ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τεχνητή Νοημοσύνη και Κυβερνοασφάλεια σε Εφαρμογές Άμυνας και Ασφάλειας: Απειλές, Λύσεις και Στρατηγικές Επιπτώσεις

Αριστείδης Άγγελος Ζουμπάκης
A.M. cscyb22009

Εισηγητής: Στέφανος Γκρίτζαλης

ΜΑΪΟΣ 2024

(Κενό φύλλο)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΚΥΒΕΡΝΟΑΣΦΑΛΕΙΑ ΣΕ ΕΦΑΡΜΟΓΕΣ ΆΜΥΝΑΣ ΚΑΙ ΑΣΦΑΛΕΙΑΣ: ΑΠΕΙΛΕΣ, ΛΥΣΕΙΣ ΚΑΙ ΣΤΡΑΤΗΓΙΚΕΣ ΕΠΙΠΤΩΣΕΙΣ

Αριστείδης Άγγελος Ζουμπάκης
A.M. cscyb22009

Εισηγητής:

Στέφανος Γκρίτζαλης, Καθηγητής

STEFANOS GKRITZALIS
18/05/2024 11:33

Εξεταστική Επιτροπή:

Παναγιώτης Γιαννακόπουλος, Καθηγητής

Panagiotis Giannakopoulos
Digitally signed by Panagiotis Giannakopoulos
Date: 2024.05.18 19:02:59 +03'00'

Εμμανουήλ Θ. Μιχαηλίδης, Εντεταλμένος Διδάσκων

EMMANOUIL MICHAILIDIS
Digitally signed by EMMANOUIL MICHAILIDIS
Date: 2024.05.19 10:00:30 +03'00'

Ημερομηνία εξέτασης: 02/05/2024

(Κενό φύλλο)

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος **Αριστείδης Άγγελος Ζουμπάκης** του **Γεωργίου**, με αριθμό μητρώου **cscyb22009** φοιτητής του Προγράμματος Μεταπτυχιακών Σπουδών «Κυβερνοασφάλεια» του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Βεβαιώνω ότι είμαι συγγραφέας της παρούσας διπλωματικής εργασίας και ότι έχω αναφέρει ή παραπέμψει σε αυτή, ρητά και συγκεκριμένα, όλες τις πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, προτάσεων ή λέξεων, είτε αυτές μεταφέρονται επακριβώς (στο πρωτότυπο ή μεταφρασμένες) είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για την συγκεκριμένη διπλωματική εργασία».

Ο Δηλών

(Κενό φύλλο)

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία ολοκληρώθηκε μετά από επίμονες προσπάθειες, σε ένα ενδιαφέρον γνωστικό αντικείμενο, όπως αυτό της τεχνητής νοημοσύνης και της κυβερνοασφάλειας. Την προσπάθειά μου αυτή ενέπνευσε και υποστήριξε ο επιβλέπων καθηγητής μου, τον οποίο θα ήθελα να ευχαριστήσω.

Ακόμα θα ήθελα να ευχαριστήσω την οικογένειά μου για την υποστήριξη και τη συμπαράσταση κατά τη διάρκεια των σπουδών μου.

(Κενό φύλλο)

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία ασχολείται με την αλληλεπίδραση τεχνητής νοημοσύνης και κυβερνοασφάλειας με έμφαση στις εφαρμογές σε περιβάλλοντα άμυνας και ασφάλειας. Η εργασία επικεντρώθηκε στην κατανόηση των ιδιαίτερων απειλών που τίθενται στα στρατιωτικά συστήματα και στα συστήματα ασφαλείας που βασίζονται στην τεχνητή νοημοσύνη, στον εντοπισμό των προσφερόμενων λύσεων και στην ανάλυση των ευρύτερων στρατηγικών επιπτώσεων για την εθνική άμυνα και ασφάλεια από τη διάχυση της τεχνητής νοημοσύνης.

Για την ανάλυση επικινδυνότητας ενός συστήματος τεχνητής νοημοσύνης καθοριστική σημασία έχει η εξέταση του κύκλου ζωής του και η λήψη κατάλληλων αντιμέτρων σε όλα τα επιμέρους στάδια. Η χρήση συστημάτων τεχνητής νοημοσύνης σε περιβάλλοντα άμυνας και ασφάλειας αυξάνει την επιφάνεια επίθεσης των εν λόγω συστημάτων. Ταυτόχρονα, η τεχνητή νοημοσύνη αποτελεί χρήσιμο εργαλείο στον τομέα της κυβερνοασφάλειας τόσο σε επιθετικές όσο και σε αμυντικές επιχειρήσεις. Οι χρήσεις τεχνητής νοημοσύνης εναντίον τεχνητής νοημοσύνης, η ανάπτυξη πλαισίου διακυβέρνησης και ο έλεγχος των εξοπλισμών στον τομέα της τεχνητής νοημοσύνης αποτελούν τις σημαντικότερες προκλήσεις για το επόμενο διάστημα.

ΕΠΙΣΤΗΜΟΝΙΚΗ ΠΕΡΙΟΧΗ: Κυβερνοασφάλεια
ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Τεχνητή Νοημοσύνη, Αυτόνομα συστήματα, Κυβερνοασφάλεια

ABSTRACT

The present thesis concerns the interaction between artificial intelligence and cybersecurity with emphasis on applications in defence and security environments. The thesis focused on understanding the specific threats posed to military and security systems based on artificial intelligence, identifying the solutions offered, and analyzing the broader strategic implications for national defense and security from the diffusion of AI.

For the risk analysis of an AI system, it is crucial to consider its life cycle and to take appropriate countermeasures at all individual stages. The use of AI systems in defence and security environments increases the attack surface of these systems. At the same time, AI is a useful tool in the field of cybersecurity in both offensive and defensive operations. The uses of AI vs. AI, the development of a governance and arms control framework in the field of AI are the most important challenges for the immediate future.

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ.....1

1.1. Αντικείμενο Εργασίας..... 1

1.2. Μεθοδολογία..... 2

1.3 Δομή Εργασίας..... 3

2. ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ 4

2.1. Βασικές Έννοιες και Χαρακτηριστικά της ΤΝ..... 4

2.2. Αλγόριθμοι Μηχανικής Μάθησης..... 8

2.3. Ο Κύκλος Ζωής της ΤΝ..... 13

3. ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΚΥΒΕΡΝΟΑΣΦΑΛΕΙΑ.....20

3.1. Κυβερνοασφάλεια της ΤΝ..... 20

3.2. ΤΝ ως εργαλείο Κυβερνοεπιθέσεων..... 27

3.3. ΤΝ ως εργαλείο Κυβερνοασφάλειας και Κυβερνοάμυνας..... 30

4. ΕΦΑΡΜΟΓΕΣ ΤΝ ΣΕ ΑΣΦΑΛΕΙΑ ΚΑΙ ΆΜΥΝΑ 35

4.1. Χρήση ΤΝ σε Περιβάλλοντα Ασφαλείας..... 35

4.2. Χρήση ΤΝ σε Στρατιωτικά Περιβάλλοντα..... 38

4.3. Ανάλυση Τοπίου Απειλών από τη Χρήση ΤΝ σε Άμυνα και Ασφάλεια..... 42

4.4. Παραδείγματα Χρήσης και Προβληματισμοί..... 47

5. ΗΘΙΚΑ ΖΗΤΗΜΑΤΑ ΚΑΙ ΡΥΘΜΙΣΤΙΚΟ ΠΛΑΙΣΙΟ 51

5.1. Ηθικά Ζητήματα και Ζητήματα Ιδιωτικότητας..... 51

5.2. Ρυθμιστικό Πλαίσιο Χρήσης ΤΝ σε Περιβάλλοντα Άμυνας και Ασφάλειας .. 55

6. ΣΤΡΑΤΗΓΙΚΕΣ ΕΠΙΠΤΩΣΕΙΣ 61

6.1. Επιπτώσεις στην Εθνική Άμυνα και Ασφάλεια..... 61

6.2. Γεωπολιτικές Επιδράσεις 65

7. ΣΥΜΠΕΡΑΣΜΑΤΑ..... 69

7.1 Ανακεφαλαίωση 69

7.2 Συμπεράσματα - Προοπτικές 71

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

Σχήμα 2.1: AI systems life cycle[13] **15**

Σχήμα 2.2: CDAC AI life cycle **16**

Σχήμα 2.3: ENISA Typical AI lifecycle [20] **17**

Σχήμα 2.4: ISO/IEC 5338 Sample AI lifecycle **18**

Σχήμα 2.5: NIST AI life cycle **19**

Σχήμα 3.1: Επιθετικές και Αμυντικές Στρατηγικές Κυβερνοασφάλειας ΤΝ..... **23**

Σχήμα 4.1: Cyber Kill Chain..... **43**

Σχήμα 4.2: Πλαίσιο κυβερνοαπειλών με βάση την ΤΝ[80]..... **45**

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

Πίνακας 3.1: Τακτικές Επίθεσης κατά ATLAS..... **23**

Πίνακας 3.2: Χρήσεις ΤΝ στον τομέα της κυβερνοασφάλειας..... **33**

Πίνακας 4.1: Cyber Kill Chain Εναντίον Αυτόνομου Αεροχήματος..... **47**

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

ΕΕ Ευρωπαϊκή Ένωση

ΚΦΣ Κυβερνο-Φυσικά Συστήματα

ΟΟΣΑ Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης

ΠΣ Πληροφοριακά Συστήματα

ΤΝ Τεχνητή Νοημοσύνη

3R Robustness, Response, Resilience

AGI Artificial General Intelligence

ALTAI Assessment List for Trustworthy AI

APT Advanced Persistent Threat

ATLAS Adversarial Threat Landscape for Artificial-Intelligence Systems

ATM Automated Teller Machine

BMC3I Battle Management, Command, Control, Communications and Intelligence

CDAC Center for Data Analytics and Cognition

DAI Distributed Artificial Intelligence

DGA Domain Generation Algorithms

EASO European Asylum Support Office

ENISA European Union Agency for Cybersecurity

EUAA European Union Agency for Asylum

GAN Generative Adversarial Network

HLEG High-Level Expert Group on artificial intelligence

IATE Interactive Terminology for Europe

IDF Israel Defence Forces

IDS Intrusion Detection System

IED Improvised Explosive Device

IoBT Internet of Battlefield Things

LLM	Large Language Model
MTL	Multi-task Learning
PCA	Principal Component Analysis
PoS	Point of Sale
SAIF	Secure AI Framework
SVM	Support Vector Machines
TSVM	Transductive Support Vector Machine
IoT	Internet of Things

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

Η τεχνητή νοημοσύνη (ΤΝ) κατέχει κεντρικό ρόλο στην εξελισσόμενη 4η βιομηχανική επανάσταση, μαζί με άλλες αναδυόμενες ψηφιακές τεχνολογίες όπως το διαδίκτυο των πραγμάτων (Internet of Things - IoT), τα μέγα-δεδομένα (Big Data), η ρομποτική, οι τρισδιάστατες εκτυπώσεις, η επαυξημένη και η εικονική πραγματικότητα και οι τεχνολογίες αλυσίδας συστοιχιών (blockchain) [1]. Η ΤΝ αφορά την ανάπτυξη συστημάτων υπολογιστών που μπορούν να μιμηθούν τις γνωστικές λειτουργίες του ανθρώπου. Οι λειτουργίες αυτές περιλαμβάνουν τη μάθηση, τη συλλογιστική, την επίλυση προβλημάτων, την αντίληψη, την κατανόηση φυσικής γλώσσας και την αναγνώριση ομιλίας.

Παρόλο που ο τομέας της τεχνητής νοημοσύνης μετρά αρκετές δεκαετίες ζωής από το μοντέλο τεχνητών νευρώνων των McCulloch και Pitts το 1943 και το τεστ μίμησης του Alan Turing το 1950 [2], η αλματώδης αύξηση των δυνατοτήτων του υλικού υπολογιστών (χρήση ισχυρών καρτών γραφικών για υπολογισμούς) αλλά και η βελτιστοποίηση των αλγορίθμων μηχανικής μάθησης, έχουν καταστήσει την ΤΝ εφαρμόσιμη στην καθημερινότητα και προσβάσιμη σε όλους τους χρήστες του διαδικτύου. Την ίδια στιγμή πληθαίνουν όλο και περισσότερο οι χρήσεις της ΤΝ σε εφαρμογές όπως εικονικοί βοηθοί, αυτόνομα οχήματα, αναγνώριση φωνής και εικόνας, ιατρικές διαγνώσεις κλπ. Καθώς η τεχνητή νοημοσύνη ενσωματώνεται ταχύτατα πλέον και σε στρατιωτικές εφαρμογές και εφαρμογές ασφαλείας, αυξάνονται οι ανησυχίες τόσο για την αξιοπιστία των εφαρμογών αυτών όσο και οι ηθικές διαστάσεις σε θέματα ιδιωτικότητας και δικαίου του πολέμου. Κατά συνέπεια η κυβερνοασφάλεια αυτών των συστημάτων καθίσταται επιτακτική ανάγκη.

1.1. Αντικείμενο Εργασίας

Η παρούσα μεταπτυχιακή διατριβή επιδιώκει να διερευνήσει τη διασύνδεση της κυβερνοασφάλειας και της τεχνητής νοημοσύνης στο πλαίσιο εφαρμογών άμυνας και ασφάλειας. Η έρευνα θα επικεντρωθεί στην κατανόηση των ιδιαίτερων απειλών που τίθενται στα στρατιωτικά συστήματα και στα συστήματα ασφαλείας που βασίζονται στην τεχνητή νοημοσύνη, στον εντοπισμό των προσφερόμενων λύσεων και στην

ανάλυση των ευρύτερων στρατηγικών επιπτώσεων για την εθνική άμυνα και ασφάλεια από τη διάχυση της τεχνητής νοημοσύνης.

1.2. Μεθοδολογία

Η παρούσα μελέτη υιοθετεί έναν ερευνητικό σχεδιασμό μεικτής μεθόδου για να διερευνήσει τη διασύνδεση της τεχνητής νοημοσύνης και της κυβερνοασφάλειας στον τομέα της άμυνας και της ασφάλειας. Στο πλαίσιο της έρευνας, χρησιμοποιήθηκαν δευτερογενή δεδομένα από δημοσιευμένες εργασίες σε ελεύθερα διαθέσιμες πηγές του διαδικτύου κυρίως της τελευταίας πενταετίας. Το υλικό περιλαμβάνει εκθέσεις για την ασφάλεια στον κυβερνοχώρο, στρατηγικά σχέδια και έγγραφα πολιτικής που περιγράφουν στρατηγικές υιοθέτησης της ΤΝ στον αμυντικό τομέα από κυβερνητικούς φορείς και διεθνείς οργανισμούς, ακαδημαϊκά περιοδικά και δημοσιεύσεις, αλλά και διεθνή πρότυπα και προδιαγραφές.

Η έρευνα βασίστηκε ως επί το πλείστο σε κείμενα στην αγγλική γλώσσα και η συλλογή των πηγών έγινε από Οκτώβριο 2023 έως Φεβρουάριο 2024 με τη χρήση των όρων αναζήτησης «ai life cycle», «ai cybersecurity», «machine learning algorithms», «artificial intelligence in battlefield and vulnerabilities». Για την απόδοση των αγγλικών όρων στην ελληνική γλώσσα χρησιμοποιήθηκε η πλατφόρμα IATE και σχετική ελληνική βιβλιογραφία [3], [4].

Επιμέρους στόχοι της εργασίας είναι:

- Η ανάλυση του τοπίου απειλών (Threat Landscape) από την εφαρμογή της ΤΝ στους τομείς της άμυνας και της ασφάλειας.
- Η διερεύνηση των τρωτών σημείων των μοντέλων ΤΝ και οι επιπτώσεις τους στα εν λόγω συστήματα.
- Η διερεύνηση του υφιστάμενου ρυθμιστικού πλαισίου, των ηθικών ζητημάτων καθώς και των ζητημάτων ιδιωτικότητας που σχετίζονται με τη χρήση ΤΝ σε εφαρμογές άμυνας και ασφάλειας.
- Η διερεύνηση του αντικτύπου της ΤΝ στην εθνική άμυνα και ασφάλεια, στις γεωπολιτικές παραμέτρους και στο εξελισσόμενο πεδίο επιχειρήσεων του κυβερνοχώρου.
- Η πρόταση βελτιώσεων για τη διασφάλιση της υπεύθυνης και ηθικής χρήσης των τεχνολογιών ΤΝ με παράλληλη διατήρηση των στόχων εθνικής ασφάλειας.

1.3 Δομή Εργασίας

Η εργασία είναι δομημένη σε 7 κεφάλαια. Το εισαγωγικό κεφάλαιο περιγράφει το πεδίο εφαρμογής, τη μεθοδολογία και τη συνολική δομή της έρευνας. Στο δεύτερο κεφάλαιο, αναλύονται οι θεμελιώδεις έννοιες της ΤΝ, οι αλγόριθμοι μηχανικής μάθησης και ο κύκλος ζωής των συστημάτων ΤΝ, παρέχοντας μια θεμελιώδη κατανόηση των συγκεκριμένων θεμάτων. Στο τρίτο κεφάλαιο διερευνάται η πολύπλευρη σχέση μεταξύ της ΤΝ και της ασφάλειας στον κυβερνοχώρο. Το τέταρτο κεφάλαιο εμβαθύνει στις πρακτικές εφαρμογές της τεχνητής νοημοσύνης στην ασφάλεια και την άμυνα, εξετάζοντας διαφορετικές περιπτώσεις χρήσης, από γενικές εφαρμογές ασφαλείας πληροφοριακών συστημάτων έως εφαρμογές στο πεδίο της μάχης. Στη συνέχεια, πραγματοποιείται μια εμπεριστατωμένη ανάλυση του τοπίου απειλών που προκύπτουν από τη χρήση της ΤΝ στην άμυνα και την ασφάλεια, παρέχοντας πληροφορίες σχετικά με πιθανούς κινδύνους και προκλήσεις. Τα ρυθμιστικά πλαίσια και θέματα δεοντολογίας έρχονται στο επίκεντρο του πέμπτου κεφαλαίου, εξετάζοντας τους υφιστάμενους κανονισμούς και τα δεοντολογικά διλήμματα που περιβάλλουν τις εφαρμογές ΤΝ. Επιπλέον εξετάζονται οι ηθικές επιπτώσεις της ανάπτυξης της ΤΝ σε ευαίσθητα και υψηλού κινδύνου περιβάλλοντα. Με την ολοκληρωμένη εξέταση των ρυθμιστικών και ηθικών πτυχών, το κεφάλαιο αυτό συμβάλλει στην ολοκληρωμένη κατανόηση της υπεύθυνης χρήσης της ΤΝ σε περιβάλλοντα άμυνας και ασφάλειας, τονίζοντας τη σημασία της ευθυγράμμισης των τεχνολογικών εξελίξεων με δεοντολογικά ζητήματα και νομικά πλαίσια. Στο έκτο κεφάλαιο εξετάζονται διεξοδικά οι στρατηγικές επιπτώσεις της ΤΝ, από τον ρόλο της στη λήψη αποφάσεων έως το ευρύτερο αντίκτυπό της στην εθνική άμυνα, την ασφάλεια και το γεωπολιτικό περιβάλλον. Το κεφάλαιο διερευνά περαιτέρω τις ευρύτερες επιπτώσεις της ΤΝ στην εθνική άμυνα και ασφάλεια, εξετάζοντας διεξοδικά τις επιπτώσεις της στον στρατηγικό σχεδιασμό, τις επιχειρησιακές δυνατότητες και τους εθνικούς αμυντικούς μηχανισμούς. Επιπλέον, εμβαθύνει στις γεωπολιτικές επιπτώσεις, διερευνώντας πώς οι εξελίξεις στην ΤΝ επηρεάζουν τη δυναμική και τις σχέσεις παγκόσμιας ισχύος. Τέλος, το καταληκτικό κεφάλαιο συνθέτει τα βασικά ευρήματα, παρέχοντας μια συνολική ανακεφαλαίωση της εργασίας και προτείνοντας θέματα για μελλοντική έρευνα και συστάσεις πολιτικής.

ΚΕΦΑΛΑΙΟ 2

ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Η ανάπτυξη της τεχνητής νοημοσύνης παρουσιάζει μια σειρά από παράδοξα. Ενώ στοχεύει στην αύξηση της αυτοματοποίησης, απαιτεί ταυτόχρονα στενή ανθρώπινη παρακολούθηση για τον έλεγχο του παραγόμενου αποτελέσματος. Επιπλέον αποτελεί μια τεχνολογία διπλής χρήσης καθώς μια εφαρμογή που έχει δημιουργηθεί με ΤΝ μπορεί να χρησιμοποιηθεί τόσο για το δημόσιο καλό όσο και για να προκαλέσει βλάβη [5]. Το ίδιο ισχύει και στη σχέση ΤΝ και κυβερνοασφάλειας. Ενώ η ΤΝ δημιουργεί νέες δυνατότητες για την αυτοματοποίηση και τη βελτιστοποίηση των εργασιών κυβερνοασφάλειας, ταυτόχρονα δημιουργεί νέες απειλές όταν χρησιμοποιείται από κακόβουλους δρώντες. Επιπρόσθετα, η ανάπτυξη συστημάτων ΤΝ δημιουργεί νέες απαιτήσεις κυβερνοασφάλειας και διευρύνει την επιφάνεια έκθεσης (attack surface) των πληροφοριακών συστημάτων σε πιθανές επιθέσεις. Για την καλύτερη κατανόηση των αναδυόμενων προκλήσεων, αναλύονται στη συνέχεια τα βασικά στοιχεία της ΤΝ και η σχέση της με την κυβερνοασφάλεια.

2.1. Βασικές Έννοιες και Χαρακτηριστικά της ΤΝ

Όπως συμβαίνει με τις περισσότερες αναδυόμενες τεχνολογίες, δεν υφίσταται ένας κοινά αποδεκτός ορισμός για την τεχνητή νοημοσύνη. Η Ευρωπαϊκή Επιτροπή έχει εντοπίσει τα εξής κοινά χαρακτηριστικά σε όλους τους προτεινόμενους ορισμούς [6]:

- Αντίληψη του περιβάλλοντος και της πολυπλοκότητας του πραγματικού κόσμου.
- Επεξεργασία δεδομένων που περιλαμβάνει συλλογή και ερμηνεία.
- Δυνατότητα λήψης αποφάσεων και ανάληψης δράσης.
- Επίτευξη συγκεκριμένων στόχων, χαρακτηριστικό το οποίο αποτελεί και τον απώτερο σκοπό των συστημάτων ΤΝ.

Ο ΟΟΣΑ, ορίζοντας την ΤΝ κάνει αναφορά σε συστήματα βασισμένα στη μηχανή, που έχουν σχεδιαστεί να λειτουργούν με διαφορετικά επίπεδα αυτονομίας και είναι σε θέση για ένα δεδομένο σύνολο στόχων που θα τεθούν από τον άνθρωπο, να κάνουν προβλέψεις, συστάσεις ή να λάβουν αποφάσεις, επηρεάζοντας πραγματικά ή εικονικά

περιβάλλοντα [7]. Αντίστοιχα, ένας κοινά αποδεκτός ορισμός στο επίπεδο της ΕΕ συμπεριλαμβάνεται στην «Πράξη για την Τεχνητή Νοημοσύνη» (AI Act) η οποία πρόσφατα έγινε αποδεκτή από το Ευρωπαϊκό Κοινοβούλιο. Ο παρεχόμενος ορισμός αναφέρει ότι, ένα σύστημα ΤΝ είναι λογισμικό που έχει αναπτυχθεί για ένα σύνολο καθορισμένων από τον άνθρωπο στόχων που παράγουν αποτελέσματα όπως περιεχόμενο, προβλέψεις, συστάσεις ή αποφάσεις και επηρεάζουν τα περιβάλλοντα με τα οποία αλληλοεπιδρούν [8]. Στο κείμενο της πράξης, διακρίνονται 3 προσεγγίσεις στην ανάπτυξη συστημάτων ΤΝ: μηχανική μάθηση, συμβολικές προσεγγίσεις (λογικής/γνώσης) και στατιστικές προσεγγίσεις. Το Υπουργείο Άμυνας των ΗΠΑ, έχει υιοθετήσει έναν απλοποιημένο ορισμό περιγράφοντας την ΤΝ ως «τη δυνατότητα μηχανών να εκτελούν δραστηριότητες που κανονικά απαιτούν ανθρώπινη νοημοσύνη» [9]. Μελέτη της υπηρεσίας ερευνών του Κογκρέσου των ΗΠΑ, κατηγοριοποιεί τους ορισμούς που έχουν προταθεί διαχρονικά για την ΤΝ σε 4 κατηγορίες: συστήματα που σκέφτονται σαν άνθρωποι, συστήματα που σκέφτονται λογικά, συστήματα που δρουν σαν άνθρωποι και συστήματα που δρουν λογικά [10]. Από την ίδια μελέτη αλλά και τη μελέτη του ΟΟΣΑ που προαναφέρθηκε, ενδιαφέρον παρουσιάζει ο διαχωρισμός των συστημάτων ΤΝ που έχουν αναπτυχθεί για μια συγκεκριμένη εργασία ή ένα στενό κύκλο εργασιών (*weak/narrow AI*) και των συστημάτων ΤΝ που έχουν γνωστικές ικανότητες συγκρίσιμες με εκείνες του ανθρώπου (*strong/general AI*), που τους επιτρέπουν να προσαρμόζονται σε διαφορετικά περιβάλλοντα και να εκτελούν καθήκοντα πέρα από τον αρχικό τους προγραμματισμό. Εξετάζοντας τη σχετική βιβλιογραφία σε συνδυασμό με το αντικείμενο της εργασίας, μια πρώτη παρατήρηση είναι ότι στον τομέα της άμυνας, οι περισσότερες εφαρμογές ΤΝ δεν αφορούν μόνο λογισμικό αλλά και υλικό. Επιπλέον σε πολλές περιπτώσεις η διάκριση μεταξύ ανθρώπινης και λογικής απόφασης είναι διφορούμενη σε περιβάλλοντα που λαμβάνονται κρίσιμες αποφάσεις. Κατά συνέπεια, ένας πληρέστερος ορισμός της ΤΝ, ο οποίος καλύπτει τις ανάγκες της εργασίας, είναι αυτός που έχει δοθεί από την Ομάδα Εμπειρογνωμόνων Υψηλού Επιπέδου για την τεχνητή νοημοσύνη (HLEG) της ΕΕ. Σύμφωνα με τον εν λόγω ορισμό, συστήματα ΤΝ λογίζονται, υλικό και λογισμικό σχεδιασμένα από τον άνθρωπο, που μπορούν να αντιλαμβάνονται το περιβάλλον και να δρουν στο φυσικό και ψηφιακό κόσμο και είναι σε θέση να επιλέξουν τις βέλτιστες ενέργειες για να επιτύχουν ένα τιθέμενο σκοπό και

επιπλέον μπορούν να προσαρμόσουν τη συμπεριφορά τους αναλύοντας τον τρόπο με τον οποίο επιδρούν στο περιβάλλον με τις πράξεις τους [6].

Στη διάρκεια της εξέλιξης της ΤΝ, έχουν εμφανιστεί διαφορετικές προσεγγίσεις για την επίλυση προβλημάτων με τη βοήθεια αλγορίθμων και μεθοδολογιών έξυπνων συστημάτων, οι οποίες κατατάσσονται σε τρεις κατηγορίες [11]:

- Συμβολικές προσεγγίσεις (symbolic AI): περιλαμβάνουν προσεγγίσεις βασισμένες στη λογική και προσεγγίσεις βασισμένες στη γνώση. Η πρώτη περίπτωση αφορά εργαλεία που χρησιμοποιούνται για την αναπαράσταση της γνώσης και την επίλυση προβλημάτων. Η δεύτερη χρησιμοποιεί μια βάση γνώσης που βασίζεται σε δηλώσεις, διαδικασίες, ευρετήρια ή δομές δεδομένων και μια μηχανή εξαγωγής συμπερασμάτων που περιλαμβάνει τεχνικές όπως η συλλογιστική βάσει κανόνων, η συλλογιστική βάσει μοντέλων και η συλλογιστική βάσει περιπτώσεων για την εξαγωγή νέων γνώσεων/αποφάσεων. Παράδειγμα αποτελούν τα λεγόμενα έμπειρα συστήματα (expert systems) τα οποία βασίζονται σε κανόνες (rule-based) και εξομοιώνουν την ανθρώπινη σκέψη κατά τη λήψη αποφάσεων [12].
- Στατιστικές προσεγγίσεις (statistical AI): περιλαμβάνουν πιθανοτικές προσεγγίσεις, που αποτυπώνουν την αβεβαιότητα σε πολύπλοκες σχέσεις και γνώσεις όπου οι αποφάσεις λαμβάνονται μέσω τεχνικών στατιστικής επαγωγής και προσεγγίσεων μηχανικής μάθησης που αναλύονται στη συνέχεια. Οι συγκεκριμένες προσεγγίσεις κερδίζουν όλο και περισσότερο έδαφος και έχουν σημαντικό ρόλο στην εξάπλωση της ΤΝ τις τελευταίες δεκαετίες.
- Υποσυμβολικές προσεγγίσεις (sub-symbolic AI): περιλαμβάνουν προσεγγίσεις που μιμούνται τους νευρώνες του ανθρώπινου εγκεφάλου και λαμβάνουν αποφάσεις με βάση τα βάρη των συνδέσεων μεταξύ κόμβων. Παράδειγμα αποτελούν οι γενετικοί αλγόριθμοι οι οποίοι αντλούν έμπνευση από τη διαδικασία της φυσικής επιλογής και εξέλιξης, χρησιμοποιώντας ένα σύνολο πιθανών λύσεων το οποίο εξελίσσουν επαναληπτικά για να βρουν βέλτιστες ή σχεδόν βέλτιστες λύσεις σε ένα δεδομένο πρόβλημα [12].

