022, https://plot4.ai/about. Accessed 11 Jan. 2026.

23. **ISO** (2022). *ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection — Guidance on managing information security risks*. ISO. Accessed 27 December 2025.
24. **ISO** (2023). *ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on Risk Management*. ISO. Accessed 06 January 2026.
25. **ISO** (2023). *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system*. ISO. Accessed 06 January 2026.
26. **Kumar, Anil, et al.** "Advances in Data Protection and Artificial Intelligence: Trends and Challenges." *International Journal of Advanced Engineering Technologies and Innovations*, vol. 01, no. 1, 2023, p. 1. Accessed 13 Apr. 2025.
27. **Kumar , Anil , and Srikanth Suryadevara**. "Emerging Frontiers: Data Protection Challenges and Innovations in Artificial Intelligence." Jan. 2024.
28. **Lee, Hao-Ping (Hank), et al.** "Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks." *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 11 May 2024, pp. 1–19, https://doi.org/10.1145/3613904.3642116. Accessed 13 Apr. 2025.
29. **Lupicinio.** "Artificial Intelligence and Data Protection: Regulatory Challenges | Lupicinio." *Lupicinio*, 3 Mar. 2025, https://lupicinio.com/en/artificial-intelligence-and-data-protection-regulatory-challenges/. Accessed 13 Apr. 2025.
30. **Marengo, Federico.** *PROTECTING INDIVIDUALS in the AGE of AI.* 1st ed., 16 Aug. 2023, pp. 14–18. Accessed 13 Apr. 2025.
31. **Mathew, Alex.** "Multidisciplinary International Journal of Research and Development 84 All Rights Are Reserved by Www.mijrd.com AI TRiSM: Trust, Risk, and Security Management in Cybersecurity." *MIJRD*, 11 Mar. 2025.
32. **McCormack, J., Gifford, T., & Hutchings, P.** (2019). *Autonomy, Authenticity, Authorship and Intention in Computer Generated Art.* In *The Oxford Handbook of Algorithmic Music*.
33. **Microsoft.** (2023). *Responsible AI Standard: Putting our AI principles into practice.* https://www.microsoft.com/en-us/ai/responsible-ai
34. **MITRE**. "MITRE Adversarial Threat Landscape for AI Systems (ATLAS)." 2023, Accessed 11 Jan. 2026.
35. **NIST**. "NIST Risk Management Framework." NIST, 24 Sept. 2024, csrc.nist.gov/projects/risk-management/about-rmf. Accessed 4 Jan. 2026.

36. **OECD Publishing.** "AI, DATA GOVERNANCE and PRIVACY SYNERGIES and AREAS of INTERNATIONAL CO-OPERATION OECD ARTIFICIAL INTELLIGENCE PAPERS." June 2024, https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/ai-data-governance-and-privacy_2ac13a42/2476b1a4-en.pdf. Accessed 19 May 2025
37. **OECD Publishing**. "COMMON GUIDEPOSTS to PROMOTE NTEROPERABILITY in AI RISK MANAGEMENT." Nov. 2023, https://www.oecd.org/en/publications/common-guideposts-to-promote-interoperability-in-ai-risk-management_ba602d18-en.html. Accessed 06 January 2026.
38. **OpenCV.** (2023). *What Is Computer Vision in 2024? A Beginners Guide.* https://opencv.org/blog/what-is-computer-vision/. Accessed 19 May 2025.
39. **OVIC.** "Artificial Intelligence and Privacy - Issues and Challenges." *Office of the Victorian Information Commissioner*, 2018, https://ovic.vic.gov.au/privacy/resources-for-organisations/artificial-intelligence-and-privacy-issues-and-challenges/. Accessed 13 Apr. 2025.
40. **OWASP.** "AI Security Overview." *Owaspai.org*, 2025, https://owaspai.org/docs/ai_security_overview/#ai-security-matrix. Accessed 31 May 2025.
41. **Cyber Security Agency of Singapore.** (2024, July). Companion guide on securing AI systems. Retrieved from https://www.csa.gov.sg, Accessed 25 Sept. 2025.
42. **Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A.** (19 March 2017). *Practical Black-Box Attacks against Machine Learning*. Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 506–519, https://doi.org/10.48550/arXiv.1602.02697, Accessed 26 Sept 2025
43. **Pearce, Will, et al.** "AI Security Risk Assessment Best Practices and Guidance to Secure AI Systems." *Microsoft*, 19 Mar. 2024.
44. **Rassin, E., & Willems, C.** (2023). *Copyright challenges in the age of generative AI. AI & Society*, 1–9.
45. **Shahriar, Sakib, et al.** "A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle." *IEEE Access*, vol. 11, 19 June 2023, pp. 61829–61854, https://doi.org/10.1109/access.2023.3287195. Accessed 13 Apr. 2025.
46. **Shokri, R., Stronati, M., Song, C., & Shmatikov, V.** (2017). Membership inference attacks against machine learning models. 2017 IEEE Symposium on Security and Privacy (SP), 3–18. https://doi.org/10.1109/SP.2017.41, Accessed 26 Sept. 2025