Οι παραπάνω προσεγγίσεις δεν είναι αλληλοαποκλειόμενες αλλά συμπληρωματικές μεταξύ τους και πλέον τα σύγχρονα συστήματα ΤΝ χρησιμοποιούν υβριδικές προσεγγίσεις, δηλαδή συνδυασμό δύο ή περισσοτέρων τεχνικών.

Η κυρίαρχη τεχνολογία που χρησιμοποιείται σήμερα στον τομέα της ΤΝ είναι η μηχανική μάθηση (machine learning). Αφορά σε τεχνικές που επιτρέπουν σε μηχανές να μαθαίνουν με αυτοματοποιημένο τρόπο, μέσω προτύπων και συμπερασμάτων αντί ρητών εντολών που έχουν δοθεί από τον άνθρωπο. Συχνά οι τεχνικές αυτές επιτυγχάνουν τη μάθηση μέσω της τροφοδότησης της μηχανής με πλήθος παραδειγμάτων σωστών συμπερασμάτων. Ταυτόχρονα δίνονται και ορισμένοι κανόνες τους οποίους χρησιμοποιεί η μηχανή για να εκπαιδευθεί μέσω δοκιμών και σφαλμάτων [13].

Τα συστήματα ΤΝ αναλόγως των δυνατοτήτων τους διακρίνονται σε δύο ευρείες κατηγορίες: την προγνωστική ΤΝ (predictive AI) και την παραγωγική ΤΝ (generative AI). Η προγνωστική ΤΝ επιχειρεί να προβλέψει μελλοντικά αποτελέσματα με βάση ιστορικά δεδομένα και συσχετίσεις. Παράδειγμα εφαρμογών προγνωστικής ΤΝ είναι η υπολογιστική όραση. Η παραγωγική ΤΝ μπορεί να εντοπίσει πολύπλοκες σχέσεις σε μεγάλα σύνολα δεδομένων εκπαίδευσης (σώματα δεδομένων-corpus) και στη συνέχεια να γενικεύσει αυτά που μαθαίνει για να δημιουργήσει νέα δεδομένα. Παράδειγμα αποτελούν μοντέλα όπως το ChatGPT και το BardAI.

Πολλά σύγχρονα συστήματα ΤΝ βασίζονται σε γλωσσικά μοντέλα. Πρόκειται για υπολογιστικά πλαίσια που αποσκοπούν στην κατανόηση και τη δημιουργία κειμένου που μοιάζει με ανθρώπινο κείμενο. Βασίζονται στην αρχή της πιθανολογικής πρόβλεψης, όπου το υπολογιστικό σύστημα μαθαίνει μοτίβα και εξαρτήσεις σε ακολουθίες λέξεων για να εκτιμήσει την πιθανότητα μιας συγκεκριμένης λέξης δεδομένου του προηγούμενου πλαισίου με αποτέλεσμα να μπορεί να παράγει συνεκτικό και σχετικό με το πλαίσιο κείμενο. Αυτό επιτυγχάνεται με την εκπαίδευση του μοντέλου σε μεγάλες ποσότητες δεδομένων κειμένου, επιτρέποντάς του να μάθει την κατανομή των λέξεων, φράσεων και συντακτικές δομές σε κάθε γλώσσα. Τα μεγάλα γλωσσικά μοντέλα (Large Language Models - LLMs) έχουν εκπαιδευτεί σε εκτεταμένα δεδομένα κειμένου και διαθέτουν πολλαπλές παραμέτρους που τους

δίνουν τη δυνατότητα παραγωγής πολύπλοκων κειμένων σε ένα ευρύ φάσμα θεμάτων [14].

Μια σημαντική τεχνική μηχανικής μάθησης είναι η μεταφορά μάθησης (transfer learning) [15]. Αφορά στη χρήση ενός προεκπαιδευμένου μοντέλου ΤΝ για την εκπαίδευση ενός άλλου μοντέλου με σκοπό την επέκταση ή βελτίωση των δυνατοτήτων του. Χρησιμοποιείται σε εφαρμογές όπως η υπολογιστική όραση, η επεξεργασία φυσικής γλώσσας και γενικά σε τομείς όπου δεν είναι πάντα άμεσα διαθέσιμες μεγάλες ποσότητες επισημασμένων δεδομένων. Μπορεί να μειώσει σημαντικά τον όγκο του dataset που απαιτείται για την εκπαίδευση ενός μοντέλου και συχνά οδηγεί σε ταχύτερη σύγκλιση στα επιθυμητά αποτελέσματα.

2.2. Αλγόριθμοι Μηχανικής Μάθησης

Ο αριθμός των αλγορίθμων μηχανικής μάθησης είναι συνεχώς αυξανόμενος, καθώς παρουσιάζονται νέοι ή βελτιωμένες εκδόσεις παλαιότερων αλγορίθμων, οι οποίοι όπως προαναφέρθηκε έχουν συμβάλει στην ταχεία ανάπτυξη των δυνατοτήτων της ΤΝ. Οι αλγόριθμοι μηχανικής μάθησης διακρίνονται σε 4 κύριες κατηγορίες [9],[13]:

- Επιβλεπόμενη Μάθηση (Supervised Learning), όπου τα δεδομένα που δίνονται στον αλγόριθμο περιέχουν ήδη τη σωστή απάντηση. Στην περίπτωση αυτή τα δεδομένα έχουν επισημανθεί από ανθρώπους «επιβλέποντες». Ωστόσο η διαδικασία τροφοδοσίας του συστήματος με επαρκή, επισημασμένα δεδομένα είναι συνήθως δυσχερής, χρονοβόρα και δαπανηρή.
- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning), όπου η ομαδοποίηση των δεδομένων γίνεται από τον αλγόριθμο χωρίς πρότερη πληροφόρηση σχετικά με τον τρόπο διαίρεσής τους σε ομάδες. Η συγκεκριμένη προσέγγιση έχει χαμηλότερη απόδοση από την επιβλεπόμενη μάθηση αλλά μπορεί να χρησιμοποιηθεί σε περιπτώσεις που η εφαρμογή επιβλεπόμενης μάθησης δεν είναι εφικτή.
- Μερικώς Επιβλεπόμενη Μάθηση (Semi-supervised Learning), η οποία συνδυάζει τα πλεονεκτήματα των δύο προηγουμένων.
- Ενισχυτική Μάθηση (Reinforcement Learning), η οποία συνεπάγεται τη δημιουργία ενός συστήματος ανταμοιβών μέσα σε ένα τεχνητό περιβάλλον για

να διδάξει σε έναν νοήμονα πράκτορα πώς να δρα σε διαφορετικές καταστάσεις και να ενεργεί σε ένα δεδομένο πλαίσιο. Οι νοήμονες πράκτορες συλλέγουν αυτόνομα δεδομένα και αυτοβελτιώνονται μέσω μιας διαδικασίας δοκιμών και σφαλμάτων (trial and error). Αν και πρόκειται για μια πολλά υποσχόμενη μέθοδο, συναντά ακόμα πολλές προκλήσεις στην υλοποίησή της.

Επιπλέον κατηγορίες αλγορίθμων μηχανικής μάθησης είναι οι εξής [16]:

- Μάθηση πολλαπλών καθηκόντων (Multi-task learning - MTL), όπου ένα μοντέλο εκπαιδεύεται για να εκτελεί ταυτόχρονα πολλαπλές εργασίες εκμεταλλευόμενο τις ομοιότητες μεταξύ των διαφορετικών εργασιών. Η ιδέα είναι ότι η κοινόχρηστη γνώση που αποκτάται από μια εργασία μπορεί να ωφελήσει την απόδοση σε άλλες συναφείς εργασίες, οδηγώντας σε βελτιωμένη γενίκευση και αποτελεσματικότητα.
- Συλλογική μάθηση (Ensemble Learning), όπου γίνεται συνδυασμός των προβλέψεων πολλαπλών μοντέλων για τη βελτίωση της συνολικής απόδοσης και της ευρωστίας. Με τη συγκεκριμένη μέθοδο επιχειρείται η αντιστάθμιση των αδυναμιών επιμέρους μοντέλων για συγκεκριμένα προβλήματα.
- Νευρωνικά δίκτυα (Neural Networks), όπου επιχειρείται να αναγνωρισθούν οι υποκείμενες σχέσεις σε ένα σύνολο δεδομένων μέσω μιας διαδικασίας που μιμείται τον τρόπο με τον οποίο λειτουργεί ο ανθρώπινος εγκέφαλος. Είναι ιδιαίτερα αποδοτικά σε εφαρμογές όπως η αναγνώριση ομιλίας και εικόνας και οι μεταφράσεις.
- Μάθηση κατά περίπτωση (Instance-based Learning), όπου σε αντίθεση με άλλες μεθόδους, τα δεδομένα εκπαίδευσης αποθηκεύονται χωρίς να εξάγονται συσχετίσεις. Τη στιγμή της ερώτησης αντλείται μια απάντηση από την εξέταση των πλησιέστερων γειτόνων της ερώτησης. Με τη συγκεκριμένη τεχνική ο αλγόριθμος μπορεί εύκολα να προσαρμοστεί με νέα δεδομένα εκπαίδευσης.

Μια υποκατηγορία των νευρωνικών δικτύων είναι η Βαθιά Μάθηση (Deep Learning). Πρόκειται για συνδυασμό πολλαπλών νευρωνικών δικτύων σε ιεραρχική δομή που αυξάνουν την πολυπλοκότητα μεταξύ εισόδου και εξόδου [13]. Κάθε επίπεδο νευρώνων χρησιμοποιεί τα δεδομένα από το επίπεδο που βρίσκεται κάτω από αυτό, κάνει υπολογισμούς και προσφέρει την έξοδό του στα επίπεδα που βρίσκονται πάνω

από αυτό. Οι δυνατότητες της συγκεκριμένης προσέγγισης, απορρέουν από την ικανότητά του μοντέλου να μαθαίνει αυτόματα ιεραρχικά και αφηρημένα χαρακτηριστικά από μεγάλες ποσότητες δεδομένων. Η βελτιστοποίηση των συγκεκριμένων αλγορίθμων τα τελευταία χρόνια, έχει σαν αποτέλεσμα η βαθιά μάθηση να αποτελεί μια από τις πιο δημοφιλείς προσεγγίσεις μηχανικής μάθησης με πολλαπλές εφαρμογές τόσο στην κυβερνοασφάλεια όσο και στον αμυντικό τομέα.

Μελέτη του ENISA έχει καταγράψει 42 αλγορίθμους που χρησιμοποιούνται στην ανάπτυξη μοντέλων ΤΝ. Για την κατανόηση των τεχνικών εκπαίδευσης ενός μοντέλου ΤΝ, κρίνεται σκόπιμο να γίνει μια σύντομη αναφορά στους κυριότερους αλγορίθμους κάθε κατηγορίας. Στην κατηγορία της επιβλεπόμενης μάθησης, οι χρησιμοποιούμενοι αλγόριθμοι ασχολούνται κυρίως με ταξινόμηση (classification), δηλαδή κατηγοριοποίηση των δεδομένων εισόδου σε συγκεκριμένες κλάσεις. Οι κυριότεροι από αυτούς είναι [16]–[19]:

- Δέντρα απόφασης (Decision Trees): ο αλγόριθμος δημιουργεί αναδρομική κατάτμηση των δεδομένων με βάση τις τιμές των χαρακτηριστικών γνωρισμάτων τους, σχηματίζοντας δενδροειδείς μορφές. Στο δέντρο που σχηματίζεται, η ρίζα είναι η είσοδος, ο κάθε εσωτερικός κόμβος αντιπροσωπεύει μια απόφαση με βάση ένα χαρακτηριστικό, κάθε κλάδος αναπαριστά το αποτέλεσμα της απόφασης και κάθε φύλλο την εκτιμώμενη κλάση ή την τιμή πρόβλεψης για την περίπτωση που ο αλγόριθμος χρησιμοποιείται για παλινδρόμηση (regression). Σημαντικό πλεονέκτημά του συγκεκριμένου αλγορίθμου είναι ότι είναι εύκολα κατανοητός και ερμηνεύσιμος καθώς μοιάζει με την ανθρώπινη διαδικασία λήψης απόφασης. Μειονέκτημά του είναι ότι κατά την εκπαίδευση των μοντέλων μπορεί να προκληθεί υπερπροσαρμογή (overfitting), δηλαδή αδυναμία γενίκευσης για νέα και άγνωστα δεδομένα εισόδου.
- Απλή ταξινόμηση Bayes (Naive Bayes): στηρίζεται σε κατανομές πιθανότητας με την «αφελή» παραδοχή ότι τα χαρακτηριστικά της εισόδου είναι ανεξάρτητα μεταξύ τους. Κατά συνέπεια είναι κατάλληλος όταν τηρείται κατ’ ελάχιστο σε μεγάλο βαθμό η συγκεκριμένη παραδοχή. Είναι ιδιαίτερα αποδοτικός σε

εφαρμογές εντοπισμού spam και σε περιπτώσεις που υπάρχουν περιορισμοί στην υπολογιστική ισχύ.

- Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM): βασίζονται στην έννοια της μεγιστοποίησης του περιθωρίου (margin) που διαχωρίζει τις κλάσεις δεδομένων. Τα δεδομένα εισόδου αναπαρίστανται ως πολυδιάστατα διανύσματα με βάση τα χαρακτηριστικά τους. Η απόκριση του αλγορίθμου προκύπτει από την εύρεση του υπερεπιπέδου που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων. Ο αλγόριθμός είναι ιδιαίτερα χρήσιμος όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, ενώ λειτουργεί και σε προβλήματα παλινδρόμησης. Μειονέκτημα αποτελεί η χαμηλή απόδοση για μεγάλα σύνολα δεδομένων ή εισόδων με ύπαρξη «θορύβου», δηλαδή dataset με σφάλματα, ασυνέπειες ή άσχετες πληροφορίες.

Στην κατηγορία της μη επιβλεπόμενης μάθησης, οι χρησιμοποιούμενοι αλγόριθμοι επιλύνουν προβλήματα ομαδοποίησης (clustering) και μείωσης διαστάσεων (dimensionality reduction), δηλαδή τεχνικές για τη μείωση του αριθμού των μεταβλητών εισόδου στα δεδομένα εκπαίδευσης [20]. Οι κυριότεροι από αυτούς είναι [16]–[19]:

- Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis- PCA): αποσκοπεί στη μείωση των διαστάσεων των δεδομένων μέσω της γραμμικής προβολής σε χαμηλότερες διαστάσεις. Με τον τρόπο αυτό τα δεδομένα από ένα σύνολο συσχετιζόμενων μεταβλητών μετασχηματίζονται σε ένα σύνολο τιμών γραμμικά ασυσχέτιστων μεταβλητών που ονομάζονται κύριες συνιστώσες. Μια από τις εφαρμογές του αλγορίθμου είναι στην αφαίρεση θορύβου από ένα dataset.
- Αλγόριθμος κ-Μέσων (K-Means): αποσκοπεί στον εντοπισμό εγγενών προτύπων ή δομών στα δεδομένα εισόδου χωρίς να διαθέτει προκαθορισμένες ετικέτες (ως μη επιβλεπόμενος αλγόριθμος). Επιλύει προβλήματα ομαδοποίησης (clustering), όπου κ ο αριθμός των ομάδων. Το κ πρέπει να δίνεται ως είσοδος του αλγορίθμου, η εκτίμησή του ωστόσο αποτελεί μια πρόκληση και μπορεί να απαιτηθούν δοκιμές διαφόρων τιμών μέχρι να επιτευχθεί το βέλτιστο αποτέλεσμα.

Στην κατηγορία της μερικώς επιβλεπόμενης μάθησης κυριότεροι αλγόριθμοι είναι:

- Παραγωγικά Μοντέλα (Generative Models): δημιουργούν νέα ρεαλιστικά δείγματα δεδομένων (χαρακτηριστικά και κλάσεις) που ομοιάζουν με τα δεδομένα εκπαίδευσης. Έχουν εφαρμογές στη σύνθεση εικόνων, στην παραγωγή κειμένου και στην επαύξηση δεδομένων.
- Αυτοεκπαιδευόμενα μοντέλα (Self-Training Models): τροφοδοτούνται αρχικά με επισημασμένα δεδομένα εκπαίδευσης, τα οποία χρησιμοποιούνται στη συνέχεια για την ταξινόμηση μη επισημασμένων δεδομένων με τα οποία το μοντέλο επανεκπαιδεύεται. Αποτελεί μια αποδοτική λύση όταν δεν υφίσταται μεγάλο πλήθος επισημασμένων δεδομένων και είναι σημαντικό να αξιοποιηθούν τα διαθέσιμα μη επισημασμένα δεδομένα. Ωστόσο απαιτείται ταυτόχρονα στενή παρακολούθηση και επικύρωση για να ελεγχθεί η ποιότητα των ψευδοετικετών που δημιουργεί ο αλγόριθμος.
- Επαγωγικές Μηχανές Διανυσμάτων Υποστήριξης (Transductive Support Vector Machine - TSVM): πρόκειται για επέκταση των SVM για το χειρισμό dataset που είναι μερικώς επισημασμένα. Χρησιμοποιείται για την επισήμανση των δεδομένων με τέτοιο τρόπο ώστε να μεγιστοποιείται το περιθώριο μεταξύ επισημασμένων και μη επισημασμένων δεδομένων. Και σε αυτή την περίπτωση απαιτείται στενή παρακολούθηση και επικύρωση των αποτελεσμάτων.

Στην κατηγορία της συλλογικής μάθησης οι κυριότεροι αλγόριθμοι είναι:

- Ενδυνάμωση (Boosting): συνδυάζουν τις προβλέψεις αδύναμων μοντέλων (weak learners), δηλαδή ταξινομήσεων που δεν ανταποκρίνονται επαρκώς στην πραγματικότητα για να δημιουργήσουν ισχυρότερα μοντέλα (strong learners) στα οποία οι ταξινομήσεις είναι περισσότερο ρεαλιστικές. Μέσω της ενδυνάμωσης μπορεί να προκύψουν μοντέλα ποιο αποδοτικά και ποιο ανθεκτικά σε υπερπροσαρμογή σε σύγκριση με περίπλοκα μεμονωμένα μοντέλα.
- Επανατοποθέτηση (Bagging): χρησιμοποιούνται για τη βελτίωση της ακρίβειας και της σταθερότητας ενός μοντέλου. Λειτουργεί με την εκπαίδευση πολλαπλών

στιγμιότυπων του ίδιου αλγορίθμου σε διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης και τον συνδυασμό στη συνέχεια των εξαγόμενων προβλέψεων.

Στην κατηγορία της ενισχυτικής μάθησης οι κυριότεροι αλγόριθμοι είναι [19]–[21]:

- Μοντέλα Monte Carlo: χρησιμοποιούν τυχαία δειγματοληψία για την εξαγωγή αριθμητικών αποτελεσμάτων. Εφαρμόζονται για την εκτίμηση της αξίας των καταστάσεων ή των ζευγών καταστάσεων-δράσεων και τη βελτίωση της πολιτικής ενός έξυπνου πράκτορα που αλληλεπιδρά με ένα περιβάλλον.
- Hidden Markov Model – HMM: χρησιμοποιείται για τη μοντελοποίηση συστημάτων όπου η υποκείμενη κατάσταση δεν είναι άμεσα ανιχνεύσιμη αλλά μπορεί να συναχθεί από τις παρατηρούμενες εξόδους.
- Q-Learning: χρησιμοποιείται για την εκμάθηση της τιμής μιας ενέργειας σε μια συγκεκριμένη κατάσταση χωρίς να απαιτεί μοντέλο του περιβάλλοντος.
- Proximal Policy Optimization – PPO: χρησιμοποιείται για να βρεθεί μια βέλτιστη πολιτική για έναν πράκτορα που αλληλεπιδρά με ένα περιβάλλον ώστε να μεγιστοποιηθούν οι αθροιστικές ανταμοιβές.

Στην κατηγορία μάθησης κατά περίπτωση ο αλγόριθμος που χρησιμοποιείται κυρίως είναι ο αλγόριθμος κ-πλησιέστερων γειτόνων (k-Nearest Neighbor), όπου κ ο αριθμός των πλησιέστερων γειτόνων ενός μη ταξινομημένου σημείου. Ο αλγόριθμος βασίζεται στην ευκλείδεια απόσταση μεταξύ των δειγμάτων εισόδου και του δείγματος εκπαίδευσης. Διενεργεί προβλέψεις με βάση την κλάση πλειοψηφίας ή το μέσο όρο των k-κοντινότερων σημείων δεδομένων στο χώρο χαρακτηριστικών. Η βασική ιδέα είναι ότι περιπτώσεις με παρόμοια χαρακτηριστικά τείνουν να έχουν παρόμοιες τιμές στόχου. Χρησιμοποιείται τόσο για ταξινόμηση όσο και για παλινδρόμηση. Μειονέκτημά του είναι ότι έχει υψηλό υπολογιστικό κόστος καθώς χρειάζεται να αποθηκεύσει όλα τα δεδομένα και να εκτελεί υπολογισμούς με το σύνολο τους, ενώ καταλαμβάνει συγκριτικά περισσότερο χώρο για να αποθηκεύσει όλα τα δεδομένα εκπαίδευσης.

2.3. Ο Κύκλος Ζωής της ΤΝ

Ο κύκλος ζωής της ΤΝ αφορά στα στάδια ανάπτυξης, υλοποίησης και συντήρησης ενός συστήματος ΤΝ. Η κατανόηση του κύκλου ζωής είναι σημαντική για να εντοπισθούν σημεία αδυναμίας αλλά και τρόποι επίθεσης σε ένα σύστημα ΤΝ. Τα

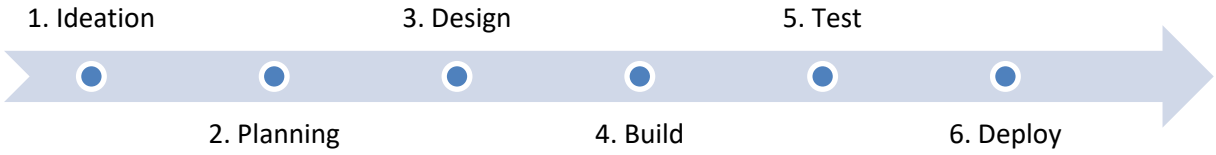
επιμέρους στάδια μπορεί να διαφέρουν αναλόγως της χρησιμοποιούμενης μεθοδολογίας ή των απαιτήσεων και των χαρακτηριστικών ενός έργου.

Η διαδικασία ανάπτυξης ενός μοντέλου μηχανικής μάθησης περιλαμβάνει μια σειρά από αποφάσεις σχετικά με την κατανόηση του προβλήματος που πρόκειται να επιλύσει (π.χ. ομαδοποίηση ή παλινδρόμηση), την επιλογή του ή των αλγορίθμων που θα χρησιμοποιηθούν, την ανάπτυξη του μοντέλου, την εκπαίδευσή του με κατάλληλα σύνολα δεδομένων και τον έλεγχο των προβλέψεων και κατά πόσο πληρούν τις απαιτήσεις. Στη συνέχεια γίνεται η επανεκπαίδευση του μοντέλου με νεότερα στοιχεία.

Ένα γενικό πλαίσιο των σταδίων που απαιτούνται για την ανάπτυξη ενός συστήματος ΤΝ περιλαμβάνει 4 βήματα [1]: τον καθορισμό των απαιτήσεων, την επιλογή και προετοιμασία των δεδομένων εκπαίδευσης, τη μοντελοποίηση που εμπεριέχει την ανάπτυξη, την αξιολόγηση και τη βελτιστοποίηση του μοντέλου και την πρόβλεψη, όπου επιβεβαιώνεται η εγκυρότητα των αποτελεσμάτων και ελέγχεται η απόδοση του μοντέλου για νέα δεδομένα. Είναι σημαντικό να σημειωθεί ότι δεν πρόκειται για μια γραμμική διαδικασία τύπου καταρράκτη, αλλά για μια επαναληπτική διαδικασία, όπου κάθε βήμα θα πρέπει να εκτελείται περισσότερες από μία φορές προκειμένου να επιτευχθεί μια ικανοποιητική λύση.

Ένα περισσότερο λεπτομερές πλαίσιο περιλαμβάνει 6 στάδια: διαμόρφωση ιδεών, προγραμματισμό, σχεδιασμό, ανάπτυξη, δοκιμή και λειτουργία [13]. Στο πρώτο στάδιο λαμβάνονται αποφάσεις ως προς το επίπεδο αυτονομίας του συστήματος, καθορίζονται οι απαιτήσεις και εξετάζονται τυχόν κίνδυνοι από επιπλέον χρήσεις που μπορεί να έχει το υπό ανάπτυξη σύστημα. Στο δεύτερο στάδιο εξετάζεται ο τύπος της ανάδρασης που χρειάζεται το σύστημα για να είναι αποδοτικό καθώς και οι απαιτήσεις ελέγχου του συστήματος τόσο σε επίπεδο αλγορίθμου όσο και σε επίπεδο εξόδων. Καθορίζονται επίσης οι απαιτήσεις ασφαλείας. Στο τρίτο στάδιο εξετάζονται και προετοιμάζονται τα δεδομένα εκπαίδευσης, γίνεται ανάλυση της επιφάνειας επίθεσης και μοντελοποίηση των απειλών. Στο τέταρτο στάδιο, γίνεται η εκπαίδευση του μοντέλου και η παρακολούθηση της προόδου με βάση τις εξόδους. Στο συγκεκριμένο στάδιο προτείνεται να υπάρχει σε λειτουργία μηχανισμός παύσης (kill-switch) για να διατηρηθεί ο ανθρώπινος έλεγχος των διαδικασιών ΤΝ. Στο πέμπτο στάδιο εξετάζεται η απόδοση του συστήματος και η επίτευξη των στόχων του. Στο τελευταίο στάδιο το

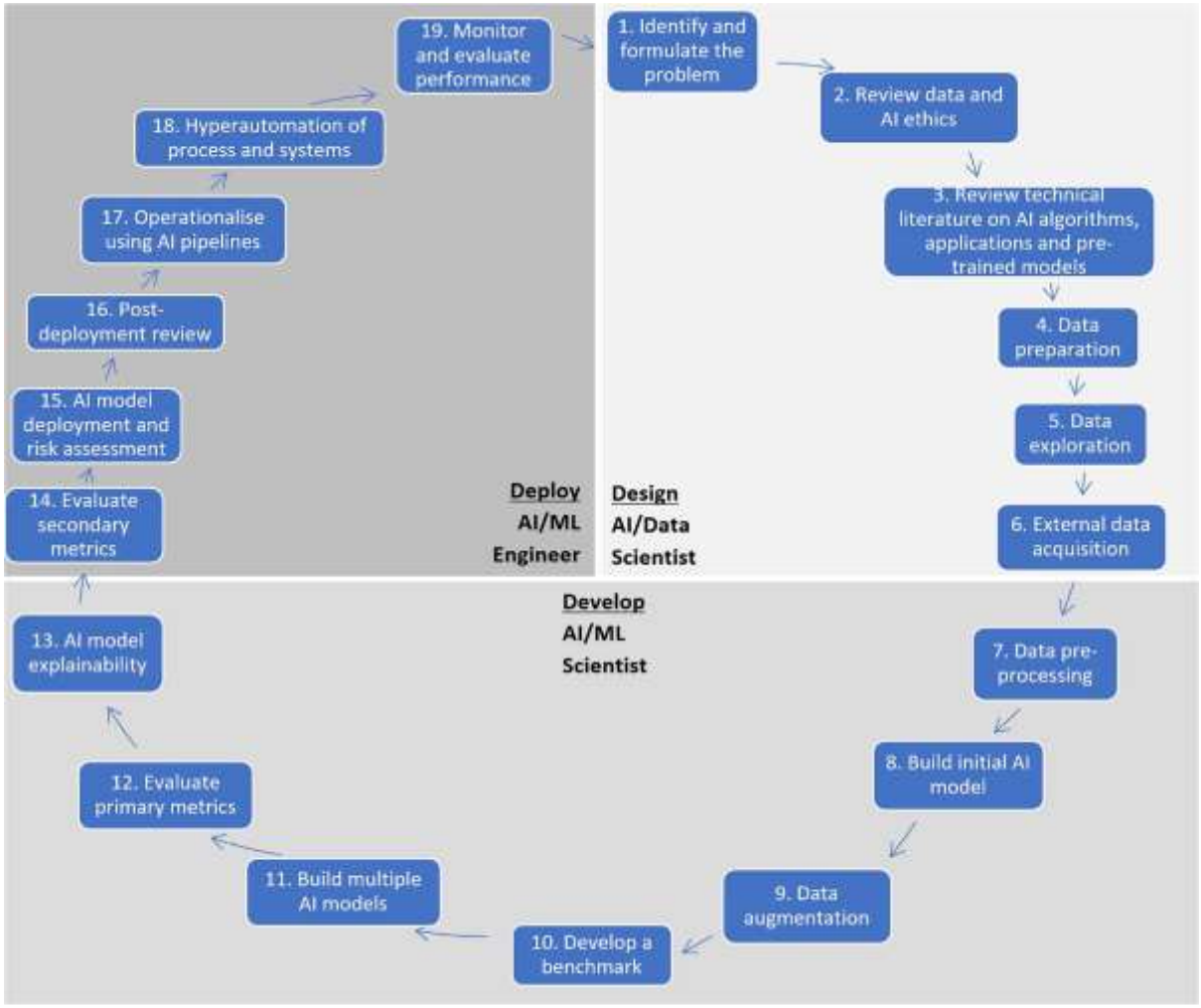
σύστημα παρακολουθείται συνεχώς για την απόδοση, την ανθεκτικότητα και τυχόν ύποπτη συμπεριφορά. Παρατηρούμε ότι το συγκεκριμένο μοντέλο επεκτείνει τη γενική διαδικασία που αναφέρθηκε παραπάνω με έμφαση σε θέματα ασφαλείας.



Σχήμα 2.1: AI systems life cycle[13]

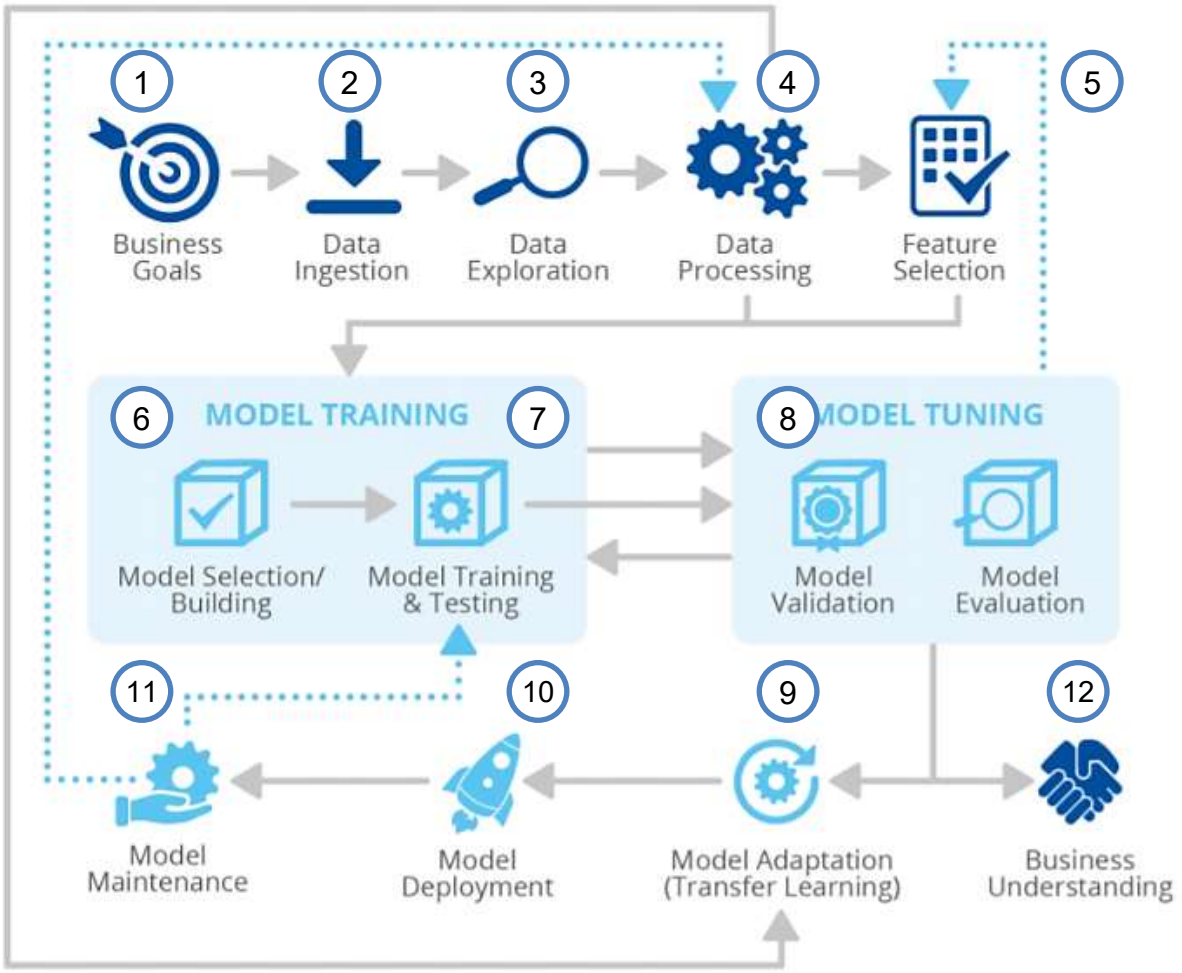
Το κέντρο CDAC προτείνει έναν αναλυτικότερο κύκλο ζωής ΤΝ που αποτελείται από τρεις φάσεις και 19 στάδια [22] όπως φαίνεται στο επόμενο σχήμα. Προ της εφαρμογής της διαδικασίας, προτείνεται η εκτίμηση ρίσκου για θέματα ιδιωτικότητας, κυβερνοασφάλειας, εμπιστοσύνης, ερμηνευσιμότητας, επεξηγησιμότητας, χρηστικότητας και ευρωστίας του μοντέλου αλλά και τυχόν κοινωνικές επιπτώσεις. Στη φάση "Σχεδιασμός" γίνεται η ανάλυση του προβλήματος, επιλέγονται οι αλγόριθμοι και τα προ-εκπαιδευμένα μοντέλα που θα χρησιμοποιηθούν και εξετάζονται οι κατευθυντήριες γραμμές και τα πλαίσια δεοντολογίας, που πρέπει να διέπουν την ανάπτυξη του συστήματος. Ακολουθεί η προετοιμασία των δεδομένων και η απόκτηση εξωτερικών δεδομένων που απαιτούνται για την εκπαίδευση του μοντέλου. Η φάση "Ανάπτυξη" είναι προσανατολισμένη σε τεχνικά θέματα, καθώς μετατρέπει τα δεδομένα και τους αλγορίθμους σε μοντέλα ΤΝ που συγκρίνονται, αξιολογούνται και ερμηνεύονται. Στη φάση "Εφαρμογή" αξιολογούνται οι επιδόσεις του μοντέλου, το οποίο στη συνέχεια τίθεται σε λειτουργία. Ακολουθεί εκ νέου ανάλυση επικινδυνότητας και έλεγχος της απόδοσης με το σύστημα σε λειτουργία. Κατά τη διάρκεια της λειτουργίας γίνονται βελτιώσεις για την αυτοματοποίηση των διαδικασιών και τη βελτίωση της απόδοσης. Η ολοκληρωμένη λύση ΤΝ, παρακολουθείται και αξιολογείται συνεχώς για να τροφοδοτήσει την επόμενη επανάληψη του κύκλου ζωής. Παρατηρούμε ότι με βάση το συγκεκριμένο μοντέλο, ένα σύστημα ΤΝ αντιμετωπίζεται σαν ένας ζωντανός οργανισμός ο οποίος αναπτύσσεται συνεχώς καθώς ο κύκλος επανατροφοδοτείται με βάση την προηγούμενη λειτουργία του μοντέλου. Επιπλέον

δίνεται έμφαση σε θέματα δεοντολογίας και κυβερνοασφάλειας τόσο πριν την έναρξη της ανάπτυξης του συστήματος όσο και κατά τη σχεδίαση αλλά και κατά τη λειτουργία.



Σχήμα 2.2: CDAC AI life cycle

Ο ENISA έχει αναπτύξει ένα πλαίσιο 12 σταδίων για την ανάπτυξη ενός συστήματος ΤΝ [20]. Στο πρώτο στάδιο λαμβάνει χώρα η κατανόηση του επιχειρηματικού πλαισίου του συστήματος και των δεδομένων που απαιτούνται για την επίτευξη των επιχειρηματικών στόχων. Επιπλέον καθορίζονται μετρικές αξιολόγησης της επίτευξης των στόχων αυτών. Στο δεύτερο στάδιο το σύστημα τροφοδοτείται με τα δεδομένα τα οποία μπορεί να είναι σε διάφορες μορφές και να συλλέγονται σε πραγματικό χρόνο ή από κάποιο αποθετήριο δεδομένων. Στο τρίτο στάδιο εξετάζονται τα ληφθέντα δεδομένα και επικυρώνονται.

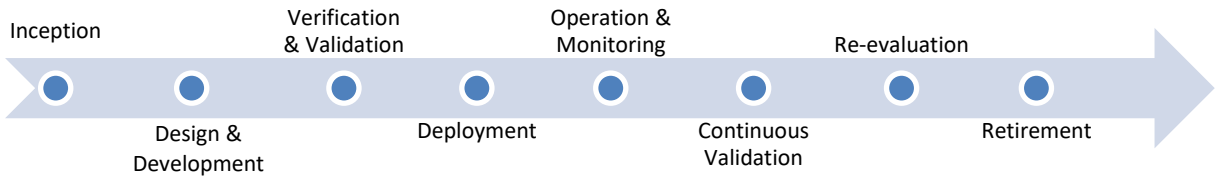


Σχήμα 2.3: ENISA Typical AI lifecycle [20]

Στο τέταρτο στάδιο γίνεται επεξεργασία και μετασχηματισμός των δεδομένων ώστε να είναι επεξεργάσιμα από τους αλγόριθμους. Στο πέμπτο στάδιο απορρίπτονται στοιχεία των δεδομένων τα οποία δεν έχουν κάποια χρησιμότητα για τον αλγόριθμο και δημιουργείται το βελτιωμένο dataset που θα χρησιμοποιηθεί από το μοντέλο. Στο έκτο στάδιο επιλέγεται ο καταλληλότερος αλγόριθμος με βάση τις απαιτήσεις του συστήματος και το dataset που έχει δημιουργηθεί. Στο έβδομο στάδιο εκτελείται ο αλγόριθμος με τις κατάλληλες παραμέτρους και σύμφωνα με τα δεδομένα εκπαίδευσης. Επίσης εκτελούνται δοκιμές επικύρωσης της ορθής εκπαίδευσης. Στο όγδοο στάδιο το μοντέλο προσαρμόζεται χρησιμοποιώντας ένα σύνολο δεδομένων επικύρωσης, σύμφωνα με τις συνθήκες ανάπτυξης. Στο ένατο στάδιο χρησιμοποιείται ένα προ-εκπαιδευμένο μοντέλο για τη βελτίωση της ακρίβειας. Στο δέκατο στάδιο το μοντέλο τίθεται σε χρήση. Στο ενδέκατο στάδιο παρακολουθούνται τα αποτελέσματα

των συμπερασμάτων του συστήματος καθώς και τα δεδομένα εισόδου που λαμβάνει για να εντοπιστούν πιθανές αλλαγές ή παρεκκλίσεις. Εφόσον απαιτείται γίνεται επανεκπαίδευση του μοντέλου, επαναλαμβάνοντας τον κύκλο από το στάδιο 4. Στο δωδέκατο στάδιο αξιολογείται η αξία του ανεπτυγμένου συστήματος και ο αντίκτυπος που έχει.

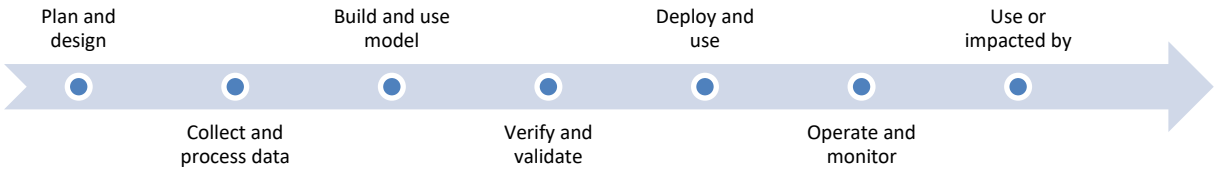
Μια προσπάθεια τυποποίησης του κύκλου ζωής ενός συστήματος ΤΝ επιχειρείται με το πρότυπο ISO/IEC 5338, στο οποίο αντί να περιγράφονται συγκεκριμένα στάδια, αναλύονται 33 επιμέρους διαδικασίες οι οποίες λαμβάνουν χώρα στον κύκλο ζωής ενός συστήματος [23]. Οι διαδικασίες αυτές μπορεί να λαμβάνουν χώρα σε περισσότερα από ένα στάδια του κύκλου και επαναλαμβάνονται καθ' όλη τη διάρκεια ζωής ενός συστήματος ΤΝ. Οι διαδικασίες χωρίζονται σε 4 κατηγορίες: διαδικασίες συμφωνιών (2), οργανωτικές διαδικασίες (6), διαδικασίες τεχνικής διαχείρισης (8) και τεχνικές διαδικασίες (17). Παρέχεται επίσης ένας ενδεικτικός κύκλος ζωής ενός συστήματος ΤΝ που περιλαμβάνει 8 στάδια όπως φαίνονται στο επόμενο σχήμα. Επιπρόσθετα, το πρότυπο προτείνει υψηλού επιπέδου διεργασίες που εκτείνονται σε όλα στάδια του κύκλου ζωής και είναι: διαφάνεια και επεξηγησιμότητα, ασφάλεια και ιδιωτικότητα, διαχείριση κινδύνου και διακυβέρνηση. Σε σχέση με τα μοντέλα που αναλύθηκαν παραπάνω, το πρότυπο εισάγει μια έννοια που δεν υπάρχει στα προηγούμενα προτεινόμενα μοντέλα και αφορά στην απόσυρση ενός συστήματος ΤΝ και εξετάζει θέματα διαχείρισης των dataset που έχουν δημιουργηθεί κατά τη διάρκεια της ζωής του συστήματος.



Σχήμα 2.4: ISO/IEC 5338 Sample AI lifecycle

Τέλος, το Ινστιτούτο NIST προτείνει έναν κύκλο ζωής 7 σταδίων [24]: προγραμματισμός και σχεδίαση, συλλογή και επεξεργασία δεδομένων, κατασκευή και χρήση μοντέλου, επαλήθευση και επικύρωση, ανάπτυξη και χρήση, λειτουργία και

παρακολούθηση, χρήση ή επηρεασμός. Το πρώτο στάδιο έχει να κάνει με το πεδίο εφαρμογής του συστήματος ΤΝ και τον καθορισμό των απαιτήσεων. Το δεύτερο στάδιο αφορά στη συλλογή και την επεξεργασία των δεδομένων εκπαίδευσης. Στο τρίτο και το τέταρτο στάδιο πραγματοποιείται η εκπαίδευση του μοντέλου. Στο πέμπτο στάδιο ενεργοποιείται το σύστημα και ελέγχονται οι έξοδοι σε διαφορετικά σενάρια χρήσης. Στο έκτο στάδιο το σύστημα τίθεται σε πλήρη λειτουργία και παρακολουθείται συνεχώς η απόδοσή του. Στο τελευταίο στάδιο εκτιμάται ο αντίκτυπος του συστήματος στο περιβάλλον στο οποίο χρησιμοποιείται και εξετάζεται αν απαιτούνται διορθωτικές ενέργειες.



Σχήμα 2.5: NIST AI life cycle

Εξετάζοντας τα διάφορα προτεινόμενα μοντέλα κύκλου ζωής συστημάτων ΤΝ, διαπιστώνεται ότι παρουσιάζουν πολλές ομοιότητες και εισέρχονται σε διαφορετικό επίπεδο λεπτομέρειας το καθένα. Πέρα από τις κοινές αρχές, κάθε σύστημα ΤΝ μπορεί να έχει διαφοροποιήσεις στον κύκλο ζωής του αναλόγως της περίπτωσης χρήσης.

ΚΕΦΑΛΑΙΟ 3

ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΚΥΒΕΡΝΟΑΣΦΑΛΕΙΑ

Καθώς οι απειλές στον κυβερνοχώρο γίνονται όλο και πιο εξελιγμένες και ποικίλες, οι τεχνολογίες τεχνητής νοημοσύνης έχουν αναδειχθεί σε σημαντικά εργαλεία για την ενίσχυση των μηχανισμών ανίχνευσης, πρόληψης και αντιμετώπισης απειλών στο πεδίο της κυβερνοασφάλειας. Την ίδια στιγμή αυξάνονται οι κακόβουλες ενέργειες που επιδιώκουν να εκμεταλλευτούν τα τρωτά σημεία των μοντέλων τεχνητής νοημοσύνης, δημιουργώντας νέες απειλές στα ήδη πολλαπλώς προσβαλλόμενα πληροφοριακά συστήματα.

3.1. Κυβερνοασφάλεια της ΤΝ

Καθώς η ΤΝ χρησιμοποιείται σε όλο και περισσότερες εφαρμογές και υπάρχοντα συστήματα, αυξάνεται η επιφάνεια επίθεσης των συστημάτων πέρα από αυτή των παραδοσιακών κυβερνοεπιθέσεων. Συνεπώς υφίσταται μια διττή ανάγκη, αφενός για την ασφάλεια (safety) των συστημάτων ΤΝ, δηλαδή την αποφυγή δυσάρεστων και καταστροφικών συνεπειών από τη χρήση της, αλλά και για την προστασία (security) των συστημάτων αυτών από κακόβουλες ενέργειες. Επιπρόσθετα, ζητήματα ασφάλειας και προστασίας είναι καθοριστικά για την εμπιστοσύνη σε ένα σύστημα ΤΝ η οποία παίζει καίριο ρόλο στην αποδοχή του [7], [25].

Η κυβερνοασφάλεια των συστημάτων ΤΝ διαφέρει από αυτή των λοιπών πληροφοριακών συστημάτων καθώς η ΤΝ εισάγει νέες μορφές κυβερνοεπιθέσεων μέσω των ιδιαίτερων χαρακτηριστικών των αλγορίθμων μηχανικής μάθησης. Ακόμα και αν ένα σύστημα ΤΝ έχει εκπαιδευθεί σύμφωνα με όλα τα πρότυπα και προφυλάξεις, μπορεί να επηρεασθεί η χρήση του από επιθέσεις κατά τη διάρκεια της λειτουργίας του. Το γεγονός αυτό δεν σημαίνει ότι δεν έχουν εφαρμογή τα γνωστά αντίμετρα κυβερνοεπιθέσεων αλλά ότι απαιτούνται επιπλέον ενέργειες και αντίμετρα που είναι προσαρμοσμένα στον κύκλο ζωής και στα ιδιαίτερα χαρακτηριστικά ενός συστήματος ΤΝ.

Στον τομέα της βιομηχανικής κατασκοπίας, καθώς η ΤΝ εισέρχεται σε οπλικά συστήματα, οι αντίπαλοι δεν αναζητούν πλέον τα σχέδια των οπλικών συστημάτων

αλλά τους αλγορίθμους και τα δεδομένα με τα οποία αυτά αναπτύσσονται. Κατά συνέπεια αυξάνονται τα κίνητρα επιθέσεων στα σημεία ανάπτυξης των στρατιωτικών συστημάτων ΤΝ [26].

Τα θέματα ασφαλείας είναι διάχυτα σε όλα τα στάδια του κύκλου ζωής ενός συστήματος ΤΝ. Όπως αναλύθηκε στο προηγούμενο κεφάλαιο, σε όλα τα προτεινόμενα μοντέλα κύκλου ζωής συστημάτων ΤΝ υπάρχουν προβλέψεις για ελέγχους και διορθωτικές ενέργειες. Στο πλαίσιο αυτό, απαιτείται να εκτελείται εκτίμηση επικινδυνότητας των συστημάτων ΤΝ και να εφαρμόζονται τα κατάλληλα αντίμετρα σε όλες τις επιμέρους φάσεις του κύκλου ζωής.

Οι αποτυχίες συστημάτων ΤΝ παρατηρούνται σε όλα τα στάδια εξέλιξης των σχετικών τεχνολογιών με αυξανόμενο ρυθμό, γεγονός που είναι αναμενόμενο καθώς εμφανίζονται συνεχώς νέες περιπτώσεις χρήσης. Από το μπλοκάρισμα νόμιμων email από φίλτρα spam, μέχρι αυτοοδηγούμενα οχήματα που προκάλεσαν δυστύχημα, υπάρχουν ήδη δεκάδες παραδείγματα συστημάτων που εμφάνισαν αστοχίες. Οι αστοχίες αυτές μπορεί να οφείλονται, σε σφάλματα υλοποίησης, σε επίδραση του περιβάλλοντος στο οποίο δραστηριοποιείται το σύστημα ΤΝ ή σε κακόβουλες ενέργειες. Σε επίπεδο υλοποίησης εμφανίζονται προβλήματα αλγοριθμικής μεροληψίας, κακής απόδοσης ή βασικές δυσλειτουργίες [27]. Η μεροληψία μπορεί να προκύψει από λανθασμένη επιλογή και προετοιμασία των δεδομένων εκπαίδευσης και αναλόγως με τη χρήση του συστήματος μπορεί να οδηγήσει σε παραπλάνηση των χρηστών [28]. Η μειωμένη απόδοση μπορεί να οφείλεται επίσης στη χρήση μη αντιπροσωπευτικών δεδομένων εκπαίδευσης. Σε πολλές περιπτώσεις η ενισχυτική εκπαίδευση μπορεί να βελτιώσει την απόδοση ενός μοντέλου.

Για την ανάλυση των κακόβουλων ενεργειών σε συστήματα ΤΝ, ο οργανισμός MITRE έχει αναπτύξει τη βάση γνώσης ATLAS στην οποία καταγράφονται παρατηρήσεις από πραγματικές επιθέσεις και ρεαλιστικές επιδείξεις επιθέσεων από ομάδες ασφαλείας. Στην παρούσα έκδοση του πίνακα ATLAS[1] καταγράφονται 14 τακτικές και 56 τεχνικές επίθεσης. Οι τακτικές που έχουν εντοπισθεί συνοψίζονται στον ακόλουθο πίνακα:

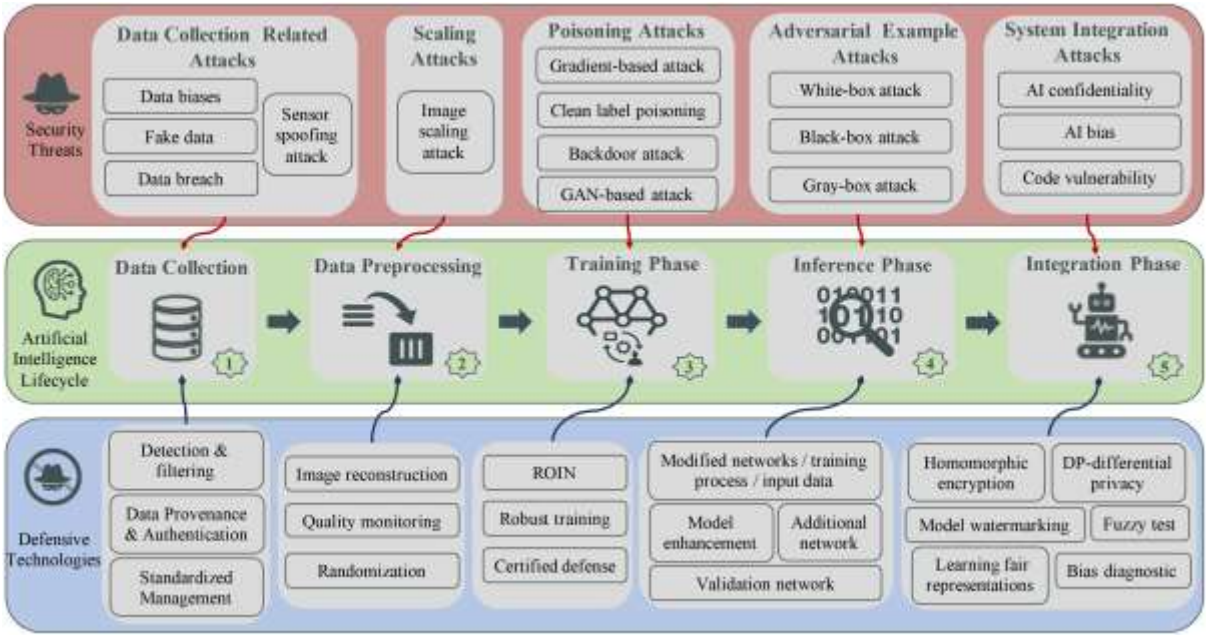
[1] https://atlas.mitre.org/

Τακτική Επίθεσης	Περιγραφή
Αναγνώριση	Ενεργητική ή παθητική συλλογή πληροφοριών του συστήματος ΤΝ για μελλοντική χρήση
Ανάπτυξη Πόρων	Δημιουργία, αγορά ή υποκλοπή/κλοπή πόρων που μπορούν να χρησιμοποιηθούν για την υποστήριξη της επίθεσης. Τέτοιοι πόροι περιλαμβάνουν αντικείμενα μηχανικής μάθησης, υποδομές, λογαριασμούς ή δυνατότητες
Αρχική Πρόσβαση	Το σύστημα-στόχος μπορεί να είναι ένα δίκτυο, μια κινητή συσκευή ή μια περιφερειακή συσκευή, όπως μια πλατφόρμα αισθητήρων
Πρόσβαση στο Μοντέλο Μηχανικής Μάθησης	Απόκτηση πρόσβασης με σκοπό τη συλλογή πληροφοριών, την εκτέλεση επιθέσεων ή την εισαγωγή δεδομένων στο μοντέλο
Εκτέλεση Κακόβουλου Κώδικα	Εκτέλεση κώδικα του επιτιθέμενου σε ένα τοπικό ή μεμακρυσμένο σύστημα
Παραμονή	Διατήρηση πρόσβασης στο σύστημα κατά την επανεκκίνηση, την αλλαγή διαπιστευτηρίων και άλλες διακοπές
Κλιμάκωση Προνομίων	Απόκτηση δικαιωμάτων υψηλότερου επιπέδου σε ένα σύστημα ή δίκτυο
Αποφυγή Άμυνας	Αποφυγή ανίχνευσης από λογισμικό ασφαλείας
Πρόσβαση σε διαπιστευτήρια	Κλοπή ονομάτων λογαριασμών και κωδικών πρόσβασης
Αποκάλυψη	Απόκτηση γνώσης για το σύστημα και το εσωτερικό δίκτυο
Συλλογή	Συγκέντρωση αντικειμένων μηχανικής μάθησης ή πληροφοριών
Σταδιακή Επίθεση στο Μοντέλο Μηχανικής Μάθησης	Αξιοποίηση γνώσης και πρόσβασή στο σύστημα-στόχο για προσαρμογή της επίθεσης
Εξαγωγή	Κλοπή αντικειμένων μηχανικής μάθησης ή πληροφοριών

Τακτική Επίθεσης	Περιγραφή
Επίδραση	Χειραγώγηση, διακοπή, υπονόμευση εμπιστοσύνης ή καταστροφή συστήματος ΤΝ

Πίνακας 3.1: Τακτικές Επίθεσης κατά ATLAS

Με βάση τις τακτικές και τεχνικές επίθεσης που έχουν καταγραφεί στη βάση γνώσης ATLAS έχει αναπτυχθεί ένα πλαίσιο στρατηγικών επίθεσης στις φάσεις του κύκλου ζωής ενός συστήματος ΤΝ που φαίνεται στο σχήμα που ακολουθεί [28]. Για την περιγραφή των κινδύνων χρησιμοποιείται ένας κύκλος ζωής συστήματος ΤΝ που περιλαμβάνει 5 φάσεις: συλλογή δεδομένων, προεπεξεργασία δεδομένων, εκπαίδευση μοντέλων, εξαγωγή συμπερασμάτων και ολοκλήρωση του συστήματος.



Σχήμα 3.1: Επιθετικές και Αμυντικές Στρατηγικές Κυβερνοασφάλειας ΤΝ

Στη φάση της συλλογής υφίσταται ο κίνδυνος επίθεσης εξαπάτησης αισθητήρων, όταν χρησιμοποιούνται συσκευές υλικού ή υποκλοπής και παραποίησης δεδομένων, όταν χρησιμοποιούνται συστήματα λογισμικού. Στη φάση της προεπεξεργασίας έχουν καταγραφεί επιθέσεις κλιμάκωσης (scaling attacks), οι οποίες μέσα από μικρές αλλαγές στα δεδομένα εισόδου επιτυγχάνουν το μοντέλο να κάνει λάθος προβλέψεις. Στη φάση εκπαίδευσης, επιχειρείται τροποποίηση των δεδομένων ή του ίδιου του μοντέλου (data poisoning/model poisoning) με αποτέλεσμα το μοντέλο είτε να

αποτυγχάνει την εκπαίδευσή του είτε να προκαλούνται λανθασμένες απαντήσεις. Προσβάλλεται δηλαδή η διαθεσιμότητα (availability) ή η ακεραιότητα (integrity) του μοντέλου. Επιπλέον οι επιθέσεις poisoning μπορεί να προκαλούν παραβιάσεις εμπιστευτικότητας (confidentiality), αυθεντικοποίησης (authentication), εξουσιοδότησης (authorization), και ιδιωτικότητας (privacy) [29]. Η εισαγωγή των τροποποιημένων δεδομένων επιχειρείται μέσω trojan, δηλαδή κακόβουλου κώδικα που παρουσιάζεται ως νομότυπο λογισμικό και backdoor, δηλαδή μη εξουσιοδοτημένη πρόσβαση μέσω παράκαμψης της κρυπτογράφησης και της αυθεντικοποίησης ενός συστήματος. Στη φάση της εξαγωγής συμπερασμάτων επιχειρείται να υποβαθμιστούν ή να παρεμβληθούν τα συμπεράσματα του μοντέλου με την τροφοδότησή του με ελαφρώς τροποποιημένα δεδομένα τα οποία όμως μπορούν να οδηγήσουν σε αντίθετα συμπεράσματα (adversarial examples). Στη φάση της ολοκλήρωσης του συστήματος, η επιφάνεια επίθεσης επεκτείνεται καθώς τα μοντέλα ενσωματώνονται σε συσκευές υλικού ή είναι προσβάσιμα μέσω δικτύων. Οι επιθέσεις επικεντρώνονται στην παραβίαση της εμπιστευτικότητας μοντέλων και δεδομένων εκπαίδευσης, σε ευπάθειες του κώδικα που καλεί το μοντέλο ΤΝ για εκτέλεση εργασιών και στην παραγωγή μεροληπτικών απαντήσεων από το μοντέλο μέσω της τροποποίησης των εισόδων – ερωτημάτων.

Από πλευράς αντικτύπου, οι κακόβουλοι χρήστες που αποκτούν πρόσβαση σε ένα μοντέλο ΤΝ έχουν τις εξής επιλογές [29]:

- Αλλοίωση της λογικής των αλγορίθμων (logic corruption), που οδηγεί σε ανεπιθύμητα αποτελέσματα και αποφάσεις.
- Χειραγώγηση δεδομένων (data manipulation), με την τροποποίηση των ετικετών των δεδομένων εκπαίδευσης και λοιπών παραμέτρων μάθησης.
- Εισαγωγή δεδομένων (data injection), όπου εισάγονται νέα δεδομένα στο dataset με σκοπό να δημιουργήσουν μεροληψία του αλγορίθμου ή λανθασμένα αποτελέσματα.
- Μεταφορά μάθησης (transfer learning), όπου επιχειρείται η εισαγωγή ενός προεκπαιδευμένου μοντέλου για την αλλαγή της συμπεριφοράς του μοντέλου που εκπαιδεύεται. Επιπλέον ο επιτιθέμενος μπορεί να αποσπάσει ένα εκπαιδευμένο μοντέλο για να βελτιώσει ένα δικό του.