47. **Singh, K.** (2024). *Building Blocks for AI Part 4: Types of Neural Networks.* Medium. https://medium.com/@kamalmeet/building-blocks-for-ai-part-4-types-of-neural-networks-1fe1ed234132. Accessed 14 May 2025.
48. **Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N.** (2024). *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence*. https://doi.org/10.48550/arXiv.2408.12622, Accessed 12 Jan 2026.
49. **SNOWFLAKE**. "AI Security Framework." Snowflake, 2025, www.snowflake.com/en/resources/white-paper/snowflake-ai-security-framework/. Accessed 24 Sept. 2025.
50. **Solove, Daniel.** "A Regulatory Roadmap to AI and Privacy." *IAPP News*, 24 Apr. 2024. SSRN.
51. **Tabassi, Elham.** Artificial Intelligence Risk Management Framework (AI RMF 1.0). 26 Jan. 2023, www.nist.gov/itl/ai-risk-management-framework, https://doi.org/10.6028/nist.ai.100-1. Accessed 5 Jan. 2026.
52. **Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T.** (2016). *Stealing Machine Learning Models via Prediction APIs*. 25th USENIX Security Symposium, 601–618. Accessed 24 Sept. 2025.
53. **Vassilev A, Oprea A, Fordyce A, Anderson H** (2024) Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Artifcial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2023. https://doi.org/10.6028/NIST.AI.100-2e2023, Accessed 31 May 2025
54. **Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., … Gabriel, I.** (2021). Ethical and social risks of harm from Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2112.04359, Accessed 28 Sept. 2025
55. **Yeung, Louis .** "Guidance for the Development of AI Risk and Impact Assessments." July 2021, https://cltc.berkeley.edu/ai-risk-and-impact/, Accessed 4 June 2025
56. **Zhang, C., and Lu, Y.** (2021). *Study on Artificial Intelligence: The State of the Art and Future Prospects. Journal of Industrial Information Integration*, 23(23), p.100224. https://doi.org/10.1016/j.jii.2021.100224.

8. Παράρτημα Α. Συγκριτικός Πίνακας Πλαισίων Αξιολόγησης Κινδύνων για Συστήματα Τεχνητής Νοημοσύνης

Framework	Σκοπός / Πεδίο Εφαρμογής	Αξιολόγηση Κινδύνου	Κάλυψη Ασφάλειας	Κάλυψη Ιδιωτικότητας	Κάλυψη Κύκλου Ζωής Συστημάτων	Νομική Υποχρέωση
ISO/IEC 27005	Asset–Threat–Vulnerability–Impact	Ισχυρή	Ισχυρή	Μέτρια προς Ισχυρή	Κυρίως σε αρχικά στάδια και επαναλαμβανόμενη ανά διαστήματα	Όχι
NIST RMF	Lifecycle-based risk management	Ισχυρή και Συνεχής	Ισχυρή	Ισχυρή	Διαρκώς επαναλαμβανόμενη διαδικασία	Όχι
NIST AI Risk Management Framework (AI RMF)	Διαχείριση κινδύνων ΤΝ σε οργανισμούς	Ισχυρή και Συνεχής	Ισχυρή	Ισχυρή	Διαρκώς επαναλαμβανόμενη διαδικασία	Όχι
EU AI Act (Risk-based Framework)	Ρύθμιση ΤΝ με βάση επίπεδα κινδύνου	Χαμηλή, αφοσιώνεται στην κατηγοριοποίηση συστημάτων	Χαμηλή	Χαμηλή	Σε αρχικά στάδια	Ναι
ISO/IEC 23894 (AI Risk Management)	Τυποποίηση διαχείρισης κινδύνων ΤΝ	Ισχυρή	Ισχυρή	Ισχυρή	Κυρίως σε αρχικά στάδια και επαναλαμβανόμενη ανά διαστήματα	Όχι

Πίνακας 1: Πλαίσια Αξιολόγησης Κινδύνων