Για τα μεγάλα γλωσσικά μοντέλα εμφανίζονται εξειδικευμένες απειλές όπως οι επιθέσεις κερκόπορτας (backdoor) [30], prompt injection και jailbreaking [31]. Στις επιθέσεις backdoor ο επιτιθέμενος έχει ως στόχο να χειραγωγήσει το μοντέλο-στόχο δηλητηριάζοντας τα δεδομένα εκπαίδευσής του, αναγκάζοντάς το να επιστρέφει το επιθυμητό αποτέλεσμα όταν δίνεται μια συγκεκριμένη είσοδος (trigger), ενώ λειτουργεί κανονικά με διαφορετικές εισόδους. Οι προτροπές συστήματος (system prompts) είναι οδηγίες που δίνονται στο μοντέλο από το σύστημα (χωρίς να είναι ορατές στο χρήστη) και αποσκοπούν στην αποτροπή ανεπιθύμητης συμπεριφοράς. Οι εντολές αυτές μπορούν να ανακτηθούν από τους χρήστες και να παρεμβληθούν με κακόβουλες εντολές οι οποίες παρακάμπτουν τις εντολές του συστήματος. Με αυτό τον τρόπο ακόμα και ένα μοντέλο που δεν έχει υποστεί data poisoning μπορεί να επιστρέψει κακόβουλες απαντήσεις. Το jailbreaking αναφέρεται στην πρακτική τροφοδότησης ερωτημάτων που προκαλούν ανεπιθύμητη συμπεριφορά του μοντέλου. Σε αντίθεση με το prompt injection, το jailbreaking δεν απαιτεί απαραίτητα από τον εισβολέα να έχει πρόσβαση στις προτροπές συστήματος του μοντέλου. Στην περίπτωση αυτή χρησιμοποιούνται κατάλληλα ερωτήματα στο μοντέλο τα οποία επιφέρουν απαντήσεις οι οποίες αποκαλύπτουν προσωπικές ή εμπιστευτικές πληροφορίες ή παρακάμπτουν τα system prompts [32].

Από την πλευρά του αμυνόμενου υφίστανται αντίμετρα για την αντιμετώπιση των διαφορετικών τύπων επιθέσεων. Στο στάδιο της συλλογής υφίστανται υλοποιήσεις αυτοματοποιημένου ελέγχου της ποιότητας των δεδομένων και επιβολής φίλτρων ώστε να απορρίπτονται δεδομένα που είναι ύποπτα. Επίσης με μηχανισμούς αυθεντικοποίησης μπορεί να περιορισθεί η συλλογή δεδομένων από μη αξιόπιστες πηγές αλλά και δεδομένων που έχουν τροποποιηθεί. Για να περιορισθούν ανθρώπινα λάθη, η συλλογή πρέπει να ακολουθεί κάποιο πρότυπο και μια τυποποιημένη διαδικασία για να περιορισθούν οι πιθανότητες συλλογής από αμφισβητήσιμες πηγές. Στο στάδιο της προεπεξεργασίας, η τυχαιοποίηση δεδομένων μπορεί να αντιμετωπίσει επιθέσεις που βασίζονται στα χαρακτηριστικά των δεδομένων εισόδου. Για παράδειγμα όταν τα δεδομένα εκπαίδευσης περιλαμβάνουν εικόνες, η αλλαγή μεγέθους με τυχαίο τρόπο και η προσθήκη μηδενικών σε τυχαία σημεία (zero padding) μπορεί να καταστήσει ανεπιτυχείς επιθέσεις τροποποίησης δεδομένων εισόδου. Στο στάδιο της εκπαίδευσης εφαρμόζονται τακτικές εξυγίανσης δεδομένων (data

sanitization) με τις οποίες απορρίπτονται από το μοντέλο δεδομένα τα οποία έχουν σημαντικά αρνητικό αντίκτυπο στον ταξινομητή. Στη φάση εξαγωγής συμπερασμάτων εφαρμόζονται τακτικές ανταγωνιστικής εκπαίδευσης (adversarial training) δηλαδή η εκπαίδευση του μοντέλου τόσο με κανονικά όσο και με ανταγωνιστικά δεδομένα ώστε το μοντέλο να μάθει από τα λάθη του. Κατά την ολοκλήρωση του συστήματος έχουν εφαρμογή το σύνολο σχεδόν των αντιμέτρων κυβερνοασφάλειας από τη φυσική ασφάλεια μέχρι τη χρήση κρυπτογραφικών μεθόδων.

Ειδικά για την ασφάλεια των μεγάλων γλωσσικών μοντέλων, υφίστανται τεχνικές για την αντιμετώπιση των επιθέσεων prompt injection τόσο κατά τη σχεδίαση του συστήματος όσο και κατά τη λειτουργία του [33]. Αποφυγή προτροπών συστήματος που είναι εξαιρετικά σύντομες, χρήση μοναδικών και σύνθετων δομών προτροπών, εφαρμογή τεχνικών επικύρωσης εισόδου/εξόδου για το φιλτράρισμα πιθανών μοτίβων επίθεσης, τακτική ενημέρωση και τροποποίηση των προτροπών για τη μείωση της πιθανότητας να ανακαλυφθούν και να αξιοποιηθούν από επιτιθέμενους, είναι μέτρα τα οποία μπορούν να ληφθούν κατά τη σχεδίαση ενός συστήματος. Μια τεχνική ασφαλείας που εφαρμόζεται κατά τη λειτουργία ενός συστήματος είναι το prompt chaining. Η συγκεκριμένη τεχνική αφορά τη χρήση μιας εξόδου ενός LLM ως είσοδο σε ένα άλλο LLM, προκειμένου να διεκπεραιωθεί μια πιο σύνθετη ή πολλαπλών βημάτων εργασία. Με αυτό τον τρόπο αξιοποιούνται οι δυνατότητες πολλαπλών LLM και επιτυγχάνονται αποτελέσματα που δεν θα ήταν δυνατά με ένα μόνο γλωσσικό μοντέλο. Η συγκεκριμένη τεχνική αποτελεί και δικλείδα ασφαλείας, καθώς διαχωρίζοντας μια εργασία σε ξεχωριστά βήματα, καθιστά πιο δύσκολο για έναν εισβολέα να εισάγει κακόβουλο περιεχόμενο στην τελική έξοδο.

Καθίσταται σαφές ότι το πεδίο των επιθέσεων στα συστήματα ΤΝ είναι συνεχώς εξελισσόμενο όπως είναι και το πεδίο των αμυντικών μηχανισμών. Ο ENISA προτείνει 37 αντίμετρα τα οποία μπορούν να εφαρμοσθούν σε διάφορες φάσεις του κύκλου ζωής ενός συστήματος ΤΝ και χωρίζονται σε 3 κατηγορίες [20]:

- Οργανωτικά και πολιτικής, τα οποία αφορούν τα συνήθη αντίμετρα κυβερνοασφάλειας πληροφοριακών συστημάτων ή συνδέονται με πολιτικές ασφαλείας.

- Τεχνικά, τα οποία αφορούν την επιλογή μοντέλων και δεδομένων εκπαίδευσης.
- Εξειδικευμένα, τα οποία αφορούν τεχνικές εξουδετέρωσης επιθέσεων σε μοντέλα ΤΝ, όπως παρουσιάστηκαν παραπάνω.

3.2. ΤΝ ως εργαλείο Κυβερνοεπιθέσεων

Η επίδραση της ΤΝ στο τοπίο απειλών κυβερνοασφάλειας έχει τρεις μορφές: επέκταση υφιστάμενων απειλών μέσω των επαυξημένων δυνατοτήτων που παρέχει η ΤΝ, εισαγωγή νέων απειλών μέσω της εκμετάλλευσης αδυναμιών συστημάτων ΤΝ ή τρόπων επίθεσης που δεν ήταν εφικτοί παλαιότερα και αλλαγή του τυπικού χαρακτήρα των απειλών με αυξημένη αποτελεσματικότητα, στόχευση και δυσκολία απόδοσης σε κάποιο συγκεκριμένο δρώντα [34]. Οι απειλές εκτός από τον τομέα της ψηφιακής ασφάλειας επιδρούν και στη φυσική ασφάλεια αλλά και σε επίπεδο πολιτικής όπως θα αναλυθεί στα επόμενα κεφάλαια.

Καθώς η ΤΝ βασίζεται στα δεδομένα, η κυβερνοασφάλεια όπως αναλύθηκε στην προηγούμενη ενότητα περιστρέφεται γύρω από την ασφάλεια των δεδομένων. Αναφορικά με την επέκταση υφιστάμενων απειλών, η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για την απόκτηση μη εξουσιοδοτημένης πρόσβασης σε δίκτυα, για την ανάκτηση δεδομένων και πληροφοριών που μπορούν να χρησιμοποιηθούν για τη δημιουργία κακόβουλου λογισμικού που δεν ανιχνεύεται από μηχανισμούς ασφαλείας και για την πρόκληση ζημιάς στις υποδομές. Μπορεί επίσης να χρησιμοποιηθεί ως εργαλείο spear-phishing για την απόκτηση δεδομένων από επιλεγμένα άτομα για διάφορους σκοπούς. Άλλες εφαρμογές, αφορούν εργαλεία ΤΝ για την παραγωγή επιθέσεων sql injection, αυτοματοποίηση επιθέσεων με το πλαίσιο Metasploit αλλά και προηγμένων επιθέσεων (APT) [35].

Οι κυβερνοεπιθέσεις που υποστηρίζονται από ΤΝ έχουν 3 χαρακτηριστικά [13]: διάχυση, διεισδυτικότητα και προσαρμοστικότητα. Όσον αφορά στο πρώτο χαρακτηριστικό, ένα malware υποστηριζόμενο από ΤΝ, μπορεί για παράδειγμα να χρησιμοποιήσει μια εφαρμογή video conference για να εντοπίσει το στόχο με αναγνώριση προσώπου, φωνής ή γεωεντοπισμό και να μην είναι ανιχνεύσιμο από ένα λογισμικό προστασίας. Η διάχυση των συστημάτων ΤΝ σηματοδοτείται από τη

ραγδαία αύξηση των έξυπνων συσκευών με τεχνητή νοημοσύνη, οι οποίες μπορούν να αναγνωρίζουν και να αντιδρούν σε εικόνες, ήχους και άλλα μοτίβα. Οι μηχανές μαθαίνουν όλο και περισσότερα κατά τη χρήση τους και προσαρμόζονται στις μεταβαλλόμενες καταστάσεις με αποτέλεσμα να είναι σε θέση να προβλέπουν τα αποτελέσματα. Η εκμετάλλευση της παρουσίας των συστημάτων αυτών δημιουργεί νέες επιφάνειες επίθεσης. Σε θέματα προσαρμοστικότητας, όπως ένας άνθρωπος - χάκερ μαθαίνει από τα λάθη του και αναζητά νέους τρόπους διείσδυσης σε ένα σύστημα, ένα σύστημα ΤΝ προσανατολισμένο στο hacking μπορεί να βελτιώσει την απόδοσή του.

Μια νέου τύπου απειλή που αναδύεται μέσω της χρήσης της ΤΝ και έχει πολλαπλές χρήσεις είναι τα deepfakes, δηλαδή η δημιουργία ή η χειραγώγηση περιεχομένου μέσω ΤΝ, συχνά με τη μορφή βίντεο, ηχογραφήσεων ή εικόνων. Για τη δημιουργία deepfakes χρησιμοποιούνται δύο τεχνικές. Το face-swapping, επικεντρώνεται την αντικατάσταση του προσώπου σε κάποιο βίντεο ή εικόνα με τη χρήση αλγορίθμων ΤΝ. Για την εκπαίδευση των αλγορίθμων απαιτούνται χιλιάδες εικόνες των δύο προσώπων για να εντοπισθούν ομοιότητες και να γίνει η αντικατάσταση. Η δεύτερη τεχνική είναι τα δημιουργικά ανταγωνιστικά δίκτυα (GAN). Ένας αλγόριθμος, η γεννήτρια (generator), τροφοδοτείται με τυχαία δεδομένα και παράγει μια νέα εικόνα. Ο δεύτερος αλγόριθμος, ο διαχωριστής (discriminator), ελέγχει την εικόνα και τα δεδομένα για να δει αν αντιστοιχούν σε γνωστά δεδομένα (δηλαδή σε γνωστές εικόνες ή πρόσωπα). Αυτός ο ανταγωνισμός μεταξύ των δύο αλγορίθμων, ουσιαστικά καταλήγει να αναγκάζει τη γεννήτρια να δημιουργεί εξαιρετικά ρεαλιστικές εικόνες (π.χ. διασημοτήτων) που προσπαθούν να ξεγελάσουν τον διαχωριστή. Τα deepfakes χρησιμοποιούνται από το κοινό έγκλημα (πορνογραφία, εκβιασμοί, εξαπάτηση βιομετρικών συστημάτων) μέχρι πολιτικούς εκβιασμούς και στρατιωτικές χρήσεις (εξαπάτηση συστημάτων δορυφορικής τηλεπισκόπησης).

Η παραγωγική ΤΝ έχει δώσει τη δυνατότητα ανάπτυξης εργαλείων που βασίζονται σε μεγάλα γλωσσικά μοντέλα. Οι απειλές που απορρέουν από τα LLM περιλαμβάνουν την άμεση κακόβουλη χρήση των εξόδων τους για απάτη, πλαστοπροσωπία ή δημιουργία κακόβουλου λογισμικού, αλλά και μέσω πράξεων χειραγώγησης του μοντέλου (π.χ. μέσω δηλητηρίασης των δεδομένων) [31]. Πιο συγκεκριμένα, μοντέλα

γενικής χρήσης όπως το ChatGPT αλλά και εξειδικευμένα μοντέλα όπως το FraudGPT και το WormGPT μπορούν να παράγουν από καμπάνιες phishing που μιμούνται αξιόπιστους οργανισμούς μέχρι επιθέσεις deepfake που υποδύονται γνωστές προσωπικότητες [36]. Σε γενικές γραμμές τα LLM που είναι ελεύθερα διαθέσιμα στο κοινό έχουν δικλείδες ασφαλείας ώστε να μην επιστρέφουν απαντήσεις οι οποίες μπορούν να χρησιμοποιηθούν για κακόβουλη χρήση. Η διάθεση LLM χωρίς τους συγκεκριμένους περιορισμούς ή η παράκαμψη των περιορισμών που έχουν τα ελεύθερα διαθέσιμα LLM μπορεί να τα μετατρέψει σε εργαλεία επιθέσεων. Πρόσφατη έρευνα [37] κατέδειξε ότι είναι εφικτή η δημιουργία backdoors κατά την εκπαίδευση γλωσσικών μοντέλων τα οποία δεν εντοπίζονται από τις δικλείδες ασφαλείας ενώ τεχνικές όπως το adversarial training αντί να εντοπίζουν τα backdoor βοηθούν τα μοντέλα να τα αποκρύπτουν καλύτερα.

Παραδείγματα LLM με κακόβουλη συμπεριφορά αποτελούν τα FraudGPT, WormGPT και PoisonGPT. Το FraudGPT, είναι ένα συνδρομητικό εργαλείο δημιουργίας αληθοφανών email και ιστοσελίδων μέσω των οποίων μπορούν οι χρήστες να παραπλανηθούν και να εισάγουν προσωπικά τους στοιχεία ή να κατεβάσουν κακόβουλο λογισμικό στον υπολογιστή τους. Καθώς δεν έχει γίνει γνωστό σε ποιο LLM στηρίζεται, είναι δυσκολότερη η ανάλυση των δυνατοτήτων του και ο εντοπισμός επιθέσεων που βασίζονται σε αυτό. Ταυτόχρονα όμως αποτελεί ένα εργαλείο που μπορεί να επιτρέψει σε επίδοξους χάκερ να οργανώσουν αποτελεσματικές επιθέσεις, χωρίς ιδιαίτερες τεχνικές γνώσεις [38]. Το WormGPT βασίζεται στο ανοικτού κώδικα και ελεύθερα διαθέσιμο μεγάλο γλωσσικό μοντέλο GPT-J 6B και εξειδικεύεται στη δημιουργία κακόβουλου κώδικα. Καθώς είναι ανοικτού κώδικα, κακόβουλοι χρήστες μπορούν να το προσαρμόσουν στις ανάγκες τους. Μια ακόμα απειλή από τη χρήση LLM είναι η παραπληροφόρηση είτε μέσω της δημιουργίας λανθασμένων ή μερικώς σωστών αποτελεσμάτων (misinformation) είτε μέσω της σκόπιμης δημιουργίας ψεύτικων αποτελεσμάτων (disinformation). Το PoisonGPT είναι ένα κακόβουλο μοντέλο που έχει σχεδιαστεί για να διαδίδει στοχευμένες ψευδείς πληροφορίες [39]. Πρόκειται για το αποτέλεσμα της έρευνας μιας ομάδας ασφαλείας η οποία απέδειξε πως ένα ελεύθερα διαθέσιμο LLM μπορεί να τροποποιηθεί ώστε να παράγει ψευδείς πληροφορίες και να ανέβει σε ένα νόμιμο αποθετήριο και στη συνέχεια να

χρησιμοποιηθεί σε άλλες εφαρμογές χωρίς να είναι γνωστό στους χρήστες του ότι έχει παραβιαστεί η αξιοπιστία του [40].

Οι μορφές κυβερνοεπίθεσης που βασίζονται στην ΤΝ αφορούν όλους τους κακόβουλους δρώντες από hacktivists και black-hat hackers μέχρι εγκληματικές και τρομοκρατικές ομάδες και ένοπλες δυνάμεις και κρατικές υπηρεσίες πληροφοριών [41]. Στην περίπτωση τρομοκρατών ή κρατικών δυνάμεων, η ΤΝ μετατρέπεται σε κυβερνοόπλο, δημιουργώντας νέες μορφές απειλών. Μια μορφή είναι η επιτήρηση και ο εξαναγκασμός [26]. Πέρα από την απλή ανάπτυξη ενός όπλου τεχνητής νοημοσύνης για την παρακολούθηση ενός στοχοποιημένου προσώπου, οι αντίπαλοι αναμένεται να χρησιμοποιήσουν την τεχνητή νοημοσύνη για να ανακαλύψουν νέους υποψήφιους στόχους για εξαναγκασμό. Εκτός από πρόσωπα, η παρακολούθηση μπορεί να αποσκοπεί στον εντοπισμό όλων των αδύναμων σημείων ενός οργανισμού-στόχου: - ανθρώπους, τεχνολογίες, συστήματα ή συνδυασμό αυτών. Οι πληροφορίες στη συνέχεια μπορούν να χρησιμοποιηθούν για τη στρατολόγηση ενός ατόμου, είτε με τη βούληση του, είτε μέσω εκβιασμού προς όφελος του επιτιθέμενου.

3.3. ΤΝ ως εργαλείο Κυβερνοασφάλειας και Κυβερνοάμυνας

Μια από τις εφαρμογές της ΤΝ είναι και η κυβερνοασφάλεια. Με την εκθετική αύξηση των κυβερνοεπιθέσεων και την διάχυση έξυπνων συσκευών και δικτύων, οι ανάγκες αποδοτικότερων μεθόδων κυβερνοασφάλειας και κυβερνοάμυνας, αν η πρώτη αποτύχει, είναι επιτακτικές. Ο άμεσος αντίκτυπος των κυβερνοεπιθέσεων μεγάλης κλίμακας, η αυξανόμενη πολυπλοκότητα των πληροφοριακών συστημάτων αλλά και το έλλειμα εξειδικευμένου προσωπικού κυβερνοασφάλειας είναι παράγοντες που ενισχύουν την ανάγκη υλοποίησης λύσεων που βασίζονται σε ΤΝ [13].

Η ΤΝ μπορεί να βοηθήσει στην αυτοματοποίηση της ανίχνευσης και ανάλυσης απειλών, στην αποδοτικότερη και αμεσότερη αντιμετώπιση κυβερνοεπιθέσεων, στην ταχύτερη και ακριβέστερη εκτίμηση επικινδυνότητας και στην προστασία τερματικών συσκευών [42]. Στο γενικότερο πλαίσιο, η χρήση ΤΝ μπορεί να ενισχύσει τα υπάρχοντα συστήματα κυβερνοασφάλειας στο επίπεδο της αποτροπής και προστασίας, στο επίπεδο της αναγνώρισης και εντοπισμού απειλών και στο επίπεδο της αντιμετώπισης επιθέσεων [43]. Στην πρώτη περίπτωση μοντέλα ΤΝ μπορούν να

εκπαιδεύονται από αναδυόμενες απειλές και να αναπτύσσουν μηχανισμούς προστασίας με βάση την εκπαίδευσή τους. Στη δεύτερη περίπτωση μοντέλα ΤΝ μπορούν να αναγνωρίζουν πρότυπα απειλών κατά την εκπαίδευσή τους με τα οποία μπορούν στη συνέχεια να εντοπίσουν παραβιάσεις ή κενά ασφαλείας σε πληροφοριακά συστήματα. Τέλος η ΤΝ έχει τη δυνατότητα να αναλύει μεγάλο όγκο δεδομένων που προκύπτουν κατά την εξέλιξη μιας κυβερνοεπίθεσης και να προτείνει τα καταλληλότερα μέτρα θεραπείας.

Ειδικότερα τεχνικές και αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για τις παρακάτω εργασίες κυβερνοασφάλειας [44], [45]:

- Ανίχνευση και ταξινόμηση εισβολών
- Ανάλυση ανίχνευσης εισβολών
- Ταξινόμηση επιθέσεων
- Ανίχνευση και ανάλυση επιθέσεων DDoS
- Ανίχνευση ανωμαλιών
- Μείωση του ποσοστού ψευδών συναγερμών
- Ανάλυση κακόβουλης συμπεριφοράς
- Ταξινόμηση κακόβουλου λογισμικού
- Διαχείριση ασφαλείας
- Πληροφορίες απειλών (threat intelligence)
- Εκπαίδευση και ευαισθητοποίηση σε θέματα ασφάλειας

Επιπλέον για νέους τύπους απειλών που βασίζονται στην ΤΝ, οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν και σαν αντίμετρο. Για παράδειγμα μοντέλα βαθιάς μάθησης μπορούν να εκπαιδευθούν από διαθέσιμα δεδομένα για την αποτελεσματική ανίχνευση πλαστογραφίας στις περιπτώσεις deepfakes [46]. Στην περίπτωση αυτή βέβαια απαιτείται η ύπαρξη προεκπαιδευμένων μοντέλων κάτι το οποίο είναι εφικτό για μικρό αριθμό προσώπων (π.χ. διάσημοι, πολιτικοί κλπ.), είναι όμως πρακτικά ανέφικτο για το γενικό πληθυσμό. Για το λόγο αυτό, εκτός από την ανίχνευση deepfake με βάση τα χαρακτηριστικά, αναπτύσσονται και προσεγγίσεις που δεν εξαρτώνται από χαρακτηριστικά καθώς και προσεγγίσεις βασισμένες σε πολιτικές.

Η ΤΝ μπορεί να συμβάλει στην ευρωστία, ανταπόκριση και ανθεκτικότητα (3R: robustness, response, resilience) των πληροφοριακών συστημάτων [47]. Η ευρωστία συνδέεται άμεσα με την επιχειρησιακή συνέχεια (business continuity) καθώς μέσω της ΤΝ μπορούν να αναπτυχθούν μηχανισμοί αυτοελέγχου και αυτοδιόρθωσης ενός συστήματος σε περίπτωση σφάλματος με αποτέλεσμα να διατηρηθεί σε λειτουργία χωρίς ανάγκη ανθρώπινης παρέμβασης. Η ανταπόκριση αναφέρεται στην αντίδραση σε μια κυβερνοεπίθεση η οποία με τη βοήθεια της ΤΝ μπορεί να είναι περισσότερο ενεργητική και να μετατραπεί σε αντεπίθεση. Η ανθεκτικότητα αναφέρεται στην έγκαιρη ανίχνευση ανωμαλιών ώστε να αποτραπεί μια επίθεση πριν προκαλέσει καταστροφικά αποτελέσματα.

Οι πιο διαδεδομένες εφαρμογές της ΤΝ στην κυβερνοασφάλεια αφορούν τη χρήση σε συστήματα εντοπισμού εισβολών (Intrusion Detection System - IDS), εντοπισμό spam (ανεπιθύμητων μηνυμάτων) και αναγνώριση και ταυτοποίηση κακόβουλου λογισμικού [48]. Ο εντοπισμός εισβολών μπορεί να επιτευχθεί με βάση γνωστές επιθέσεις (exploit-based), εντοπισμό μη ομαλής συμπεριφοράς (anomaly-based) ή με συνδυασμό αυτών (hybrid-based). Κατάλληλοι αλγόριθμοι είναι τόσο αλγόριθμοι επιβλεπόμενης μάθησης (δέντρα απόφασης, SVM) όσο και νευρωνικά δίκτυα. Για τον εντοπισμό spam καταλληλότεροι είναι οι αλγόριθμοι ταξινόμησης.

Οι αλγόριθμοι βαθιάς μάθησης έχουν εφαρμογή σε συστήματα εντοπισμού επιθέσεων και έχουν αποδεδειγμένη αποτελεσματικότητα σε μια μεγάλη γκάμα επιθέσεων όπως επιθέσεις άρνησης υπηρεσιών (DoS), zero-day, ανίχνευσης, phishing, επιθέσεις ακεραιότητας κ.ά. [49]. Επιπλέον βρίσκουν εφαρμογή σε εντοπισμό απάτης (fraud detection) σε επικοινωνιακά και τραπεζικά συστήματα [50].

Συνοπτικά, οι επιμέρους προσεγγίσεις, τεχνικές και μοντέλα της ΤΝ μπορούν να χρησιμοποιηθούν για του σκοπούς που φαίνονται στον παρακάτω πίνακα [44], [46], [51], [52]:

Προσέγγιση ΤΝ	Αλγόριθμος	Σκοπός
Συστήματα Κανόνων		Συστήματα ανίχνευσης εισβολών στο δίκτυο
Επιβλεπόμενη Μάθηση	Δέντρα απόφασης	Ανάλυση κακόβουλης συμπεριφοράς Σύστημα ανίχνευσης εισβολών Σύστημα ανίχνευσης ανωμαλιών

Προσέγγιση ΤΝ	Αλγόριθμος	Σκοπός
	Απλή ταξινόμηση Bayes	Σύστημα ανίχνευσης εισβολών
	Μηχανές Διανυσμάτων Υποστήριξης	Ταξινόμηση επίθεσης Ανίχνευση και ταξινόμηση εισβολών Ανίχνευση και ανάλυση DDoS Συστήματα ανίχνευσης ανωμαλιών
Μη επιβλεπόμενη μάθηση		Ανάλυση ανίχνευσης εισβολής Ανάλυση κακόβουλου λογισμικού Ανίχνευση Spam
Μερικώς επιβλεπόμενη μάθηση		Ανίχνευση ανωμαλιών δικτύου
Ενισχυτική μάθηση		Ανίχνευση κακόβουλων δραστηριοτήτων και εισβολών
	Hidden Markov Model (HMM)	Σύστημα ανίχνευσης εισβολής
Μάθηση πολλαπλών καθηκόντων	Deep Neural Networks (DNN)	Ανίχνευση deep fake
Συλλογική μάθηση	Adaptive Boosting (ADABoost)	Ανίχνευση ανωμαλιών δικτύου Ανίχνευση εισβολής Ανάλυση κακόβουλου λογισμικού Ανίχνευση spam και phishing
	Random forests	Συστήματα ανίχνευσης εισβολών στο δίκτυο
Νευρωνικά δίκτυα	Deep Learning	Ανίχνευση ανωμαλιών Ανάλυση ανίχνευσης εισβολής Ταξινόμηση επιθέσεων Ταξινόμηση της κυκλοφορίας κακόβουλου λογισμικού Ανίχνευση spam Ανίχνευση απάτης
	ANN classifier	Ανίχνευση εισβολής Ανάλυση κακόβουλου λογισμικού Ανίχνευση spam και phishing
Μάθηση κατά περίπτωση	k-Nearest Neighbour	Σύστημα ανίχνευσης εισβολής στο δίκτυο Μείωση του ποσοστού ψευδών συναγερμών Σύστημα ανίχνευσης εισβολής

Πίνακας 3.2: Χρήσεις ΤΝ στον τομέα της κυβερνοασφάλειας

Εκτός από τα πλεονεκτήματα της ΤΝ στην κυβερνοασφάλεια υφίσταται και μια σειρά από προκλήσεις. Η ανάπτυξη συστημάτων ΤΝ είναι μια χρονοβόρα διαδικασία η οποία απαιτεί πολλούς πόρους και σε θέματα κόστους και σε θέματα υποδομών [53]. Οι περιορισμένοι πόροι είναι επίσης ένα σημαίνον ζήτημα σε περιβάλλοντα IoT.

Επιπλέον η αποτελεσματικότητα ενός συστήματος ΤΝ εξαρτάται από τα δεδομένα εκπαίδευσης του [54]. Ελλιπή ή μεροληπτικά δεδομένα δημιουργούν προκαταλήψεις και δυσλειτουργίες. Για παράδειγμα, επιθέσεις οι οποίες είναι σπάνιες δεν παρέχουν αρκετά δεδομένα για την εκπαίδευση των μοντέλων μηχανικής μάθησης. Τέλος, ο αριθμός των ψευδών εντοπισμών (false positives) μπορεί να μειώσει την αξιοπιστία ενός μοντέλου και απαιτεί επανεκπαίδευση [50].

Κατά συνέπεια συνεκτιμώντας και το πλήθος απειλών κυβερνοασφάλειας κατά την ανάπτυξη συστημάτων ΤΝ, ο ανθρώπινος παράγοντας εξακολουθεί να παίζει σημαντικό ρόλο στην αποτελεσματικότητα της χρήσης ΤΝ στην κυβερνοασφάλεια.

ΚΕΦΑΛΑΙΟ 4

ΕΦΑΡΜΟΓΕΣ ΤΝ ΣΕ ΑΣΦΑΛΕΙΑ ΚΑΙ ΆΜΥΝΑ

Η είσοδος της ΤΝ στον τομέα της άμυνας και της ασφάλειας έχει τη δυνατότητα να μετασχηματίσει υπάρχουσες δυνατότητες, συνοδεύεται όμως και από σημαντικές προκλήσεις. Οι τεχνολογίες τεχνητής νοημοσύνης έχουν βρει εφαρμογές σε διάφορους τομείς, υποσχόμενες να ενισχύσουν την αποτελεσματικότητα, την ακρίβεια και τις διαδικασίες λήψης στρατηγικών αποφάσεων των ενόπλων δυνάμεων και των υπηρεσιών επιβολής του νόμου.

4.1. Χρήση ΤΝ σε Περιβάλλοντα Ασφαλείας

Εφαρμογές της ΤΝ έχουν ήδη εξετασθεί σε λειτουργίες ασφαλείας. Η ΕΕ εξετάζει τη χρήση ΤΝ στον έλεγχο και ασφάλεια συνόρων [55]. Πιθανές εφαρμογές αποτελούν η βιομετρική ταυτοποίηση, η ανίχνευση συναισθημάτων, η εκτίμηση κινδύνου και η παρακολούθηση μετανάστευσης. Η βιομετρική ταυτοποίηση περιλαμβάνει την αναγνώριση προσώπου, φωνής και βαδίσματος και μπορεί να χρησιμοποιηθεί για σκοπούς επιτήρησης, εντοπισμού υπόπτων, εύρεσης αγνοουμένων και εγκληματολογικές έρευνες [56]. Η ανίχνευση συναισθημάτων είναι μια τεχνική επεξεργασίας φυσικής γλώσσας η οποία με βάση την είσοδο (κείμενο ή φωνή) εξάγει τα συναισθήματα του πομπού του μηνύματος. Στον τομέα της ασφάλειας συνόρων η συγκεκριμένη τεχνική μπορεί να συνεισφέρει στην αναγνώριση πιθανών απειλών, όπως και στην εκτίμηση της στάσης του κοινού απέναντι σε ένα γεγονός [57]. Η ανάλυση πρόβλεψης και η εκτίμηση κινδύνου στην περίπτωση της ασφάλειας συνόρων, υλοποιούνται με την ανάλυση ιστορικών δεδομένων διέλευσης συνόρων, αναρτήσεων μέσων κοινωνικής δικτύωσης και άλλων σχετικών πληροφοριών, με βάση τα οποία αλγόριθμοι ΤΝ μπορούν να συνεισφέρουν στον εντοπισμό πιθανών απειλών ή ανωμαλιών.

Ο συνδυασμός IoT και ΤΝ παρέχει νέες δυνατότητες επίγνωσης της κατάστασης σε αστυνομικές δυνάμεις στο πεδίο. Η «έξυπνη αστυνόμευση» (smart policing) συνδυάζει δεδομένα από αισθητήρες και κάμερες σε περιπολικά οχήματα, αστυνομικούς και εγκαταστάσεις με τη ροή πληροφοριών από κοινωνικά δίκτυα για να παρέχει

αναβαθμισμένες δυνατότητες διεύθυνσης επιχειρήσεων και επίγνωσης κατάστασης στις αστυνομικές δυνάμεις στο πεδίο. Ο ρόλος της ΤΝ αφορά στην πρόγνωση περιστατικών ασφαλείας, την έγκαιρη προειδοποίηση και την παροχή προτάσεων τρόπων ενεργείας στον διευθύνοντα της επιχείρησης ή και απευθείας στις δυνάμεις στο πεδίο [58]. Εκτός από το τακτικό επίπεδο, η έξυπνη αστυνόμευση έχει εφαρμογές και σε επιχειρησιακό και στρατηγικό επίπεδο. Με τη χρήση ιστορικών στοιχείων εγκληματικότητας εντοπίζονται τα θερμά σημεία (hotspots) σε μια περιοχή και γίνεται καλύτερη ανάθεση πόρων σε θέματα περιπολιών και επιτήρησης. Η ανάλυση εγκληματικών δικτύων μπορεί να εντοπίσει τις διασυνδέσεις εγκληματικών ή και τρομοκρατικών ομάδων για τον καθορισμό προτεραιοτήτων και την πρόβλεψη της πιθανότητας διάπραξης αξιόποινων πράξεων. Η παρακολούθηση των εντάσεων (tension monitoring) εφαρμόζει ένα μείγμα διαδικασιών επεξεργασίας φυσικής γλώσσας, τεχνικών βαθιάς μάθησης και ανάλυσης δικτύων σε δεδομένα κοινωνικών δικτύων για τη βαθμονόμηση της κοινωνικής συνοχής και τον εντοπισμό αιχμών έντασης και παραγόντων με ισχυρή επίδραση στο κοινό, με σκοπό τη διατήρηση της δημόσιας τάξης [59].

Μια άλλη εφαρμογή της ΤΝ στον τομέα της ασφάλειας είναι η προγνωστική αστυνόμευση (predictive policing) [60]. Πρόκειται για τη χρήση δυνατοτήτων ΤΝ για τη χωρική χαρτογράφηση των πιθανών περιοχών εγκλήματος, την αναγνώριση προτύπων εγκληματικής συμπεριφοράς και την ανάλυση υπόπτων. Η εφαρμογή των συγκεκριμένων τεχνικών συμβάλει στην αποτροπή εγκλημάτων, στην ταχύτερη και αποδοτικότερη κατανομή πόρων αλλά και στη βελτίωση της εμπιστοσύνης των πολιτών στις δυνάμεις ασφαλείας. Επιπλέον οι τεχνικές αυτές έχουν εφαρμογή στην οδική ασφάλεια. Σε αναλογία με τον εντοπισμό hotspot εγκληματικότητας, συστήματα ΤΝ μπορούν να αναλύσουν ιστορικά δεδομένα αλλά και δεδομένα πραγματικού χρόνου ώστε να προβλέψουν τροχαία ατυχήματα [61].

Η ανάλυση συμπεριφοράς (behavioral analysis) είναι μια από τις δυνατότητες της ΤΝ που βρίσκει εφαρμογές στην επιβολή του νόμου [62]. Με παρόμοιο τρόπο με τον οποίο οι συνήθειες, η ιδιοσυγκρασία και οι συμπεριφορές των ανθρώπων μπορούν να αναλυθούν από αλγόριθμους ΤΝ για λόγους μάρκετινγκ, μπορούν να συμβάλουν και στην πρόβλεψη, αναγνώριση, διερεύνηση και δίωξη εγκλημάτων και τρομοκρατικών

ενεργειών. Οι πηγές δεδομένων που μπορούν να τροφοδοτήσουν τους αλγόριθμους ΤΝ είναι δεκάδες και περιλαμβάνουν έξυπνες συσκευές (π.χ. wearables), προσωπικούς υπολογιστές, κινητά τηλέφωνα, ATM, PoS κλπ. Ωστόσο η συντριπτική πλειονότητα των δεδομένων, των συνομιλιών και των ενεργειών που δύναται να καταγραφούν είναι άσχετες με τους σκοπούς του νόμου, της τάξης και της ασφάλειας, και η μεγάλη πρόκληση είναι ο εντοπισμός των συγκεκριμένων πληροφοριών που θα υποδηλώνουν μια μακροπρόθεσμη ή βραχυπρόθεσμη απειλή. Με βάση τον συγκεκριμένο εντοπισμό επιλέγονται στη συνέχεια οι περιπτώσεις που συνιστούν άμεσο κίνδυνο για την έγκαιρη επέμβαση των δυνάμεων ασφαλείας. Οι συνεχώς αυξανόμενες τρομοκρατικές ενέργειες από «μοναχικούς λύκους», δηλαδή άτομα που ριζοσπαστικοποιούνται και αποφασίζουν τη διεξαγωγή μεμονωμένων επιθέσεων χωρίς να έχουν επιδείξει εγκληματική συμπεριφορά στο παρελθόν, καταδεικνύουν τη χρησιμότητα της συγκεκριμένης χρήσης της ΤΝ.

Η αστυνόμευση του σκοτεινού διαδικτύου (dark web) είναι μια ακόμα περίπτωση που η ΤΝ μπορεί να συνεισφέρει. Καθώς στο dark web δεν γίνεται ευρετηρίαση από μηχανές αναζήτησης, ενώ για την πρόσβαση χρησιμοποιούνται φυλλομετρητές ιστού που βασίζονται στην ανωνυμία, ο εντοπισμός κακόβουλων χρηστών είναι δυσχερής. Με δεδομένο ότι η παράνομη δραστηριότητα επικεντρώνεται στο σκοτεινό διαδίκτυο, το dark web αποτελεί χώρο δραστηριοποίησης των δυνάμεων ασφαλείας [63]. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να συνεισφέρουν μέσω της ανάλυσης δεδομένων και συμπεριφορών ώστε να συσχετισθεί δραστηριότητα χρηστών του σκοτεινού διαδικτύου με δραστηριότητα στο ανοικτό διαδίκτυο, αποκαλύπτοντας την ταυτότητά τους [64]. Παράδειγμα ενός τέτοιου συστήματος αποτελεί το MEMEX το οποίο χρησιμοποιείται από τουλάχιστον 30 υπηρεσίες ασφαλείας παγκοσμίως για την παρακολούθηση του σκοτεινού διαδικτύου και έχει συνεισφέρει μέχρι στιγμής στον εντοπισμό χιλιάδων περιπτώσεων εμπορίας ανθρώπων [65].

Όλες οι παραπάνω χρήσεις είναι πολλά υποσχόμενες, ωστόσο ταυτόχρονα ανακύπτουν ηθικά ζητήματα και ζητήματα ιδιωτικότητας τα οποία αναπτύσσονται στο επόμενο κεφάλαιο. Επιπρόσθετα σε ορισμένες περιπτώσεις αμφισβητείται η αποτελεσματικότητά τους. Σημαντική πρόκληση αποτελούν τα δεδομένα εκπαίδευσης τα οποία σε πολλές περιπτώσεις δεν επαρκούν ή δεν είναι αντιπροσωπευτικά. Ομοίως

δεν υπάρχουν επαρκή εμπειρικά δεδομένα για να καταγραφεί το άμεσο όφελος από τη χρήση των συγκεκριμένων τεχνικών, ενώ η νομοθεσία μπορεί να αποτελέσει περιοριστικό παράγοντα χρήσης τους.

4.2. Χρήση ΤΝ σε Στρατιωτικά Περιβάλλοντα

Στον στρατιωτικό τομέα η ΤΝ βρίσκει εφαρμογές σε όλα τα πεδία στρατιωτικών επιχειρήσεων (ξηρά, θάλασσα, αέρα, διάστημα, κυβερνοχώρο, πληροφορίες) και όλα τα επίπεδα (πολιτικό, στρατηγικό, επιχειρησιακό, τακτικό). Στο πολιτικο-στρατηγικό επίπεδο, η ΤΝ μπορεί να υποστηρίξει πληροφοριακές επιχειρήσεις παραπληροφόρησης και προπαγάνδας με σκοπό την αποσταθεροποίηση ενός κράτους. Εργαλεία του κυβερνοχώρου με την υποστήριξη της ΤΝ είναι ικανά να παράγουν υλικό παραπληροφόρησης (συμπεριλαμβανομένων των deepfakes) τόσο λεπτομερές και αληθοφανές και σε τέτοιες ποσότητες, ώστε να καθίσταται πολύ δύσκολη η διάκριση της αλήθειας [66]. Τα εργαλεία περιλαμβάνουν αυτοματοποιημένους λογαριασμούς (bots) και πλαστούς λογαριασμούς στα μέσα κοινωνικής δικτύωσης μέσω των οποίων επιχειρείται να υπονομεύσουν πολιτικό αφήγημα των αντιπάλων, να σπείρουν τη διχόνοια και τις διαμάχες εντός των χωρών και να θολώσουν τα όρια μεταξύ πραγματικών γεγονότων και μυθοπλασίας [67]. Είναι επίσης γνωστό ότι αλγόριθμοι ΤΝ έχουν χρησιμοποιηθεί σε προεκλογικές εκστρατείες για να επηρεάσουν τους ψηφοφόρους και να χειραγωγήσουν την κοινή γνώμη. Παραδείγματα αποτελούν τόσο οι προεδρικές εκλογές των ΗΠΑ το 2016 όσο και δημοψήφισμα για το Brexit και οι εκλογές του 2017 στο Ηνωμένο Βασίλειο [68].

Στο επιχειρησιακό/τακτικό επίπεδο η ΤΝ παρέχει νέες δυνατότητες συλλογής και επεξεργασίας πληροφοριών και υποστήριξης συστημάτων λήψης απόφασης. Η νέα γενιά αυτόνομων οχημάτων περιλαμβάνει μη επανδρωμένα χερσαία, θαλάσσια, υποθαλάσσια, εναέρια και διαστημικά συστήματα, τα οποία χρησιμοποιούν αλγόριθμους μηχανικής μάθησης για την πλοήγηση και τη συλλογή πληροφοριών. Η πλοήγηση με βάση την ΤΝ υποστηριζόμενη από πλήθος αισθητήρων όχι μόνο επιτρέπει στα μη επανδρωμένα οχήματα να κινούνται στο εχθρικό έδαφος, αλλά επιτρέπει ταυτόχρονα την εκτέλεση εξελιγμένων ελιγμών και τακτικών μάχης που σε συνδυασμό με εξελιγμένα μέσα προσβολής δίνουν τη δυνατότητα άμεσης προσαρμογής στους εχθρικούς ελιγμούς, εκμετάλλευσης ευκαιριών στο πεδίο της

μάχης και αναφοράς των μεταβαλλόμενων συνθηκών στη βάση ελέγχου. Τα ολοκληρωμένα συστήματα διαχείρισης μάχης, διοίκησης, ελέγχου, επικοινωνιών και πληροφοριών (BMC3I) μπορούν να παρέχουν νέες αμυντικές και επιθετικές δυνατότητες στους στρατιωτικούς διοικητές [69].

Οι βρετανικές ένοπλες δυνάμεις αναγνωρίζουν εφαρμογές της ΤΝ στην άμυνα σε 6 τομείς [70]. Ένας τομέας είναι η αναγνώριση προτύπων, όπου συστήματα ΤΝ μπορούν να εντοπίζουν αντικείμενα ενδιαφέροντος μέσω δεδομένων που συλλέγονται από αισθητήρες, Παραδείγματα χρήσης αποτελούν η ανάλυση δορυφορικών εικόνων, η ανάλυση εκπομπών ραδιοσυχνοτήτων αλλά και η απομάκρυνση θορύβου από συλλεγόμενα δεδομένα. Στον τομέα της ανάλυσης, η ΤΝ μπορεί να συνεισφέρει στην εξαγωγή γνώσης από μεγάλα σύνολα δομημένων και μη δομημένων δεδομένων. Παραδείγματα χρήσης αποτελούν η έξυπνη αναζήτηση πληροφοριών αλλά και η απλοποίηση και ο εναρμονισμός πολιτικών και κανονισμών με τον έλεγχο συμμόρφωσης με υπάρχουσα τεκμηρίωση. Στον τομέα της πρόβλεψης, δίνεται η δυνατότητα πρόβλεψης πιθανών αποτελεσμάτων με βάση ιστορικά δεδομένα. Παραδείγματα χρήσης αποτελούν ο υπολογισμός κλίμακας ανταλλακτικών και γεωγραφικής κατανομής τους με βάση ιστορικά δεδομένα βλαβών και ακινησιών υλικού. Στον τομέα της προσομοίωσης, η ΤΝ παρέχει τη δυνατότητα διερεύνησης εναλλακτικών σεναρίων και ανάλυσης δεδομένων για την ανάπτυξη εναλλακτικών τρόπων ενεργείας. Παράδειγμα χρήσης αποτελεί η υποβοήθηση της επιχειρησιακής σχεδίασης. Στο τομέα της δημιουργίας, εντοπίζεται η χρήση μεγάλων γλωσσικών μοντέλων σε αμυντικές εφαρμογές. Παράδειγμα χρήσης αποτελεί η μετάφραση κειμένου και ομιλίας σε πραγματικό χρόνο. Στον τομέα της λήψης αποφάσεων, η εφαρμογή της ΤΝ σε αυτόνομα συστήματα αφορά στη δημιουργία αυτοματοποιημένων συμπεριφορών για την επίτευξη ενός στόχου. Παραδείγματα χρήσης είναι αυτόνομα οχήματα μεταφοράς εφοδίων στο πεδίο της μάχης, αυτόνομα σκάφη εκκαθάρισης ναρκοπεδίων και αυτόνομα σμήνη drone σε αποστολές αναγνώρισης και εντοπισμού στόχων.

Σε αναλογία με το διαδίκτυο των πραγμάτων (IoT), στον στρατιωτικό τομέα αναπτύσσεται το διαδίκτυο των πραγμάτων στο πεδίο της μάχης (IoBT). Πρόκειται για ένα δίκτυο αισθητήρων, φορητών συσκευών και συσκευών IoT που κάνουν χρήση του

υπολογιστικού νέφους και της υπολογιστικής παρυφών (edge computing) για τη δημιουργία μιας συνεκτικής δύναμης μάχης, διασυνδέοντας, τον απλό μαχητή και τους χειριστές στα τεθωρακισμένα οχήματα και τα μέσα επικοινωνιών με τα κέντρα διοικήσεως και ελέγχου. Ενώ η υπολογιστική νέφους αποθηκεύει και επεξεργάζεται τα δεδομένα στο διαδίκτυο επιτρέποντας το διαμοιρασμό πόρων και μειώνοντας τις απαιτήσεις σε υπολογιστική ισχύ, η υπολογιστική παρυφών στοχεύει στην ελαχιστοποίηση της καθυστέρησης μέσω της τοπικής επεξεργασίας των δεδομένων. Η τοπική επεξεργασία είναι απαραίτητη σε συστήματα αντίδρασης πραγματικού χρόνου όπως τα συστήματα στοχοποίησης και επιτρέπει την αυτόνομη λειτουργία σε περιοχές με μειωμένη πρόσβαση στο διαδίκτυο. Την ίδια στιγμή η υπολογιστική νέφους είναι χρήσιμη σε εφαρμογές που διαχειρίζονται μεγάλο όγκο δεδομένων όπως συμβαίνει σε ένα σταθμό διοικήσεως. Κατά συνέπεια και οι δύο προσεγγίσεις έχουν θέση στο IoBT. Για το λόγο αυτό αναπτύσσεται η κατανεμημένη τεχνητή νοημοσύνη (Distributed AI - DAI), η οποία περιλαμβάνει την κατανομή των υπολογιστικών εργασιών σε πολλαπλούς κόμβους επιτρέποντας την ανάπτυξη και εκτέλεση αλγορίθμων τεχνητής νοημοσύνης σε συσκευές παρυφών (edge devices) [71]. Το IoBT έχει ορισμένα χαρακτηριστικά τα οποία δημιουργούν προκλήσεις που δεν υπάρχουν στο IoT [72]. Στο IoBT συναντάται μια ποικιλομορφία αποστολών και στόχων. Υφίστανται πολλαπλά δίκτυα που λειτουργούν ταυτόχρονα για την επίτευξη διαφορετικών στόχων, π.χ. επιτήρηση, εντοπισμό, προσβολή στόχων. Το περιβάλλον στο IoBT είναι δυναμικά μεταβαλλόμενο ενώ πολλές συσκευές του δικτύου έχουν περιορισμένους πόρους. Αισθητήρες και μη επανδρωμένα αεροσκάφη και οχήματα που λειτουργούν με συσσωρευτές έχουν περιορισμούς στο χρόνο λειτουργίας και στην κατανάλωση ενέργειας, ενώ άλλες μπορεί να απαιτείται να κινηθούν σε μεγάλες αποστάσεις και υπό συνθήκες παρεμβολών με αποτέλεσμα να χάσουν την επαφή με το σταθμό ελέγχου. Η ετερογένεια συστημάτων αποτελεί πρόκληση και για τους υπολογιστικούς πόρους, καθώς στο ίδιο δίκτυο απαιτείται η συνεργασία μικρών αισθητήρων με εναέριες και επίγειες πλατφόρμες που δεν έχουν ανάλογους περιορισμούς. Αντίστοιχη ετερογένεια υφίσταται και στο μέγεθος και την πυκνότητα του δικτύου. Ένα δίκτυο για παράδειγμα μπορεί να διασυνδέει ένα σμήνος μη επανδρωμένων αεροσκαφών με μεγάλη πυκνότητα και ποσότητα κόμβων και έναν ουλαμό αρμάτων που είναι διεσπαρμένα σε μια ευρεία περιοχή στο πεδίο της μάχης.

Η ΤΝ μπορεί να συνεισφέρει στην αντιμετώπιση των παραπάνω προκλήσεων σε πολλαπλά επίπεδα. Μπορεί να βελτιστοποιήσει τα δίκτυα επικοινωνίας προσαρμόζοντας δυναμικά τις διαμορφώσεις με βάση τις εξελισσόμενες συνθήκες του πεδίου μάχης, εξασφαλίζοντας αξιόπιστη και ασφαλή επικοινωνία. Επιπλέον μπορεί να βοηθήσει στην αποτελεσματική διαχείριση και κατανομή των συχνοτήτων επικοινωνίας, μειώνοντας τις παρεμβολές και ενισχύοντας την αξιοπιστία της επικοινωνίας. Μπορεί να βοηθήσει στο συνδυασμό των δεδομένων από πολλαπλούς αισθητήρες (sensor fusion) για την υποστήριξη συστημάτων υποστήριξης λήψης αποφάσεων. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την αυτόνομη πλοήγηση επίγειων οχημάτων και μη επανδρωμένων αεροσκαφών, επιτρέποντάς τους να πλοηγούνται σε πολύπλοκα εδάφη και να προσαρμόζονται σε δυναμικά περιβάλλοντα. Η τεχνητή νοημοσύνη μπορεί να αναλύει εικόνες και δεδομένα αισθητήρων για τον εντοπισμό πιθανών στόχων, βελτιώνοντας την ακρίβεια των συστημάτων στοχοποίησης. Η κυβερνοασφάλεια του IoBT είναι επίσης μια πρόκληση καθώς ο αντίπαλος θα επιχειρήσει να εκμεταλλευθεί αδυναμίες των συστημάτων για να τα εξουδετερώσει ή και να τα χρησιμοποιήσει προς όφελός του [73]. Η τεχνητή νοημοσύνη μπορεί να εντοπίζει και να ανταποκρίνεται σε κυβερνοαπειλές σε πραγματικό χρόνο, προστατεύοντας κρίσιμα στρατιωτικά δίκτυα και συστήματα από κυβερνοεπιθέσεις. Η τεχνητή νοημοσύνη μπορεί να αναλύσει τα τρωτά σημεία του δικτύου και να προτείνει προληπτικά μέτρα για την ενίσχυση της ασφάλειας στον κυβερνοχώρο.

Η τεχνητή νοημοσύνη εκτός από τις εφαρμογές στο πεδίο μάχης έχει θέση και στην αναβάθμιση της στρατιωτικής εκπαίδευσης αλλά και της διαδικασίας λήψης αποφάσεων. Τα πολεμικά παίγνια (wargaming) αφορούν την προσομοίωση ή εξομοίωση στρατιωτικών επιχειρήσεων, στρατηγικών και τακτικών σε ένα ελεγχόμενο και συχνά εικονικό περιβάλλον. Αλγόριθμοι ΤΝ μπορούν να συνεισφέρουν σε διάφορα σημεία του πολεμικού παιγνίου από την ανάπτυξη και ανάλυση σεναρίων, μέχρι την εξαγωγή τρόπων ενεργείας και τον έλεγχο αυτόνομων οχημάτων που απεικονίζουν εχθρικές δυνάμεις στο πεδίο [74], [75]. Βέβαια και σε αυτή τη χρήση υφίστανται οι περιορισμοί της μηχανικής μάθησης ως προς την ποιότητα και πληρότητα των δεδομένων εκπαίδευσης αλλά και προβληματισμοί ως προς την υπερβολική εμπιστοσύνη στις προτάσεις των αλγορίθμων [76].

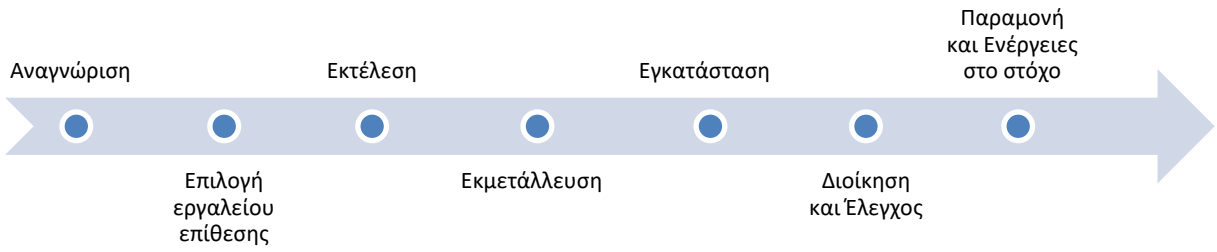
4.3. Ανάλυση Τοπίου Απειλών από τη Χρήση ΤΝ σε Άμυνα και Ασφάλεια

Το τοπίο των απειλών στον κυβερνοχώρο είναι διαρκώς εξελισσόμενο και συχνά επηρεάζεται από τις αλλαγές στην τεχνολογία, τις εξελίξεις στις γεωπολιτικές εντάσεις και τα μεγάλα παγκόσμια γεγονότα. Διάφοροι παράγοντες απειλών, δραστηριοποιούνται στον κυβερνοχώρο και επικαλούνται αλλαγές ή εξελίξεις στον πραγματικό κόσμο για να δικαιολογήσουν τις ενέργειές τους.

Καθώς τα συστήματα επιβολής του νόμου που βασίζονται στην Τεχνητή Νοημοσύνη γίνονται όλο και πιο διαδεδομένα, είναι φυσικό να γίνονται στόχοι επιθέσεων από εγκληματίες. Η κοινότητα επιβολής του νόμου αντιμετωπίζει ξεχωριστές προκλήσεις όσον αφορά στην προστασία από επιθέσεις ΤΝ. Σε αντίθεση με τον στρατιωτικό τομέα που διαθέτει εξειδικευμένο προσωπικό και δυνατότητες ανάπτυξης συστημάτων με δικά του μέσα ή εξουσιοδοτημένες εταιρείες, τα συστήματα τεχνητής νοημοσύνης των διωκτικών αρχών αποτελούν συνήθως προϊόντα που αναπτύσσονται από ιδιωτικές εταιρείες. Το γεγονός αυτό δημιουργεί προκλήσεις στη συντήρηση και την αναβάθμισή τους. Εκτός από τη χρησιμότητα στις δυνάμεις ασφαλείας, η χρήση της ΤΝ μπορεί να δώσει νέες δυνατότητες σε τρομοκρατικές ομάδες. Η χρησιμοποίηση αυτόνομων οχημάτων για την πρόκληση συγκρούσεων ή την μεταφορά εκρηκτικών υλών, οι μειωμένες απαιτήσεις ικανοτήτων ενός ελεύθερου σκοπευτή μέσω της υποβοήθησης της σκόπευσης, η δυνατότητα επιθέσεων από απόσταση με σμήνη αυτόνομων αεροχημάτων είναι μερικά μόνο παραδείγματα με τα οποία τρομοκρατικές ομάδες μπορούν να εκμεταλλευθούν την ΤΝ για να αυξήσουν τον αντίκτυπο μιας επίθεσης και να μειώσουν τον κίνδυνο αποκάλυψης τους [34].

Τα διαφιλονικούμενα περιβάλλοντα στα οποία λειτουργούν οι ένοπλες δυνάμεις δημιουργούν μοναδικούς τρόπους με τους οποίους οι αντίπαλοι μπορούν να οργανώσουν επιθέσεις εναντίον στρατιωτικών συστημάτων ΤΝ και αντίστοιχα μοναδικές προκλήσεις για την άμυνα εναντίον τους. Κακόβουλοι χρήστες μπορούν με επιθέσεις στον κώδικα των εφαρμογών που είναι ενσωματωμένες σε οπλικά συστήματα να προκαλέσουν από δυσλειτουργίες μέχρι απώλειες προσωπικού [77]. Για παράδειγμα παρεμβάλλοντας τα σήματα του δέκτη GPS ενός drone μπορεί να προκληθεί η πτώση του σε φίλια τμήματα.

Για την ανάλυση του τοπίου απειλών από τη χρήση της ΤΝ σε άμυνα και ασφάλεια, χρήσιμο εργαλείο αποτελεί η φονική αλυσίδα στον κυβερνοχώρο (cyber kill chain). Πρόκειται για την μεταφορά στον κυβερνοχώρο της διαδικασίας στοχοποίησης στον στρατιωτικό τομέα. Το συγκεκριμένο πλαίσιο αναφοράς αναπτύχθηκε από την εταιρεία Lockheed Martin και περιγράφει τις ενέργειες που πρέπει να ολοκληρώσουν οι επιτιθέμενοι σε μια κυβερνοεπίθεση για να επιτύχουν τον στόχο τους [78]. Το πλαίσιο περιλαμβάνει 7 στάδια όπως φαίνονται στο παρακάτω διάγραμμα:



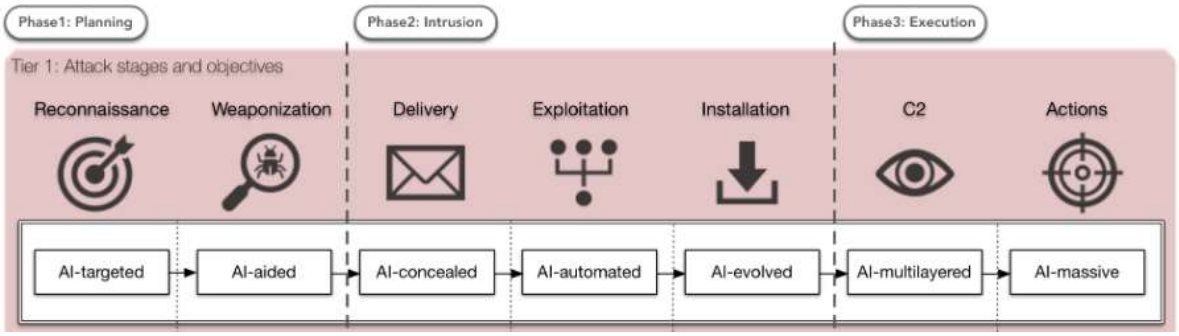
Σχήμα 4.1: Cyber Kill Chain

Στο 1ο στάδιο της αναγνώρισης (reconnaissance), ο επιτιθέμενος επιλέγει το στόχο και προσπαθεί να εντοπίσει τρωτά σημεία. Στο 2ο στάδιο (weaponization) επιλέγει ή και δημιουργεί το μέσο επίθεσης (π.χ. ιό ή worm) προσαρμοσμένο στις τρωτότητες που έχει εντοπίσει. Στη φάση της εκτέλεσης (delivery), ο επιτιθέμενος εισάγει το εργαλείο επίθεσης στο στόχο (π.χ. μέσω ενός συνημμένου σε email ή μέσω μονάδας usb). Στη φάση της εκμετάλλευσης (exploitation), ενεργοποιείται ο κακόβουλος κώδικας για εκμετάλλευση της ευπάθειας. Στη φάση της εγκατάστασης (installation), το κακόβουλο λογισμικό εγκαθιστά ένα σημείο εισόδου (π.χ. κερκόπορτα) για διατήρηση της πρόσβασης. Στη φάση της διοίκησης και ελέγχου (command & control), ο επιτιθέμενος είναι σε θέση να δίνει εντολές από απόσταση στο προσβεβλημένο σύστημα. Στην τελική φάση (persistence & actions on objective) ο επιτιθέμενος αναλαμβάνει δράση για την επίτευξη των στόχων του (π.χ. διαρροή δεδομένων, καταστροφή δεδομένων, κρυπτογράφηση δεδομένων για λύτρα). Βασικό στοιχείο του μοντέλου αυτού είναι οι ευπάθειες του συστήματος, επάνω στις οποίες βασίζονται όλοι οι κρίκοι της αλυσίδας [79]. Λαμβάνοντας υπόψη τις εγγενείς αδυναμίες των φάσεων του κύκλου ζωής ενός συστήματος ΤΝ, γίνεται αντιληπτό ότι αυξάνονται οι ευπάθειες συστημάτων που βασίζονται στην ΤΝ.

Εξετάζοντας τη φονική αλυσίδα στην περίπτωση ενός συστήματος ΤΝ, στη φάση της αναγνώρισης, οι επιτιθέμενοι μπορούν να προβούν σε ανάλυση αποθετηρίων κώδικα ή ερευνητικών εργασιών στον τομέα της ΤΝ για τον εντοπισμό πιθανών αδυναμιών ή φορέων επίθεσης. Στη φάση της επιλογής εργαλείων επίθεσης, οι επιτιθέμενοι μπορούν μέσω ανταγωνιστικών παραδειγμάτων να επιχειρήσουν την παράκαμψη συστημάτων ανίχνευσης ή ταξινόμησης βασισμένων στην τεχνητή νοημοσύνη. Στη φάση της εκτέλεσης μπορούν να χρησιμοποιηθούν payload που έχουν δημιουργηθεί με ΤΝ μέσω διαφόρων καναλιών, όπως κακόβουλα μηνύματα ηλεκτρονικού ταχυδρομείου, κακόβουλοι ιστότοποι ή επιθέσεις στην αλυσίδα εφοδιασμού που στοχεύουν σε πλαίσια ή βιβλιοθήκες ανάπτυξης συστημάτων ΤΝ. Στη φάση της εκμετάλλευσης, οι επιτιθέμενοι εκμεταλλεύονται ευπάθειες σε συστήματα ΤΝ για την απόκτηση μη εξουσιοδοτημένης πρόσβασης, χειραγώγηση των αποτελεσμάτων, ή απόσπαση ευαίσθητων πληροφοριών. Στη φάση της εγκατάστασης επιδιώκεται εγκατάσταση κερκόπορτας ή κακόβουλου λογισμικού σε παραβιασμένα συστήματα ΤΝ για τη διατήρηση της μη εξουσιοδοτημένης πρόσβασης ή τη διεξαγωγή περαιτέρω κακόβουλων δραστηριοτήτων. Στη φάση της διοίκησης και ελέγχου επιδιώκεται ο έλεγχος από απόσταση των συστημάτων ΤΝ μέσω κρυφών ή κρυπτογραφημένων καναλιών επικοινωνίας. Στην τελευταία φάση, επιτυγχάνονται κακόβουλοι στόχοι, όπως η κλοπή ευαίσθητων δεδομένων, η διατάραξη λειτουργιών ή η υπονόμευση της εμπιστοσύνης του συστήματος ΤΝ.

Όσον αφορά στη χρήση της ΤΝ για την υποβοήθηση κυβερνοεπιθέσεων, οι αλγόριθμοι μηχανικής μάθησης μπορούν να συνεισφέρουν σε όλα τα στάδια της φονικής αλυσίδας στον κυβερνοχώρο [80]. Στη φάση της αναγνώρισης, η ΤΝ μπορεί να εντοπίσει τόσο τους καταλληλότερους στόχους όσο και τις ευπάθειες που μπορεί να εκμεταλλευθεί ο επιτιθέμενος. Στη φάση της επιλογής του εργαλείου επίθεσης, η ΤΝ μπορεί να δημιουργήσει βέλτιστα ωφέλιμα φορτία (payloads) για την εκμετάλλευση των αδυναμιών. Στη φάση της εκτέλεσης, η μετάδοση του κακόβουλου ωφέλιμου φορτίου μπορεί μέσω της ΤΝ να παραμείνει μη ανιχνεύσιμη και κρυμμένη για μεγάλα χρονικά διαστήματα. Αφού αποκτηθεί πρόσβαση στο στόχο, η χρήση των δυνατοτήτων τεχνητής νοημοσύνης μπορεί να αυξήσει τον αριθμό των επιτιθέμενων που είναι σε θέση να πραγματοποιήσουν την επίθεση, μειώνοντας την ανάγκη σε ανθρώπινους πόρους. Στη φάση της εγκατάστασης, η ΤΝ μπορεί να προσφέρει

τεχνικές μετάδοσης της μόλυνσης που δεν ανιχνεύονται από συστήματα προστασίας. Η ανάγκη διαρκούς πρόσβασης μπορεί να επιτευχθεί μέσω της ικανότητας της τεχνητής νοημοσύνης να μαθαίνει με βάση τις εισροές από το περιβάλλον και να προσαρμόζει τη συμπεριφορά του. Στο τελευταίο στάδιο η ΤΝ μπορεί να αυξήσει το μέγεθος των επιθυμητών ενεργειών μέσω του συντονισμού των ενεργειών στο στόχο. Τα παραπάνω απεικονίζονται σχηματικά στο ακόλουθο διάγραμμα:



Σχήμα 4.2: Πλαίσιο κυβερνοαπειλών με βάση την ΤΝ[80]

Με βάση την ανάλυση που προηγήθηκε, μπορούμε εξετάσουμε τις ενέργειες ενός επιτιθέμενου σε ένα δυνητικό σύστημα ΤΝ στον τομέα της άμυνας. Έστω ένα σύστημα αυτόνομου αεροχήματος το οποίο βασίζεται σε ΤΝ και χρησιμοποιείται για τον εντοπισμό εχθρικών στόχων στο πεδίο της μάχης. Οι ενέργειες του επιτιθέμενου προκειμένου να καταφέρει την εξουδετέρωσή του στο πεδίο της μάχης εκμεταλλευόμενο τις αδυναμίες της ΤΝ, φαίνεται στον παρακάτω πίνακα:

Φάση	Ενέργειες Επιτιθέμενου	Συνέπειες
Αναγνώριση	• Αναγνώριση χρησιμοποιούμενης πλατφόρμας (υλικού) και γνωστών αδυναμιών (υλικού - λογισμικού) • Απόκτηση σχετικής πλατφόρμας μέσω προμηθευτή ή ανάκτηση από το πεδίο της μάχης • Εντοπισμός χρησιμοποιούμενου LLM (εμπορικού ή αποκλειστικής χρήσης) • Εντοπισμός εργαζομένων στο πρόγραμμα ανάπτυξης και εκπαιδευόμενων στο χειρισμό	• Εντοπισμός γνωστών αδυναμιών • Αντίστροφη μηχανική • Εργαστηριακές δοκιμές (έλεγχος διείσδυσης – ανάλυση ψηφιακών πειστηρίων) • Κοινωνική μηχανική για

Φάση	Ενέργειες Επιτιθέμενου	Συνέπειες
		απόκτηση πρόσβασης
Επιλογή εργαλείου επίθεσης	• Επιλογή εργαλείων (υλικό – λογισμικό) για εκμετάλλευση γνωστών ευπαθειών • Ανάπτυξη νέων εργαλείων εξειδικευμένων στην πλατφόρμα με χρήση ΤΝ • Επιλογή εισόδων στο LLM που προκαλούν δυσλειτουργία • Δημιουργία αντίπαλων παραδειγμάτων για την παράκαμψη συστημάτων ανίχνευσης ή ταξινόμησης • Ανάπτυξη κακόβουλου υλικολογισμικού (firmware) της πλατφόρμας • Επιλογή εργαζομένων – χειριστών για επίθεση	• Δυνατότητα απομακρυσμένης πρόσβασης • Δυνατότητα δυσλειτουργιών • Δυνατότητα δημιουργίας κερκόπορτας
Εκτέλεση	• Εισαγωγή κακόβουλο κώδικα στην πλατφόρμα • Κυβερνοεπίθεση στην εταιρεία-οργανισμό ανάπτυξης • Απόκτηση πρόσβασης μέσω κοινωνικής μηχανικής • Επιθέσεις σε χρησιμοποιούμενο γλωσσικό μοντέλο • Δημοσίευση (ανοικτά αποθετήρια κώδικα) – εισαγωγή κακόβουλου υλικολογισμικού (εταιρεία/οργανισμός ανάπτυξης)	• Απόκτηση πρόσβασης • Ανάκτηση γλωσσικού μοντέλου
Εκμετάλλευση	• Ανάλυση χρησιμοποιούμενου γλωσσικού μοντέλου	• Μεταφορά μάθησης σε άλλο γλωσσικό μοντέλο • Ανάκτηση διαβαθμισμένων πληροφοριών
Εγκατάσταση	• Εγκατάσταση κερκόπορτας ή κακόβουλου λογισμικού • Εγκατάσταση κακόβουλου υλικολογισμικού εν αγνοία του χρήστη	• Μη εξουσιοδοτημένη πρόσβαση • Δυνατότητα εξουδετέρωσης μέσω παρεμβολών

Φάση	Ενέργειες Επιτιθέμενου	Συνέπειες
	• Εκμετάλλευση λανθασμένων ρυθμίσεων ή αδύναμων μηχανισμών ελέγχου ταυτότητας • Εγκατάσταση εισόδων που προκαλούν δυσλειτουργία του γλωσσικού μοντέλου σε παρεμβολέα	
Διοίκηση και έλεγχος	• Απομακρυσμένη διαχείριση παραβιασμένου συστήματος • Αποκάλυψη θέσης σταθμού ελέγχου	• Μη αναμενόμενη λειτουργία
Παραμονή και ενέργειες στο στόχο	• Συλλογή πληροφοριών στόχου • Διακοπή λειτουργίας • Υπονόμευση εμπιστοσύνης	• Εξουδετέρωση συστήματος • Δημιουργία σύγχυσης

Πίνακας 4.1: Cyber Kill Chain Εναντίον Αυτόνομου Αεροχήματος

4.4. Παραδείγματα Χρήσης και Προβληματισμοί

Πλήθος εφαρμογών ΤΝ είναι σε ανάπτυξη και αναμένεται να ενταχθούν ή έχουν δοκιμασθεί σε υπηρεσίες ασφαλείας και ένοπλες δυνάμεις. Στον τομέα της ασφάλειας συνόρων, στην ΕΕ δεν υπάρχει κάποιο σύστημα ΤΝ σε χρήση, ωστόσο έχουν γίνει δοκιμές συστημάτων ανίχνευσης συναισθημάτων με σκοπό των εντοπισμό περιπτώσεων εξαπάτησης στο πλαίσιο συνοριακών ελέγχων. Το φορητό σύστημα iBorderCtrl αναπτύχθηκε από το 2013 έως το 2019 και δοκιμάστηκε πιλοτικά σε συνοριακούς ελέγχους σε Ελλάδα, Ουγγαρία και Λετονία. Η ανίχνευση βασίζεται σε συνέντευξη του ελεγχόμενου προσώπου από ένα άβαταρ και συλλογή στοιχείων από χαρακτηριστικές κινήσεις του προσώπου [57]. Η αναφερόμενη αποτελεσματικότητα ανίχνευσης ψεύδους έφτανε στο 75% ωστόσο η εγκυρότητα των αποτελεσμάτων αμφισβητείται με κύρια αδυναμία την ανεπάρκεια των δεδομένων εκπαίδευσης του συστήματος [81].

Η Ευρωπαϊκή Υπηρεσία Ασύλου EUAA (πρώην EASO), παρήγαγε εκθέσεις παρακολούθησης αναρτήσεων στα μέσα κοινωνικής δικτύωσης σε διαφορετικές γλώσσες που σχετίζονταν με θέματα ασύλου και μετανάστευσης της ΕΕ. Η χρήση του συστήματος διακόπηκε λόγω νομικών θεμάτων και ανησυχιών για την προστασία δεδομένων [57].

Το πρόγραμμα ROBORDER χρηματοδοτείται από την ΕΕ και αφορά στην ανάπτυξη ενός συστήματος που εντοπίζει και αναγνωρίζει παράνομες συνοριακές δραστηριότητες, αξιολογεί τις συνθήκες και ενημερώσει τις αρχές σχετικά με την κατάσταση της συνοριακής περιοχής. Το σύστημα βασίζεται στη συλλογή ετερογενών δεδομένων από πολλούς διαφορετικούς πόρους, όπως θερμικές και οπτικές κάμερες, παθητικά ραντάρ και αισθητήρες RF σε αυτόνομα οχήματα και σταθερές εγκαταστάσεις. Το σύστημα έχει δοκιμασθεί σε θαλάσσια σύνορα στην Ελλάδα και σε χερσαία σύνορα στη Βουλγαρία [82]. Η διάθεση της συγκεκριμένης τεχνολογίας σε τρίτους ή και η επέκταση του συστήματος σε στρατιωτικούς σκοπούς είναι ορισμένοι προβληματισμοί που έχουν διατυπωθεί για το εν λόγω σύστημα [83].

Η ΕΕ χρηματοδοτεί την έρευνα για αυτόνομα συστήματα για αμυντικούς σκοπούς [55]. Για παράδειγμα το πρόγραμμα OCEAN2020 υποστηρίζει αποστολές θαλάσσιας επιτήρησης και απαγόρευσης μέσω εναέριων μέσων, μέσων επιφανείας και υποβρύχιων μη επανδρωμένων συστημάτων. Το πρόγραμμα AIDED αφορά τη χρήση ΤΝ για τον εντοπισμό και αναγνώριση αυτοσχέδιων εκρηκτικών μηχανισμών (IED) και συμβατικών εκρηκτικών μηχανισμών (νάρκες). Το πρόγραμμα ARTUS αφορά την ανάπτυξη σμήνους αυτόνομων οχημάτων που θα μπορούν να μεταφέρουν εφόδια στο πεδίο της μάχης και να εκτελούν διακομιδές τραυματιών. Το πρόγραμμα AI4DEF εξετάζει τη χρήση ΤΝ για την επίγνωση κατάστασης, υποστήριξη λήψης αποφάσεων και σχεδίαση. Το πρόγραμμα INTEGRAL εξετάζει τη χρήση ΤΝ στην επίγνωση κατάστασης διαστήματος για τον εντοπισμό απειλών και τη στοχοποίηση με τη χρήση δορυφόρων. Το πρόγραμμα HERMES αναπτύσσεται με σκοπό την ανταλλαγή πληροφοριών επίγνωσης κατάστασης κυβερνοασφάλειας και τον έλεγχο αυτόνομων συστημάτων απόκρισης σε κυβερνοεπιθέσεων. Το πρόγραμμα SWADAR εξετάζει τον εντοπισμό και την παρακολούθηση σμήνους drone για την αποτροπή επιθέσεων.

Εκτός όμως από ερευνητικά προγράμματα υπάρχουν ήδη αναφερόμενες χρήσεις ΤΝ σε πεδία ένοπλων συρράξεων. Αναφορά του ΟΗΕ τον Μάρτιο του 2021 αναφέρει πιθανή χρήση τουρκικού drone Kargu-2 στη Λιβύη για την αυτόνομη προσβολή στόχων προσωπικού. Η κατασκευάστρια εταιρεία STM αναφέρει ότι το drone είναι εφοδιασμένο με αλγόριθμο μηχανικής μάθησης για την αναγνώριση και προσβολή στόχων προσωπικού [84]. Σύμφωνα με δηλώσεις Ρώσων αξιωματούχων, το

αντιαεροπορικό σύστημα S-350 Vityaz κατέρριψε ουκρανικό αεροσκάφος ενεργώντας αυτόνομα για τον εντοπισμό, παρακολούθηση και προσβολή του [85]. Οι ουκρανικές δυνάμεις χρησιμοποιούν το drone Saker Scout, το οποίο κατά δήλωσή τους εντοπίζει και προσβάλει στόχους αυτόνομα. Σύμφωνα με τον κατασκευαστή, το λογισμικό είναι σε θέση να αναγνωρίζει 64 διαφορετικά στρατιωτικά αντικείμενα τα οποία μπορεί να προσβάλει με βομβίδια που μεταφέρει με βάρος έως 3 κιλών [86]. Το Ισραήλ ανακοίνωσε τον Ιούνιο του 2023 ότι οι μυστικές του υπηρεσίες χρησιμοποιούν ΤΝ για την αποτροπή τρομοκρατικών επιθέσεων [87]. Μετά την τρομοκρατική επιχείρηση της Χαμάς στο Ισραήλ στις 7 Οκτωβρίου 2023, εμφανίστηκε πλήθος περιπτώσεων παραπληροφόρησης με τη χρήση του Telegram και του Twitter (X). Βίντεο και δορυφορικές φωτογραφίες από παλαιότερες συγκρούσεις, λογαριασμοί που ισχυρίζονταν ψευδώς ότι σχετίζονται με διεθνή πρακτορεία ειδήσεων, φωτογραφίες που δημιουργήθηκαν με ΤΝ είναι μερικές από τις τεχνικές που χρησιμοποιήθηκαν [88]. Κατά τη διάρκεια των πολεμικών επιχειρήσεων του ισραηλινού στρατού στη λωρίδα της Γάζας ανακοινώθηκε ότι οι IDF χρησιμοποιούν το σύστημα ΤΝ Gospel για την επιλογή στόχων για αεροπορική προσβολή και το σύστημα Fire Factory για την οργάνωση και παρακολούθηση των αεροπορικών επιδρομών [89]. Η πρώτη επιχειρησιακή χρήση του εν λόγω συστήματος ήταν κατά τη διάρκεια της επιχείρησης Guardian of the Walls το 2011, την οποία οι IDF χαρακτήρισαν τον «πρώτο πόλεμο τεχνητής νοημοσύνης» [90]. Οι ένοπλες δυνάμεις της Ινδίας χρησιμοποιούν 140 συστήματα επιτήρησης που βασίζονται στην ΤΝ για την επιτήρηση των συνόρων με την Κίνα και το Πακιστάν. Το σύστημα περιλαμβάνει κάμερες υψηλής ανάλυσης, αισθητήρες, μη επανδρωμένα αεροχήματα και ραντάρ για την τροφοδότηση μοντέλου ΤΝ για τον εντοπισμό και την κατηγοριοποίηση εισβολέων και την αναγνώριση στόχων [91].

Παρά τις πολλά υποσχόμενες δυνατότητες, η ενσωμάτωση τις ΤΝ σε συστήματα άμυνας και ασφάλειας έχει και μια σειρά από προκλήσεις οι οποίες μπορούν να καθυστερήσουν ή και να δράσουν περιοριστικά στην παροχή νέων δυνατοτήτων [92]. Τα μοντέλα μηχανικής μάθησης χαρακτηρίζονται από μειωμένη διαφάνεια και επεξηγησιμότητα. Τα συγκεκριμένα χαρακτηριστικά που αναλύονται στο επόμενο κεφάλαιο, καθιστούν δυσχερή την αιτιολόγηση των αποφάσεων ενός μοντέλου σε ένα ανθρώπινο χειριστή ο οποίος καλείται με βάση την εκτίμηση ενός μοντέλου να

εκτελέσει για παράδειγμα την προσβολή ενός στόχου. Επιπλέον το πλήθος αδυναμιών που εντοπίζονται σε όλες τις φάσεις του κύκλου ζωής ενός συστήματος ΤΝ, δίνει δυνατότητες παραπλάνησής τους ή εξουδετέρωσής τους από τον αντίπαλο. Η εκπαίδευση ενός μοντέλου είναι επίσης μια πρόκληση καθώς είναι χρονοβόρα, απαιτητική σε πόρους και απαιτεί πλήθος δεδομένων. Στο στρατιωτικό περιβάλλον μπορεί να υπάρχουν περιορισμοί στη διαθεσιμότητα των δεδομένων λόγω διαβάθμισης ασφαλείας, υφίστανται αυξημένες απαιτήσεις εκπαίδευσης των μοντέλων με νέα δεδομένα ή τα διαθέσιμα δεδομένα δεν είναι επαρκώς επισημασμένα.

Στον τομέα της διεθνούς ασφάλειας, υφίσταται η δυνατότητα σε δρώντες με χαμηλές συγκριτικά δυνατότητες να επωφεληθούν από εμπορικά ή ακαδημαϊκά μοντέλα ΤΝ σε μια προσπάθεια να αποκτήσουν υψηλό στρατιωτικό αντίκτυπο με χαμηλό κόστος και προσπάθεια [93]. Η ενσωμάτωση ηθικών και νομικών περιορισμών σε συστήματα ΤΝ στοχοποίησης ή παρακολούθησης αποτελεί μια περίπλοκη και δύσκολα εφαρμόσιμη διαδικασία και σε συνδυασμό με την έλλειψη διαφάνειας στη λειτουργία των αλγορίθμων αυξάνει την καχυποψία για την αποτελεσματικότητά τους.

Σε συστήματα άμυνας και ασφάλειας υπάρχουν αυξημένες απαιτήσεις ανθρώπινης παρέμβασης σε όλες τις φάσεις του κύκλου ζωής ώστε να διασφαλίζεται ότι τηρούνται οι ηθικοί και νομικοί περιορισμοί ή ότι τυχόν σφάλματα των μοντέλων δεν θα έχουν επιπτώσεις σε ανθρώπινες ζωές. Ακόμα και αν έχει προβλεφθεί ανθρώπινη παρέμβαση, σε περιπτώσεις που ο χρόνος λήψης απόφασης είναι πιεστικός υπάρχει ο κίνδυνος η υπερβολική εμπιστοσύνη στο σύστημα να έχει αρνητικές επιπτώσεις.

ΚΕΦΑΛΑΙΟ 5

ΗΘΙΚΑ ΖΗΤΗΜΑΤΑ ΚΑΙ ΡΥΘΜΙΣΤΙΚΟ ΠΛΑΙΣΙΟ

Η χρήση της τεχνητής νοημοσύνης στην άμυνα και την ασφάλεια εγείρει ηθικά ζητήματα και προκλήσεις. Το ενδεχόμενο μεροληψίας των αλγορίθμων μηχανικής μάθησης, η παραβίαση της ιδιωτικότητας και η ανάπτυξη αυτόνομων όπλων θέτουν ηθικούς προβληματισμούς που απαιτούν προσεκτική εξέταση. Η εξεύρεση ισορροπίας μεταξύ της αξιοποίησης της τεχνητής νοημοσύνης για την ενίσχυση της ασφάλειας και της αντιμετώπισης των ηθικών επιπτώσεων παραμένει μια κρίσιμη πτυχή της υιοθέτησης των τεχνολογιών ΤΝ στην άμυνα και την επιβολή του νόμου.

5.1. Ηθικά Ζητήματα και Ζητήματα Ιδιωτικότητας

Ένα από τα ηθικά ζητήματα που ανακύπτουν από τη χρήση της ΤΝ, έχει να κάνει με την εμπιστοσύνη στην αποτελεσματικότητα των συστημάτων αυτών. Οι ηθικές προκλήσεις που εγείρουν θέματα εμπιστοσύνης μπορούν να κατηγοριοποιηθούν ως εξής [7]:

- Διαφάνεια και επεξηγησιμότητα: η λογική του «μαύρου κουτιού» είναι ένα εγγενές χαρακτηριστικό των αλγορίθμων βαθιάς μάθησης, το οποίο δυσχεραίνει την αποκατάσταση εμπιστοσύνης. Ταυτόχρονα στην περίπτωση που παρέχονται επεξηγήσεις για τις εξόδους ενός συστήματος, μπορεί να δημιουργηθεί υπερβολική εμπιστοσύνη. Για το λόγο αυτό θα πρέπει να επιδιώκεται η διαφάνεια σε όλη τη διαδικασία εξαγωγής αποτελεσμάτων και οι άνθρωποι που είναι επιφορτισμένοι με τον έλεγχο να είναι κατάλληλα εκπαιδευμένοι.
- Ακρίβεια και αξιοπιστία: η μειωμένη ακρίβεια των εξόδων ενός μοντέλου προφανώς μειώνει την εμπιστοσύνη ιδιαίτερα όταν παρουσιάζεται στα πρώιμα στάδια της χρήσης ενός συστήματος. Οι έλεγχοι των εξόδων στα ενδιάμεσα στάδια του κύκλου ζωής και η επανεκπαίδευση μπορούν να βελτιώσουν την αξιοπιστία.
- Αυτοματοποίηση έναντι επαύξησης δυνατοτήτων: η «τυφλή» εμπιστοσύνη σε ένα σύστημα ΤΝ μπορεί να οδηγήσει στην απώλεια δεξιοτήτων από τους

αρμόδιους χειριστές. Αντίθετα ένα σύστημα ΤΝ που επαυξάνει τις δυνατότητες των ανθρώπων είναι θεμιτό. Ειδικά σε περιβάλλοντα ασφαλείας και άμυνας η ανάθεση εργασιών εξολοκλήρου σε συστήματα ΤΝ δεν ενδείκνυται.

- Ανθρωπομορφισμός: αφορά την ενσωμάτωση ανθρώπινων χαρακτηριστικών σε συστήματα ΤΝ. Σε αυτή την περίπτωση υφίστανται ανησυχίες για τη δημιουργία υπερβολικής εμπιστοσύνης αλλά και χειραγώγησης.
- Μαζική εξαγωγή δεδομένων: η προέλευση και η μορφή των δεδομένων εκπαίδευσης και επανεκπαίδευσης των μοντέλων ΤΝ μπορεί να δημιουργήσει θέματα προστασίας ιδιωτικότητας. Ειδικά σε συστήματα ασφαλείας τα οποία μπορεί να τροφοδοτούνται σε πραγματικό χρόνο με βιομετρικά στοιχεία και στοιχεία συμπεριφοράς, εγείρονται ζητήματα ορθής διακυβέρνησης των συστημάτων για την ασφάλεια και προστασία των δεδομένων αυτών, την ανωνυμοποίηση αλλά και την καταστροφή των δεδομένων όταν πλέον δεν απαιτούνται.

Επεκτείνοντας το ζήτημα της χειραγώγησης αποτελεσμάτων, ένα από τα ηθικά ζητήματα που ανέκυψαν ιδιαίτερα με την ανάπτυξη της παραγωγικής ΤΝ είναι αυτό των αλγοριθμικών προκαταλήψεων [94]. Τα δεδομένα εκπαίδευσης ενός μεγάλου γλωσσικού μοντέλου παίζουν σημαντικό ρόλο στη συμπεριφορά του. Έχουν ήδη καταγραφεί περιπτώσεις αρνητικής μεροληψίας, δηλαδή διακρίσεων σε βάρος προσώπων με βάση ορισμένα χαρακτηριστικά τους όπως το φύλο, η εθνική, εθνοτική ή φυλετική καταγωγή και το θρήσκευμα. Στην περίπτωση συστημάτων που χρησιμοποιούνται σε περιβάλλον αρχών ασφαλείας, τέτοιου είδους προκαταλήψεις μπορεί να έχουν ως συνέπεια μεροληπτικά αποτελέσματα ή και παράλειψη αποτελεσμάτων λόγω των προκαταλήψεων. Για παράδειγμα ένα μεροληπτικό σύστημα που εξετάζει τις εισόδους προσώπων στην χώρα, μπορεί να μεροληπτεί κατά κάποιων ομάδων ανθρώπων, ενώ κάποιες άλλες ομάδες να περνούν χωρίς την εμφάνιση προειδοποιήσεων, παρά το γεγονός ότι πληρούν άλλες προϋποθέσεις που έχουν τεθεί. Αντίστοιχα μια εφαρμογή αναγνώρισης εχθρικών οχημάτων που έχει για παράδειγμα εκπαιδευθεί μεροληπτικά με εικόνες οχημάτων συγκεκριμένης χώρας προέλευσης, πιθανότατα δεν θα αναγνωρίσει ως εχθρικά οχήματα που έχουν κοινό κατασκευαστή με τη χώρα προέλευσης του συστήματος.

Η χρήση της ΤΝ παρέχει πολλαπλές ευκαιρίες οι οποίες όμως δημιουργούν και αντίστοιχες ηθικές προκλήσεις οι οποίες προκαλούνται από την υπερβολική ή την κακόβουλη χρήση της ΤΝ [95]. Από την μια πλευρά η ΤΝ μειώνει το φόρτο από επίπονες γνωστικές διαδικασίες από τους ανθρώπους δίνοντας τη δυνατότητα να αναπτυχθούν σε άλλους τομείς, από την άλλη πλευρά όμως δημιουργείται ο κίνδυνος απώλειας δεξιοτήτων και αντίληψης διαδικασιών για τις οποίες απασχολούνται συστήματα ΤΝ. Για παράδειγμα στον στρατιωτικό τομέα ένας χειριστής οπλικού συστήματος επαναπαυόμενος στις προτάσεις ενός συστήματος μπορεί να απωλέσει την κριτική του ικανότητα περί της νομιμότητας ή της αναγκαιότητας προσβολής ενός στόχου. Η ΤΝ βελτιώνει τις ανθρώπινες δυνατότητες επεξεργασίας και ανάλυσης δεδομένων ωστόσο δεν θα πρέπει να αποτελεί πρόφαση για την αποποίηση ευθύνης. Για παράδειγμα η απόφαση για την προσβολή ενός στόχου που στηρίζεται σε ΤΝ δεν απαλλάσσει από ευθύνες τους χειριστές ή και τους προγραμματιστές ενός οπλικού συστήματος.

Οι αυξανόμενες δυνατότητες που παρέχει η ΤΝ δίνουν τη δυνατότητα νέων λύσεων σε παλαιά προβλήματα. Ωστόσο το γεγονός ότι πολλές φορές οι αλγόριθμοι λειτουργούν με τρόπο που δεν είναι αντιληπτός ή κατανοητός στο μέσο χρήστη δεν θα πρέπει να απομακρύνει τον έλεγχο από τους ανθρώπους. Για παράδειγμα συστήματα επιτήρησης δεν θα πρέπει να είναι σε θέση να συλλέγουν ανεξέλεγκτα πληροφορίες. Τέλος ενώ η ΤΝ μπορεί να προτείνει βέλτιστες πρακτικές με βάση ιστορικά γεγονότα και να προβλέψει αποτελέσματα για την αντιμετώπιση πλήθους προβλημάτων, δεν θα πρέπει να μειώνει τον ανθρώπινο αυτοέλεγχο. Για παράδειγμα ένα σύστημα υποστήριξης της επιχειρησιακής σχεδίασης, μπορεί να κατευθύνει τους στρατιωτικούς διοικητές στην υιοθέτηση μιας απόφασης στερώντας τους την πρωτοβουλία και την επινοητικότητα.

Όπως προαναφέρθηκε, η ΤΝ στον τομέα της κυβερνοασφάλειας συνεισφέρει στην ευρωστία, ανταπόκριση και ανθεκτικότητα (3R) των πληροφοριακών συστημάτων. Στους ίδιους τομείς όμως δημιουργούνται παρόμοια ηθικά ζητήματα με αυτά που παρουσιάσθηκαν παραπάνω [96]. Για παράδειγμα, η ανάθεση του ελέγχου λογισμικού και συστημάτων στην ΤΝ στο πλαίσιο διασφάλισης της ευρωστίας μπορεί να οδηγήσει στη μείωση δεξιοτήτων των εμπειρογνωμόνων. Η ανταπόκριση σε κυβερνοεπιθέσεις

διευκολύνεται μέσω της ΤΝ, καθώς με βάση τα στοιχεία που συλλέγονται από μια επίθεση είναι εφικτή η ανάπτυξη ανταποδοτικών κυβερνοεπιθέσεων. Το γεγονός αυτό σε επίπεδο κρατών μπορεί να οδηγήσει σε κλιμάκωση της αντιπαράθεσης και πρόκληση συγκρούσεων και σε άλλα πεδία εκτός του κυβερνοχώρου ως απάντηση. Η ανθεκτικότητα των συστημάτων μέσω της ΤΝ απαιτεί τη συνεχή παρακολούθηση των συστημάτων και τη συλλογή πλήθους δεδομένων. Το γεγονός αυτό θέτει σε κίνδυνο την ιδιωτικότητα των χρηστών, δημιουργώντας νέους κινδύνους όταν παραβιάζεται η εμπιστευτικότητα των συλλεγόμενων δεδομένων και με τη διάχυση των συστημάτων ΤΝ σε όλο και περισσότερες δραστηριότητες οδηγεί στο φαινόμενο της μαζικής παρακολούθησης.

Σε θέματα ιδιωτικότητας, τεχνολογίες όπως η δημιουργική ΤΝ δημιουργούν ανησυχίες ως προς τον τρόπο χρήσης των συλλεγόμενων δεδομένων για την ανάπτυξη των μεγάλων γλωσσικών μοντέλων [97]. Η μαζική συλλογή δεδομένων δεν έχει τη συναίνεση των υποκειμένων, ενώ δημιουργούνται και θέματα πνευματικών δικαιωμάτων καθώς χρησιμοποιείται υλικό χωρίς τη γνώση, την άδεια ή την αποζημίωση των δημιουργών. Επιπλέον τα συλλεγόμενα δεδομένα μπορεί να περιλαμβάνουν ευαίσθητα δεδομένα τα οποία αποκαλύπτουν προσωπικές πληροφορίες. Οι προκλήσεις αυτές είναι ακόμα μεγαλύτερες στο πλαίσιο αστυνομικών υπηρεσιών και ενόπλων δυνάμεων όπου τα χρησιμοποιούμενα δεδομένα μπορεί να δημιουργούν θέματα εθνικής ασφάλειας. Ζητήματα δημιουργούνται και από τις προτροπές που δίνονται σε chatbots, οι οποίες μπορεί να περιέχουν προσωπικές ή εμπιστευτικές πληροφορίες που χρησιμοποιούνται για την επανεκπαίδευση και τον έλεγχο της απόδοσης των LLMs.

Ένας διαφορετικός ηθικός προβληματισμός αφορά στην αποκάλυψη των ευπαθειών των συστημάτων ΤΝ. Στον τομέα της κυβερνοασφάλειας έχει επικρατήσει η κοινοποίηση των ευπαθειών που εντοπίζονται σε πληροφοριακά συστήματα ώστε να ενημερώνονται οι κατασκευαστές και να αναλαμβάνουν μέτρα αποκατάστασης. Με τον τρόπο αυτό επιδιώκεται να μειωθεί η δυνατότητα εκμετάλλευσης της αδυναμίας από κακόβουλους δρώντες. Ωστόσο μια ευπάθεια που μπορεί να αποκατασταθεί με μια ενημέρωση του λογισμικού ενός πληροφοριακού συστήματος δεν αφορά πάντα τα συστήματα ΤΝ. Λόγω της φύσης τους, κάποιες ευπάθειες μπορεί να μην είναι άμεσα

επιδιορθώσιμες. Σε αυτή την περίπτωση οι κρατικοί δρώντες έχουν όφελος να μην δημοσιοποιούν ευπάθειες που εντοπίζουν σε συστήματα ΤΝ προκειμένου να είναι σε θέση να τις εκμεταλλεύονται για επιθετικούς σκοπούς [25].

5.2. Ρυθμιστικό Πλαίσιο Χρήσης ΤΝ σε Περιβάλλοντα Άμυνας και Ασφάλειας

Η ανάγκη θέσπισης πλαισίου κανόνων για την ανάπτυξη συστημάτων ΤΝ καταγράφεται εκτενώς στη μελετηθείσα βιβλιογραφία. Το πρόσφατο διάστημα έχουν δημοσιευθεί πλήθος προτύπων και οδηγιών σχετικά με την τεχνητή νοημοσύνη. Ωστόσο δεν ισχύει το ίδιο με το νομοθετικό πλαίσιο. Η αλματώδης πρόοδος του τομέα έχει σαν αποτέλεσμα οι κυβερνήσεις να τρέχουν πίσω από τις εξελίξεις και πολλά θέματα να επαφίονται στην υπευθυνότητα των εταιρειών που αναπτύσσουν τα σχετικά συστήματα.

Σε ακαδημαϊκό επίπεδο, ένα προτεινόμενο ηθικό πλαίσιο για τη χρήση της ΤΝ βασίζεται στις 4 θεμελιώδεις αρχές της βιοηθικής (αυτονομία, μη κακόβουλη συμπεριφορά, ευεργεσία, δικαιοσύνη) και προσθέτει την επεξηγησιμότητα [95]. Η αρχή της αυτονομίας έγκειται στην ισορροπία μεταξύ της εξουσίας λήψης απόφασης που διατηρείται από τον άνθρωπο με αυτή που αποδίδεται σε ένα σύστημα ΤΝ. Η μη κακόβουλη συμπεριφορά έγκειται σε θέματα διαφύλαξης ιδιωτικότητας, ασφάλειας αλλά και ελέγχου των δυνατοτήτων της ΤΝ. Η αρχή της ευεργεσίας έγκειται στο προσδοκώμενο όφελος από τη χρήση ενός συστήματος ΤΝ, η οποία θα πρέπει να εξετάζεται σε συνδυασμό με την προηγούμενη αρχή, ειδικά για συστήματα που έχουν διττή χρήση. Η αρχή της δικαιοσύνης αφορά στην ισότιμη ανάπτυξη των δυνατοτήτων που παρέχει η ΤΝ προς όφελος όλης της ανθρωπότητας χωρίς διακρίσεις. Ειδικότερα στον τομέα της άμυνας και της ασφάλειας, η ΤΝ τείνει να δημιουργήσει ένα νέο πεδίο διεθνούς ανταγωνισμού εξοπλισμών και εργαλείο παγκόσμιας ισχύος, όπως αναλύεται στο επόμενο κεφάλαιο. Η επιπρόσθετη αρχή της επεξηγησιμότητας εκφράζει την ανάγκη κατανόησης και λογοδοσίας στις διαδικασίες λήψης απόφασης των συστημάτων ΤΝ.

Στο πολιτικό επίπεδο, μόλις τον Οκτώβριο του 2023 εκδόθηκε από την προεδρία των ΗΠΑ εκτελεστικό διάταγμα[2] για την ασφαλή και αξιόπιστη χρήση της τεχνητής

[2] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

νοημοσύνης. Στο διάταγμα καθορίζεται η υποχρέωση των εταιρειών ΤΝ να κοινοποιούν τα αποτελέσματα ελέγχων και άλλες κρίσιμες πληροφορίες των μοντέλων στην αμερικανική κυβέρνηση, η ανάπτυξη προτύπων για την ασφάλεια, προστασία και αξιοπιστία των συστημάτων ΤΝ αλλά και την αναγνώριση απάτης και παραπλάνησης μέσω ΤΝ και η ανάπτυξη εργαλείων κυβερνοασφάλειας για την αντιμετώπιση ευπαθειών. Ειδική μνεία γίνεται στα θέματα ιδιωτικότητας αλλά και στη χρήση της ΤΝ στην αντιμετώπιση του εγκλήματος καθώς και στη στρατιωτική χρήση της ΤΝ [98]. Χαρακτηριστική είναι επίσης η αναφορά στη δυνατότητα των ενόπλων δυνάμεων να αντιμετωπίσουν τη χρήση ΤΝ από τους αντιπάλους.

Σε ευρωπαϊκό επίπεδο, η Οδηγία NIS αποτέλεσε το πρώτο δεσμευτικό κανονιστικό κείμενο για την καθιέρωση της κυβερνοασφάλειας δικτύων και πληροφοριακών συστημάτων (ΠΣ) σε επίπεδο ΕΕ [99]. Η δεύτερη έκδοσή της NIS2 εφαρμόζεται από το 2023 και επεκτείνει το πεδίο εφαρμογής της σε ένα ευρύ πλαίσιο δραστηριοτήτων στις οποίες χρησιμοποιούνται ΠΣ, ωστόσο δεν καλύπτει θέματα ασφαλείας που ανακύπτουν από την εισαγωγή της ΤΝ. Αντίθετα κάνει αναφορές σε χρήση της ΤΝ στο πλαίσιο της ανίχνευσης και πρόληψης κυβερνοεπιθέσεων. Η πρώτη προσπάθεια στο συγκεκριμένο τομέα σε επίπεδο ΕΕ, είναι η Πράξη περί Τεχνητής Νοημοσύνης (AI Act), για την οποία το Ευρωπαϊκό Κοινοβούλιο και το Συμβούλιο της Ευρώπης κατέληξαν σε προκαταρκτική συμφωνία τον Δεκέμβριο του 2023 [100]. Στο σχέδιο νόμου περιλαμβάνονται απαγορευμένες χρήσεις συστημάτων ΤΝ που αφορούν βιομετρικά χαρακτηριστικά, μη στοχευμένη απόσπαση εικόνων προσώπου για τη δημιουργία βάσεων δεδομένων αναγνώρισης προσώπου, αναγνώριση συναισθημάτων στο χώρο εργασίας και στα εκπαιδευτικά ιδρύματα, χειραγώγηση ανθρώπινης συμπεριφοράς και εκμετάλλευση ανθρώπινων αδυναμιών. Προβλέπει ωστόσο την ίδια στιγμή εξαιρέσεις για συλλογή και χρήση βιομετρικών δεδομένων ακόμα και σε πραγματικό χρόνο εφόσον υπάρχει δικαστική έγκριση και για συγκεκριμένα εγκλήματα. Ταυτόχρονα στο νόμο αναφέρεται ρητώς ότι οι προβλέψεις του δεν αφορούν συστήματα ΤΝ που αναπτύσσονται για στρατιωτικούς σκοπούς [8].

Η Ευρωπαϊκή Επιτροπή έχει εκδώσει οδηγίες για την αξιόπιστη ΤΝ[3] στις οποίες παρέχεται ένα πλαίσιο που βασίζεται σε 4 ηθικές αρχές (σεβασμός της ανθρώπινης

[3] Ethics guidelines for trustworthy AI

αυτονομίας, πρόληψη βλάβης, δικαιοσύνη, επεξηγησιμότητα) και 7 απαιτήσεις (ανθρώπινη δράση και εποπτεία, τεχνική ευρωστία και ασφάλεια, ιδιωτικότητα και διακυβέρνηση δεδομένων, διαφάνεια, διαφορετικότητα, μη διάκριση και δικαιοσύνη, κοινωνική και περιβαλλοντική ευημερία, λογοδοσία) [101]. Σε συνέχεια αυτού εκδόθηκε η λίστα αυτοαξιολόγησης ALTAI με την οποία παρέχονται οδηγίες εφαρμογής του πλαισίου [102]. Η ALTAI αποτελεί ένα χρήσιμο εργαλείο που συνεισφέρει στη διακυβέρνηση της ΤΝ σε έναν οργανισμό και την εκτίμηση του επιπέδου ωριμότητας σε θέματα ΤΝ, ωστόσο μια αδυναμία είναι ότι δεν παρέχει κατευθύνσεις για τη βελτίωση των τομέων στους οποίους εντοπίζονται ελλείψεις [103].

Εκτός όμως από χώρες και διεθνείς οργανισμούς η ανάγκη θέσπισης κανόνων έχει αναγνωρισθεί και από τις επιχειρήσεις που δραστηριοποιούνται στην ΤΝ. Η Google έχει αναπτύξει το δικό της πλαίσιο ασφαλούς χρήσης ΤΝ[4] το οποίο περιλαμβάνει 6 στοιχεία [104]. Το πρώτο στοιχείο αφορά στην ανάπτυξη μέτρων και ελέγχων ασφαλείας σε ολόκληρο το οικοσύστημα ΤΝ του οργανισμού. Η ανάπτυξη μέτρων ασφαλείας αποτελεί μια δυναμική διαδικασία καθώς ο οργανισμός πρέπει να παρακολουθεί τις εξελίξεις και να προσαρμόζεται στα αναπτυσσόμενα μοντέλα απειλών. Το δεύτερο στοιχείο αφορά στον εντοπισμό και στην έγκαιρη αντίδραση σε κυβερνοπεριστατικά με επέκταση των υπαρχόντων συστημάτων κυβερνοασφάλειας στα συστήματα ΤΝ. Τρίτο στοιχείο αποτελεί η χρήση της ΤΝ για την αντιμετώπιση κυβερνοπεριστατικών. Θεωρώντας δεδομένη τη χρήση της ΤΝ από τους αντιπάλους, η ΤΝ μπορεί να αποτελέσει μέσο αποτελεσματικής άμυνας σε νέες απειλές. Τέταρτο στοιχείο αποτελεί η εναρμόνιση των ελέγχων σε όλα τα στάδια του κύκλου ζωής ενός συστήματος ΤΝ. Γενικότερα τα μέτρα ασφαλείας δεν θα πρέπει να είναι αποσπασματικά και όταν χρησιμοποιούνται διαφορετικά πλαίσια και κανονισμοί ασφαλείας θα πρέπει να εξασφαλίζεται ότι δεν υπάρχουν κενά σε συγκεκριμένα στάδια της διαδικασίας ανάπτυξης και χρήσης ενός συστήματος ΤΝ. Πέμπτο στοιχείο αποτελεί η προσαρμογή των ελέγχων σε νέες απειλές και η ανατροφοδότηση από περιστατικά ασφαλείας. Στο στοιχείο αυτό περιλαμβάνονται τεχνικές όπως η ενισχυτική μάθηση με βάση τα περιστατικά και η ανατροφοδότηση των χρηστών. Περιλαμβάνονται επίσης βήματα όπως η επικαιροποίηση των συνόλων δεδομένων

[4] Secure AI Framework

εκπαίδευσης, η αναπροσαρμογή των μοντέλων ώστε να ανταποκρίνονται καλύτερα στις επιθέσεις και η ενσωμάτωση δυνατοτήτων ανίχνευσης ανωμαλιών. Έκτο στοιχείο αποτελεί η ανάλυση επικινδυνότητας σε όλες τις επιχειρησιακές διαδικασίες στις οποίες εμπλέκεται ΤΝ για τον καθορισμό ελέγχων απόδοσης.

Ειδικά για τον στρατιωτικό τομέα, το ΝΑΤΟ έχει αναπτύξει στρατηγική για τη στρατιωτική χρήση της ΤΝ η οποία βασίζεται σε 6 αρχές: νομιμότητα, ευθύνη και υπευθυνότητα, επεξηγησιμότητα και ιχνηλασιμότητα, αξιοπιστία, κυβερνησιμότητα, και μετριασμός της προκατάληψης [105]. Η αρχή της νομιμότητας αφορά τη συμμόρφωση των συστημάτων με την εθνική και παγκόσμια νομοθεσία. Η αρχή της ευθύνης και της υπευθυνότητας αφορά στη διατήρηση του ελέγχου και τη μη αποποίηση των ευθυνών από τους ανθρώπους. Η αρχή επεξηγησιμότητας και ιχνηλασιμότητας αφορά στη διασφάλιση της διαφάνειας των διαδικασιών και την ανάπτυξη μηχανισμών ελέγχου και επικύρωσης της απόδοσης. Η αρχή της αξιοπιστίας αφορά τον αυστηρό καθορισμό περιπτώσεων χρήσης σε όλο τον κύκλο ζωής ενός συστήματος ΤΝ ο οποίος θα πιστοποιείται με καθορισμένες διαδικασίες. Η αρχή της κυβερνησιμότητας αφορά στη διατήρηση του ελέγχου το συστημάτων από τους ανθρώπους και τη δυνατότητα απενεργοποίησης σε περίπτωση μη αναμενόμενης συμπεριφοράς. Τέλος, η αρχή του μετριασμού της προκατάληψης αφορά στην επιλογή των δεδομένων εκπαίδευσης ώστε να είναι αντιπροσωπευτικά του προβλήματος και κατά το δυνατόν απαλλαγμένα από προκαταλήψεις.

Στον τομέα της τυποποίησης, ο οργανισμός ISO, είτε έχει εκδώσει, είτε βρίσκεται στη διαδικασία έκδοσης μιας σειράς προτύπων σχετικών με την ΤΝ. Σύμφωνα με μελέτη του ENISA[106], ο αριθμός των προτύπων ISO που σχετίζονται με την κυβερνοασφάλεια και την ΤΝ ανέρχεται σε 33, στα οποία προστίθενται άλλα 27 της σειράς ISO 27000 τα οποία έχουν εφαρμογή σε όλα τα πληροφοριακά συστήματα και κατ' επέκταση στα συστήματα ΤΝ. Στη συνέχεια γίνεται μια σύντομη αναφορά στα κυριότερα από αυτά.

Το πρότυπο ISO/IEC 42001 αποτελεί το αντίστοιχο του ISO 27001 στον τομέα της ΤΝ και παρέχει οδηγίες για ένα ολοκληρωμένο σύστημα διαχείρισης συστημάτων τεχνητής νοημοσύνης, εξετάζοντας θέματα ασφάλειας, προστασίας, δικαιοσύνης, διαφάνειας, ποιότητας δεδομένων και γενικότερης ποιότητας σε όλες τις φάσεις του

κύκλου ζωής ενός συστήματος ΤΝ. Στα αναλογία με το πρότυπο ISO 27001 για τη διαχείριση συστημάτων πληροφορικής, το ISO 42001, στο Παράρτημα Α περιλαμβάνει 37 αντίμετρα, ενώ το Παράρτημα Β περιλαμβάνει την ανάλυση των αντιμέτρων σε αναλογία με το πρότυπο ISO 27002. Ως πηγές κινδύνων αναγνωρίζονται:

- Το επίπεδο αυτοματοποίησης
- Η έλλειψη διαφάνειας και επεξηγησιμότητας
- Η πολυπλοκότητα του περιβάλλοντος ΤΠΕ
- Προβλήματα του κύκλου ζωής
- Προβλήματα υλικού συστημάτων
- Η ετοιμότητα της τεχνολογίας
- Κίνδυνοι που σχετίζονται με τη μηχανική μάθηση

Το πρότυπο ISO/IEC 24030 καταγράφει 132 περιπτώσεις χρήσης της ΤΝ. Στον τομέα της ασφάλειας και της άμυνας εντάσσονται 5 περιπτώσεις χρήσης:

- Ανάλυση συμπεριφοράς και συναισθήματος
- Λύση AI (νοημοσύνη σμήνους) για την ανίχνευση επιθέσεων σε περιβάλλον IoT
- Χρήση ρομποτικής λύσης για την αστυνόμευση και τον έλεγχο της κυκλοφορίας
- Ρομποτική λύση για την αντικατάσταση της ανθρώπινης εργασίας σε επικίνδυνες συνθήκες
- Μη παρεμβατική ανίχνευση κακόβουλου λογισμικού

Αναφορικά με τα ηθικά θέματα που ανακύπτουν από τα συστήματα ΤΝ έχει εκδοθεί το πρότυπο ISO/IEC TR24368. Το εν λόγω πρότυπο περιλαμβάνει 14 ηθικές αρχές που αφορούν όχι μόνο τους δημιουργούς και τους παρόχους συστημάτων ΤΝ αλλά και τους χειριστές και καλύπτουν θέματα όπως η υπερβολική εξάρτηση σε συστήματα ΤΝ (κατάχρηση), η ελλιπής χρήση που μπορεί να οδηγήσει σε αρνητικά αποτελέσματα (απαξίωση) και η επαναχρησιμοποίηση συστημάτων ΤΝ σε τομείς για τους οποίους δεν έχουν σχεδιαστεί ή δοκιμαστεί (κακόβουλη χρήση).

Στον τομέα της διαχείρισης κινδύνου ανάπτυξης συστημάτων ΤΝ, το ινστιτούτο NIST, έχει εκδώσει το πλαίσιο ανάλυσης κινδύνου ΤΝ[5], στο οποίο καταγράφονται 7 απαιτούμενα χαρακτηριστικά ενός αξιόπιστου συστήματος ΤΝ [24]: έγκυρο και αξιόπιστο, προστατευμένο, ασφαλές και ανθεκτικό, υπεύθυνο και διαφανές, επεξηγήσιμο και ερμηνεύσιμο, ενισχυμένο με προστασία της ιδιωτικής ζωής και δίκαιο με διαχείριση επιβλαβών προκαταλήψεων. Στον τομέα αντιμετώπισης επιθέσεων στους αλγορίθμους μηχανικής μάθησης στο πρότυπο NIST AI 100-2e2023 καταγράφονται οι τύποι επιθέσεων ανταγωνιστικής μάθησης και οι τρόποι αντιμετώπισής τους. Οι επιθέσεις ταξινομούνται σε 5 διαστάσεις: τύπο ΤΝ (προγνωστική ή παραγωγική), τύπο μάθησης και στάδιο του κύκλου ζωής κατά το οποίο πραγματοποιείται η επίθεση, στόχοι επιτιθεμένου, δυνατότητες επιτιθεμένου και γνώση επιτιθεμένου για τη διαδικασία μάθησης.

Συνοψίζοντας, διαπιστώνεται έντονη κινητικότητα στον τομέα της ανάπτυξης προτύπων και λιγότερο στην ανάπτυξη ρυθμιστικού πλαισίου το οποίο θα συμβάλει στην κυβερνοασφάλεια των συστημάτων ΤΝ και καταδεικνύεται η επίγνωση της διεθνούς κοινότητας επί των κινδύνων που συνοδεύουν τη διάχυση της ΤΝ.

[5] AI Risk Management Framework (AI RMF 1.0)

ΚΕΦΑΛΑΙΟ 6

ΣΤΡΑΤΗΓΙΚΕΣ ΕΠΙΠΤΩΣΕΙΣ

Η εξάπλωση της ΤΝ εκτός από τα τεχνολογικά, ηθικά και νομικά ζητήματα που δημιουργεί, προκαλεί και στρατηγικές επιπτώσεις. Η οπλοποίηση των αλγορίθμων μηχανικής μάθησης και των μεγάλων γλωσσικών μοντέλων δεν προσδίδει μόνο νέες δυνατότητες στο πεδίο της μάχης αλλά δημιουργεί και ένα νέο πεδίο γεωπολιτικού ανταγωνισμού. Ταυτόχρονα η διάδοση της ΤΝ σε μη κρατικούς δρώντες, η ταχεία εξέλιξη των τεχνολογιών που βασίζονται στην ΤΝ και η αυξανόμενη πολυπλοκότητα των συγκεκριμένων συστημάτων δημιουργεί κινδύνους για την εθνική ασφάλεια.

6.1. Επιπτώσεις στην Εθνική Άμυνα και Ασφάλεια

Η ΤΝ στον τομέα της άμυνας θεωρείται ένας πολλαπλασιαστής ισχύος, ενώ για αρκετούς μελετητές θα μετασχηματίσει τη φύση του πολέμου [107], [108]. Η ενσωμάτωση της ΤΝ στα συστήματα προσομοίωσης μάχης, τα επιχειρησιακά συστήματα, τα συστήματα λήψης αποφάσεων, τα συστήματα κυβερνοάμυνας και τα αυτόνομα συστήματα, υπόσχεται μια ταχύτερη και αποτελεσματικότερη λήψη αποφάσεων με τη συνεργασία ανθρώπου και μηχανής. Ταυτόχρονα καταγράφονται απόψεις ότι η ΤΝ ενισχύει ακόμα περισσότερο τη σημασία των ανθρώπινων αποφάσεων στον πόλεμο καθώς όσο καλή μπορεί να είναι η ΤΝ στη λήψη αποφάσεων τακτικής, τόσο εσφαλμένη μπορεί να είναι στη λήψη στρατηγικών αποφάσεων [109], [110].

Είναι γεγονός ότι η τεχνητή νοημοσύνη δύναται να συνεισφέρει στη στρατιωτική ισχύ, προσφέροντας δυνατότητες για βελτιωμένη επίγνωση της κατάστασης, χρήση αυτόνομων συστημάτων και στοχοποίηση ακριβείας. Τα κράτη έχουν αυξανόμενο κίνητρο να επενδύσουν στην αμυντική χρήση της ΤΝ προσδοκώντας να αποκτήσουν στρατηγικό πλεονέκτημα έναντι των αντιπάλων τους όσον αφορά στην αποτελεσματικότητα στο πεδίο της μάχης, την αποτροπή και την ασφάλεια.

Εξετάζοντας τη σύνδεση ΤΝ και κυβερνοασφάλειας από πλευράς εθνικής ασφαλείας, το ενδιαφέρον επικεντρώνεται στις αμυντικές και επιθετικές δυνατότητες που δημιουργούνται, στις αδυναμίες των αλγορίθμων ΤΝ μέσω της ανταγωνιστικής

μάθησης αλλά και στην επίδραση των συστημάτων μηχανικής μάθησης στις πληροφοριακές επιχειρήσεις [111].

Από πλευράς αμυντικών δυνατοτήτων, όπως αναλύθηκε στα προηγούμενα κεφάλαια, η ΤΝ βρίσκει εφαρμογές στον έγκαιρο εντοπισμό απειλών. Ο πρώτος κρίκος της φονικής αλυσίδας (αναγνώριση) μπορεί να προηγείται ικανό χρονικό διάστημα προ της εκδήλωσης μιας κυβερνοεπίθεσης. Κατά συνέπεια ο εντοπισμός προτύπων σε προσπάθειες συλλογής πληροφοριών μπορεί να παρέχει έγκαιρη προειδοποίηση για μελλοντικούς στόχους. Σε επίπεδο εθνικής ασφάλειας δεν αρκεί ο εντοπισμός αλλά απαιτείται η απαγόρευση κυβερνοεπιθέσεων. Με τη χρήση αλγορίθμων ΤΝ είναι εφικτή η παρενόχληση των επιτιθέμενων μέσω της εκδήλωσης επιθέσεων στα συστήματα από τα οποία εκδηλώνεται η κυβερνοεπίθεση. Ωστόσο η αυτονόμηση της συγκεκριμένης διαδικασίας εγκυμονεί κινδύνους, καθώς προκαλεί κλιμάκωση η οποία μπορεί να δημιουργήσει αντιδράσεις όταν ο επιτιθέμενος έχει αυξημένες δυνατότητες. Η απόδοση ευθυνών έχει ιδιαίτερη σημασία σε επίπεδο εθνικής ασφαλείας. Αποτελεί μια γενικότερη πρόκληση στο επίπεδο της κυβερνοσφάλειας, καθώς οι επιτιθέμενοι επιχειρούν να αποκρύψουν την ταυτότητά τους. Ωστόσο η ΤΝ παρέχει νέες δυνατότητες, καθώς για παράδειγμα αλγόριθμοι μη επιβλεπόμενης μάθησης αναλύουν τμήματα κώδικα για τον εντοπισμό APT ή τη συλλογή στοιχείων για την απόδοση ευθυνών σε μελλοντικές επιθέσεις.

Από πλευράς επιθετικών δυνατοτήτων, η ΤΝ παρέχει αυξημένες δυνατότητες εντοπισμού αδυναμιών. Η συγκεκριμένη χρήση έχει διττό ρόλο. Την ανάλυση επικινδυνότητας εγχώριων πληροφοριακών συστημάτων που χρησιμοποιούνται σε κρίσιμες υποδομές και περιβάλλοντα ασφαλείας μέσω της προσομοίωσης επιθετικών δυνατοτήτων του αντιπάλου (red teaming), αλλά και τον εντοπισμό αδυναμιών σε συστήματα του αντιπάλου τα οποία μπορούν να χρησιμοποιηθούν σε περίπτωση κλιμάκωσης της έντασης. Σημαντική τεχνική των επιθετικών επιχειρήσεων κυβερνοχώρου είναι η πρακτική spear-fishing δηλαδή η εισαγωγή κακόβουλου κώδικα σε στόχους ή η απόκτηση μη εξουσιοδοτημένης πρόσβασης μέσω μεθόδων κοινωνικής μηχανικής. Οι αλγόριθμοι της ΤΝ μπορούν να βελτιώσουν την απόδοση των συγκεκριμένων τεχνικών μέσω της προσαρμογής των μηνυμάτων με βάση το προφίλ του στόχου το οποίο δημιουργείται μέσω των αλληλεπιδράσεών του στα

κοινωνικά δίκτυα και του δικτύου επαφών του. Εκτός από την εισαγωγή κακόβουλου κώδικα σε ένα σύστημα, σημαντική είναι η επιβίωσή του και η εξάπλωσή του. Τα «σκουλήκια» που βασίζονται σε ΤΝ (AI worms) αποτελούν μια αναδυόμενη απειλή όπου αυτόνομοι πράκτορες επιχειρούν τη διάδοση σε πληροφοριακά συστήματα χωρίς την επέμβαση του χρήστη. Για παράδειγμα, σε ερευνητικό επίπεδο έχει αναπτυχθεί worm το οποίο χρησιμοποιεί μεγάλα γλωσσικά μοντέλα ελεύθερης πρόσβασης για την υποκλοπή προσωπικών στοιχείων και την αποστολή ανεπιθύμητων μηνυμάτων μέσα από μηνύματα ηλεκτρονικού ταχυδρομείου χωρίς την αλληλεπίδραση του χρήστη [112]. Ενώ η ΤΝ υπόσχεται αυξημένες δυνατότητες εντοπισμού απειλών, δημιουργεί επίσης βελτιωμένες μεθόδους απόκρυψης της ταυτότητας του επιτιθέμενου και παρεμπόδιση της προσπάθειας του αμυνόμενου για τη συγκέντρωση ψηφιακών πειστηρίων. Για παράδειγμα με τη χρήση μεθόδων ανταγωνιστικής μάθησης είναι εφικτή η δημιουργία deep-fakes τα οποία δεν εντοπίζονται από αλγόριθμους μηχανικής μάθησης [113]. Τέλος, ιδιαίτερο ενδιαφέρον σε επίπεδο εθνικής ασφάλειας έχουν οι καταστροφικές επιθέσεις. Με τη διάδοση της ΤΝ σε πολλαπλές εφαρμογές από τη χρήση σε κρίσιμες υποδομές μέχρι την ανάπτυξη αυτόνομων συστημάτων στο πεδίο της μάχης διευρύνεται η επιφάνεια επίθεσης. Για παράδειγμα η ανάκτηση του ελέγχου αυτόνομων συστημάτων στο πεδίο της μάχης από τον αντίπαλο, μπορεί να επιτρέψει την επαναχρησιμοποίησή τους εναντίον του αρχικού χρήστη τους.

Η ασφάλεια των κυβερνοφυσικών συστημάτων - ΚΦΣ (cyber-physical systems), δηλαδή έξυπνων συστημάτων που περιλαμβάνουν συνδυασμό φυσικών και υπολογιστικών δομικών στοιχείων [114] με την ενσωμάτωση στοιχείων ΤΝ γίνεται ακόμα πιο πολύπλοκη. Αναλόγως της περίπτωσης χρήσης, τα ΚΦΣ ελέγχουν τη λειτουργία κρίσιμων υποδομών (π.χ. ύδρευσης, παροχής ενέργειας, τηλεπικοινωνιών) αλλά και βιομηχανικών μονάδων. Με τη διάχυση της ΤΝ όλο και περισσότερα ΚΦΣ αναμένεται να αντικατασταθούν από νέα συστήματα με αυξημένο επίπεδο αυτονομίας δράσης και λήψης αποφάσεων. Το γεγονός αυτό έχει και μια θετική διάσταση καθώς πλήθος ΚΦΣ βασίζονται σε παρωχημένες εκδόσεις λειτουργικών συστημάτων ή λογισμικού τα οποία είναι ευάλωτα στις σύγχρονες κυβερνοεπιθέσεις. Με τη χρήση της ΤΝ αυξάνονται οι δυνατότητες άμυνας και αντίδρασης σε κυβερνοεπιθέσεις στα συγκεκριμένα συστήματα. Ωστόσο την ίδια στιγμή μειώνεται η δυνατότητα ελέγχου των

ΚΦΣ, τα οποία μπορούν είτε λόγω δυσλειτουργίας κάποιου αλγορίθμου είτε λόγω κάποιας κυβερνοεπίθεσης στο σύστημα ΤΝ να εμφανίσουν μη αναμενόμενη λειτουργία με καταστροφικές συνέπειες [80].

Καθώς οι δυνατότητες της ΤΝ, γίνονται προσιτές όχι μόνο σε προηγμένες τεχνολογικά και οικονομικά χώρες αλλά και σε μικρότερες χώρες αλλά και τρομοκρατικά δίκτυα καθίσταται όλο και πιο κρίσιμο να ανιχνεύονται, να εκτρέπονται και να περιορίζονται οι επιπτώσεις των επιθέσεων στα συστήματα ΤΝ μιας χώρας (στρατιωτικής και πολιτικής χρήσης), ενώ παράλληλα να υπονομεύονται τα αντίπαλα συστήματα ΤΝ. Στο πλαίσιο αυτό αναδύονται εμπλοκές ΤΝ εναντίον ΤΝ (κυρίως στο πλαίσιο αντιμετώπισης της παραπληροφόρησης), ενώ οι δυνατότητες που παρέχει η ΤΝ για την ενεργοποίηση εγκληματικών συμπεριφορών δεν πρέπει υποτιμάται [115].

Οι αδυναμίες των συστημάτων ΤΝ αναλύθηκαν στο προηγούμενα κεφάλαια και καταδεικνύουν τις εξειδικευμένες απαιτήσεις ασφαλείας που δημιουργεί η χρήση της ΤΝ σε κρίσιμες εφαρμογές. Σημαντικό στοιχείο της επίδρασης της ΤΝ στην εθνική ασφάλεια αποτελεί η χρήση της στο πλαίσιο πληροφοριακών επιχειρήσεων. Η αυτοματοποίηση της προπαγάνδας και της παραπλάνησης μέσω των δυνατοτήτων επεξεργασίας φυσικής γλώσσας αποτελεί απειλή για πολιτικές διαδικασίες που έχουν επιπτώσεις στην εθνική ασφάλεια. Για παράδειγμα η ΤΝ παρέχει βελτιωμένες δυνατότητες στις προσπάθειες επηρεασμού εκλογικών αποτελεσμάτων μέσα από τη διασπορά παραπλανητικών μηνυμάτων και deepfakes [116].

Μακροπρόθεσμα η πιθανότητα ανάπτυξης ΤΝ με γνωστικές ικανότητες συγκρίσιμες με εκείνες του ανθρώπου (Artificial General Intelligence - AGI), αποτελεί από μόνη της μια απειλή εθνικής ασφαλείας, καθώς μια τέτοια τεχνολογία στα χέρια μιας εταιρείας ή μιας χώρας μπορεί να έχει αποσταθεροποιητικό χαρακτήρα όχι μόνο στο πλαίσιο ενός κράτους αλλά παγκόσμια. Η εθνική κυριαρχία ενός κράτους που δεν μπορεί να ελέγξει τις δυνατότητες που αναπτύσσει μια εταιρεία ΤΝ στο έδαφός της ή δεν μπορεί να αντιμετωπίσει μια βλάβη που προκαλείται από τη δυσλειτουργία ενός συστήματος ΤΝ, μπορεί να τεθεί σε αμφισβήτηση τόσο από το εσωτερικό ακροατήριο όσο και από το εξωτερικό [117].

6.2. Γεωπολιτικές Επιδράσεις

Η κυβερνοασφάλεια έχει μεταξύ άλλων και γεωπολιτικές προεκτάσεις. Μια κυβερνοεπίθεση σε κρίσιμες υποδομές ενός κράτους μπορεί να προκαλέσει αντιδράσεις του πληττόμενου τόσο στον κυβερνοχώρο όσο και σε διαφορετικά πεδία επιχειρήσεων. Λαμβάνοντας υπόψη την αρχή της αναλογικότητας, το κράτος που έχει δεχθεί την επίθεση θα πρέπει αρχικά να εντοπίσει τον φορέα της κυβερνοεπίθεσης. Αν πρόκειται για μη κρατικό δρώντα με ατομικά κίνητρα, το κράτος μπορεί να λάβει ποινικά μέτρα ή ακόμα και να στοχοποιήσει τα συστήματα του συγκεκριμένου δρώντα. Αν πρόκειται για μη κρατικό δρώντα, ο οποίος όμως ενεργεί για λογαριασμό ή προς όφελος κάποιου κράτους, ο αμυνόμενος μπορεί να επιλέξει τα στοχοποιήσει τόσο τον επιτιθέμενο όσο και το κράτος που υποστηρίζει. Αν ο δρώντας είναι κάποιο κράτος το οποίο δρα ως ενδιάμεσος ή είναι ο δράστης της επίθεσης, το αμυνόμενο κράτος μπορεί να επικαλεστεί το δικαίωμα της αυτοάμυνας και να επιδιώξει τα χαρακτηρίσει την κυβερνοεπίθεση πολεμική ενέργεια και να αντιδράσει ανάλογα. Οι τρόποι αντίδρασης σε αυτή την περίπτωση δεν περιορίζονται στον κυβερνοχώρο και μπορεί να περιλαμβάνουν οικονομικές κυρώσεις ή ακόμα και φυσικές επιθέσεις σε υποδομές του κράτους που διέπραξε την επίθεση [118].

Λαμβάνοντας υπόψη τόσο την αυξανόμενη επιφάνεια επίθεσης από την διάχυση της τεχνητής νοημοσύνης σε όλες τις δραστηριότητες αλλά και τις επιπλέον δυνατότητες που παρέχει η ΤΝ σε στρατιωτικές εφαρμογές, γίνεται κατανοητό ότι ενισχύεται ο γεωπολιτικός αντίκτυπος της κυβερνοασφάλειας της ΤΝ. Επιπλέον καθώς μεγάλες εταιρείες πληροφορικής αναπτύσσουν και διαχειρίζονται μεγάλα γλωσσικά μοντέλα τα οποία τυγχάνουν παγκόσμιας εφαρμογής, αποτελούν δυνητικούς στόχους κυβερνοεπιθέσεων αντιπάλων κρατών. Στην περίπτωση αυτή μια δυνητική κυβερνοεπίθεση ή ακόμα και κάποια δολιοφθορά θα μπορούσε να έχει υψηλό αντίκτυπο και να προκαλέσει δυναμική αντίδραση.

Η διάχυση της ΤΝ σε στρατιωτικές εφαρμογές δημιουργεί για πολλούς αναλυτές μια νέα κούρσα εξοπλισμών [55], [93], [119]. Για παράδειγμα, οι ΗΠΑ ανακοίνωσαν τον Αύγουστο 2023 την πρωτοβουλία Replicator για την κατασκευή και παράδοση χιλιάδων αυτόνομων οχημάτων χαμηλού κόστους σε περίοδο 18 έως 24 μηνών για χρήση από τις αμερικανικές ένοπλες δυνάμεις. Η έμφαση δίνεται στις έξυπνες

δυνατότητες αλλά και στο χαμηλό κόστος και τις μεγάλες ποσότητες ώστε να υπάρχει η δυνατότητα χρήσης σε σμήνος αλλά και λόγω του μεγάλου αριθμού αναμενόμενων απωλειών. Η ανάγκες αυτές αποδίδονται ως απάντηση της ταχείας εξέλιξης της Κίνας στον τομέα αλλά και των διδαγμάτων από τον πόλεμο στην Ουκρανία όπου υπολογίζονται μέχρι και 10.000 απώλειες drone μηνιαίως [120].

Η ανάπτυξη δυνατοτήτων στον τομέα της ΤΝ ενισχύει το δίλλημα ασφαλείας μεταξύ των κρατών, καθώς αυξάνει την αβεβαιότητα για τις προθέσεις τους. Σε μια τέτοια κατάσταση, ένα κράτος εκλαμβάνει ως απειλή τις επενδύσεις στην ΤΝ ενός άλλου κράτους, με αποτέλεσμα να λαμβάνει αντίμετρα, τα οποία με τη σειρά τους κλιμακώνουν προϋπάρχουσες γεωπολιτικές εντάσεις. Η διττή χρήση των τεχνολογιών ΤΝ ενισχύει τις συγκεκριμένες ανησυχίες. Στο πλαίσιο αυτό σε πολλές περιπτώσεις η εξέλιξη της ΤΝ παραλληλίζεται με την εξάπλωση των πυρηνικών όπλων και διατυπώνονται προτάσεις για τη λήψη μέτρων περιορισμού της ΤΝ, ανάλογων με αυτά που έχουν ληφθεί για τον περιορισμό της εξάπλωσης των πυρηνικών όπλων [121], [122].

Η έννοια του ελέγχου εξοπλισμών πέρα από τα πυρηνικά και συμβατικά όπλα, με τη διάχυση της ΤΝ στο στρατιωτικό τομέα, επεκτείνεται και στον έλεγχο των δυνατοτήτων ΤΝ των κρατών [123]. Ο έλεγχος των δυνατοτήτων ΤΝ έγκειται στην ανάπτυξη και επιβολή προτύπων στο κύκλο ζωής συστημάτων ΤΝ, περιορισμό στις αποφάσεις που μπορούν να λαμβάνονται αυτόνομα από συστήματα ΤΝ, απαγορεύσεις στη χρήση ΤΝ για συγκεκριμένες δραστηριότητες κλπ. Εκτός από το πλαίσιο κανονισμών, υφίσταται ανάγκη ανάπτυξης μηχανισμών επαλήθευσης της συμμόρφωσης με αυτούς. Η υλοποίηση των παραπάνω προϋποθέτει βέβαια την πρόθεση συμμόρφωσης των κρατών, γεγονός που δεν είναι δεδομένο στο πλαίσιο των υφιστάμενων γεωπολιτικών ανταγωνισμών. Ταυτόχρονα μη κρατικοί δρώντες που δεν υπόκεινται σε σχετικούς περιορισμούς μπορούν να παρακάμπτουν τυχόν κανονισμούς που θα αναπτυχθούν αυξάνοντας τις απειλές εθνικής ασφάλειας.

Η ανάπτυξη δυνατοτήτων ΤΝ δεν εξαρτάται μόνο από τη χρήση εξελιγμένων αλγορίθμων αλλά και από τη διαθέσιμη υπολογιστική ισχύ. Η ανάπτυξη ισχυρότερων και ενεργειακά αποδοτικότερων ημιαγωγών, όπως κεντρικών επεξεργαστών (CPU) και τσιπ γραφικών (GPU) αλλά και εξειδικευμένων τσιπ τεχνητής νοημοσύνης, είναι

ζωτικής σημασίας για την ενίσχυση των επιδόσεων των συστημάτων ΤΝ και προσθέτουν ένα ακόμα πεδίο στην κούρσα εξοπλισμών. Καθώς οι δυνατότητες ανάπτυξης και παραγωγής τσιπ είναι περιορισμένες σε συγκεκριμένες χώρες, η πρόσβαση στις τεχνολογίες αυτές και η προστασία των καναλιών διανομής αποτελεί έναν ακόμη παράγοντα γεωπολιτικού ανταγωνισμού.

Ειδικότερα η παραγωγική ΤΝ αποτελεί σημείο ανταγωνισμού μεταξύ ΗΠΑ και Κίνας, σε συνδυασμό με την υφιστάμενη διαμάχη για το καθεστώς αυτονομίας της Ταιβάν. Αντίστοιχα χώρες στις οποίες δεν υπάρχει μεγάλη διαφάνεια ως προς τα επιτεύγματα της τεχνολογικής τους έρευνας όπως η Ρωσία, το Ιράν και η Βόρεια Κορέα κατά πάσα πιθανότητα θα επιδιώξουν να αναπτύξουν αντίστοιχες δυνατότητες. Επιπρόσθετα χώρες όπως το Ηνωμένο Βασίλειο, τα Ηνωμένα Αραβικά Εμιράτα, το Ισραήλ, η Ιαπωνία, η Ολλανδία, η Νότια Κορέα, και η Ινδία παρουσιάζουν έντονο ενδιαφέρον για την ανάπτυξη δυνατοτήτων στο συγκεκριμένο τομέα [115], [124]. Οι χώρες αυτές πιθανότατα θα επιδιώξουν να συνάψουν συμμαχίες και συνεργασίες με πιο ισχυρά κράτη ή και μεταξύ τους με στόχο την επίτευξη κοινών στόχων. Την ίδια στιγμή χώρες οι οποίες διεκδικούν ρόλο περιφερειακής δύναμης πιθανότατα θα επιδιώξουν να μεταφέρουν τεχνογνωσία συνεργαζόμενες με τις παραπάνω χώρες προκειμένου μεσοπρόθεσμα να αποκτήσουν εγχώριες δυνατότητες στον τομέα. Ο συνδυασμός των δυνατοτήτων που προσφέρει η παραγωγική ΤΝ με τις προϋπάρχουσες γεωπολιτικές αντιθέσεις μπορεί να έχει αποσταθεροποιητικό ρόλο. Για παράδειγμα, στο πλαίσιο των συνεχιζόμενων εντάσεων μετά την ρωσική εισβολή στην Ουκρανία, τροφοδοτείται ένας κύκλος εσφαλμένων αντιλήψεων μεταξύ ΗΠΑ και Ρωσίας σχετικά με τις προθέσεις και τις ικανότητες σε διαφόρους τομείς μεταξύ των οποίων και η ΤΝ με αποτέλεσμα να υπάρχει ο κίνδυνος ακούσιας κλιμάκωσης των πολεμικών επιχειρήσεων που θα επηρεάσει αρνητικά τη διεθνή ασφάλεια [125].

Επανερχόμενοι στην περίπτωση της AGI, το κράτος που θα αναπτύξει πρώτο αντίστοιχες δυνατότητες θα μπορούσε να αποκτήσει στρατηγικά πλεονεκτήματα στον στρατιωτικό, οικονομικό και γεωπολιτικό τομέα, επηρεάζοντας την παγκόσμια ισορροπία ισχύος. Ωστόσο εκφράζονται αμφιβολίες κατά πόσο τα συστήματα τεχνητής νοημοσύνης θα φτάσουν σε αυτό το επίπεδο γνωστικής ικανότητας μέσα στα επόμενα 20 χρόνια [115].

Συνοψίζοντας, η ΤΝ όπως και όλες οι καινοτόμες και διττής χρήσης τεχνολογίες στο παρελθόν έχει ευρύτερο αντίκτυπο πέραν του πεδίου της τεχνολογίας. Το γεγονός ότι η ΤΝ βασίζεται κατά κύριο λόγο σε αλγορίθμους και δεδομένα και κατά δεύτερο λόγο σε υλικό, αυξάνει την αβεβαιότητα των κρατών και δημιουργεί την ανάγκη ανάπτυξης καθεστώτων ελέγχου των δυνατοτήτων στον τομέα της ΤΝ σε αναλογία με την πυρηνική τεχνολογία στο παρελθόν.

ΚΕΦΑΛΑΙΟ 7

ΣΥΜΠΕΡΑΣΜΑΤΑ

7.1 Ανακεφαλαίωση

Ανακεφαλαιώνοντας, στο 2ο κεφάλαιο πέρα από τις βασικές έννοιες της τεχνητής νοημοσύνης, παρουσιάσθηκαν οι διαφορετικές προσεγγίσεις επίλυσης προβλημάτων που χρησιμοποιούνται στην ΤΝ και δόθηκε μεγαλύτερη έμφαση στη μηχανική μάθηση. Στη συνέχεια παρουσιάσθηκαν οι βασικοί αλγόριθμοι που υποστηρίζουν τις διαφορετικές προσεγγίσεις. Διαπιστώνεται ότι υπάρχει πλέον μια μεγάλη γκάμα αλγορίθμων οι οποίοι μπορούν να επιλύσουν διαφορετικά προβλήματα. Σημαντικό στοιχείο είναι ότι οι αδυναμίες ενός αλγορίθμου μπορούν να αντιμετωπισθούν από κάποιον άλλο αλγόριθμο. Το γεγονός αυτό βέβαια αυξάνει την πολυπλοκότητα των συστημάτων ΤΝ. Καταδεικνύεται επίσης η σημασία των δεδομένων εκπαίδευσης, καθώς από την ποιότητά τους κρίνεται στις περισσότερες περιπτώσεις η εκπαίδευση του μοντέλου. Από την εξέταση των διαφορετικών μοντέλων κύκλου ζωής συστημάτων ΤΝ, τεκμηριώνεται η ανάγκη ελέγχου και ασφαλείας από τη σχεδίαση μέχρι και την απόσυρση ενός συστήματος ΤΝ. Ειδικά για εφαρμογές άμυνας και ασφαλείας η ανάλυση επικινδυνότητας θα πρέπει να λαμβάνει υπόψη όλα τα επιμέρους στάδια του κύκλου ζωής και να προτείνει αντίμετρα για το κάθε ένα από αυτά. Επιπλέον θα πρέπει να λαμβάνεται υπόψη το περιβάλλον στο οποίο θα λειτουργήσει το σύστημα ΤΝ.

Στο 3ο κεφάλαιο εξετάσθηκε η αλληλεπίδραση της ΤΝ με την κυβερνοασφάλεια. Η ασφάλεια είναι συνυφασμένη με κάθε αμυντικό σύστημα καθώς θα πρέπει να εξασφαλίζεται ότι δεν διατρέχουν κίνδυνο από τη χρήση του, αφενός οι χειριστές του και αφετέρου τα φίλια τμήματα και ο άμαχος πληθυσμός. Ομοίως για ένα σύστημα αστυνόμευσης θα πρέπει να διασφαλίζεται η αμεροληψία και η συμμόρφωση με την ισχύουσα νομοθεσία. Με τη χρήση της ΤΝ στα εν λόγω συστήματα αυξάνονται και οι ανάγκες προστασίας, καθώς εισάγονται νέες ευπάθειες που είναι συνυφασμένες με την ΤΝ. Τα μοντέλα που παρουσιάσθηκαν αποτελούν χρήσιμα εργαλεία ανάλυσης επικινδυνότητας των συγκεκριμένων συστημάτων, η οποία είναι απαραίτητη για τον καθορισμό των απαιτούμενων αντιμέτρων. Από πλευράς επικινδυνότητας οι επιθέσεις

prompt injection αποτελούν μια από τις σημαντικότερες απειλές καθώς εκτελούνται κατά τη λειτουργία του συστήματος και μπορούν να οδηγήσουν σε καταστροφικά αποτελέσματα, ειδικά όταν πρόκειται για συστήματα που χειρίζονται οπλικά συστήματα. Τα παραδείγματα παραβίασης ασφαλείας μεγάλων γλωσσικών μοντέλων ή κακόβουλων γλωσσικών μοντέλων που είναι ελεύθερα διαθέσιμα καταδεικνύουν την ανάγκη οι χώρες να αναπτύσσουν δικά τους γλωσσικά μοντέλα και να αποφεύγεται η επαναχρησιμοποίηση ευρέως διαδεδομένων LLM. Η παραπληροφόρηση αποτελεί μια από τις άμεσα εφαρμόσιμες χρήσεις της ΤΝ και αποτελεί μια απειλή εθνικής ασφαλείας που αφορά όλα τα κράτη. Από την ανάλυση των χρήσεων της ΤΝ σε αμυντικό ρόλο, γίνεται αντιληπτό ότι η ανάπτυξη δυνατοτήτων στον τομέα της ΤΝ μπορεί να συνεισφέρει στην ανθεκτικότητα ενός κράτους σε κυβερνοεπιθέσεις.

Στο 4ο κεφάλαιο, εξετάζοντας τις χρήσεις της ΤΝ σε υπηρεσίες ασφαλείας και επιβολής του νόμου, διαπιστώνονται προκλήσεις όσον αφορά στην ποιότητα των δεδομένων εκπαίδευσης των μοντέλων αλλά και της νομιμότητας συλλογής των δεδομένων, η οποία θα πρέπει να συμμορφώνεται με την ισχύουσα νομοθεσία. Η χρήση της ΤΝ σε στρατιωτικές εφαρμογές, είναι συνυφασμένη με τις νέες δυνατότητες που παρέχουν τα αυτόνομα συστήματα και είναι δεδομένο ότι θα αποτελέσει στόχο για την εξουδετέρωση των συστημάτων αυτών. Στο πλαίσιο κατανόησης της συγκεκριμένης πρόκλησης, πραγματοποιήθηκε ανάλυση της εφαρμογής του cyber kil chain εναντίον ενός αυτόνομου συστήματος στο πεδίο της μάχης. Μέσω της ανάλυσης επιδιώχθηκε να καταδειχθούν τα σημεία στα οποία θα πρέπει να επικεντρωθεί η κυβερνοασφάλεια των συστημάτων αυτών. Από τα παραδείγματα χρήσεων που παρουσιάσθηκαν, προκύπτει ότι υφίστανται ήδη εφαρμόσιμες λύσεις εισαγωγής της ΤΝ σε συστήματα άμυνας και ασφάλειας. Η κυβερνοασφάλεια των συστημάτων αυτών αλλά και η τήρηση κανόνων κατά την ανάπτυξη και τη χρήση τους είναι καθοριστικές τόσο για την αποδοχή τους όσο και για την αποτελεσματικότητά τους.

Στο 5ο κεφάλαιο, αναλύθηκαν τα ηθικά ζητήματα που ανακύπτουν από την εξάπλωση της ΤΝ σε εφαρμογές άμυνας και ασφάλειας. Ιδιαίτερη σημασία έχουν τα ζητήματα εμπιστοσύνης τα οποία μπορεί να οδηγήσουν στην αποτυχία ενός συστήματος εάν δημιουργηθούν αμφιβολίες για την αποτελεσματικότητά τους από τα πρώιμα στάδια χρήσης τους. Αντίθετα η υπερβολική εμπιστοσύνη στην ΤΝ μπορεί να επιφέρει αρνητικά αποτελέσματα στο πλαίσιο χρήσης ενός συστήματος στο πεδίο της μάχης,

τα οποία μπορούν να κυμαίνονται από αδελφοκτόνα πυρά μέχρι την προσβολή άμαχου πληθυσμού. Εξετάζοντας το ρυθμιστικό πλαίσιο το οποίο αναπτύσσεται τόσο σε επίπεδο νομοθεσίας όσο και σε επίπεδο τυποποίησης και διεθνών προτύπων, διαπιστώνεται ότι λαμβάνονται υπόψη οι ηθικοί προβληματισμοί. Ωστόσο η εξαίρεση των στρατιωτικών εφαρμογών από περιορισμούς που τίθενται με υψηλού επιπέδου πρωτοβουλίες όπως το AI Act (ΕΕ) και η εκτελεστική διαταγή για την ΤΝ (ΗΠΑ), καταδεικνύει το γεγονός ότι τα κράτη αναγνωρίζουν τη σημασία των στρατιωτικών εφαρμογών της ΤΝ και θέλουν να διατηρήσουν ελευθερία κινήσεων προκειμένου να μπορούν να αντιμετωπίσουν απειλές από συστήματα ΤΝ ανταγωνιστικών ή εχθρικών χωρών.

Στο 6ο κεφάλαιο αναλύθηκαν οι στρατηγικές επιπτώσεις της ΤΝ στο επίπεδο της γεωπολιτικής και της στρατηγικής των κρατών. Η χρήση της ΤΝ στο πεδίο του κυβερνοπολέμου αλλά και η κυβερνοασφάλεια της ΤΝ έχουν διεθνές αντίκτυπο και η ανάπτυξη διαδικασιών ελέγχου των δυνατοτήτων της ΤΝ έχει ουσιαστική σημασία όχι μόνο με την προοπτική της AGI αλλά και για την αποκατάσταση της εμπιστοσύνης και της διεθνούς ασφάλειας. Η εύκολη πρόσβαση στην ΤΝ από εγκληματικές ή τρομοκρατικές ομάδες αποτελεί επίσης στοιχείο που ενισχύει την ανάγκη διεθνούς συνεργασίας σε θέματα κυβερνοασφάλειας και τεχνητής νοημοσύνης.

7.2 Συμπεράσματα - Προοπτικές

Στον τομέα της κυβερνοασφάλειας, τα συστήματα ΤΝ παρουσιάζουν καινοτομίες, οι οποίες δημιουργούν νέους κινδύνους. Οι έξοδοι ενός συστήματος ΤΝ πολύ συχνά δεν είναι εύκολο να εξηγηθούν ή να επαληθευθούν. Ο κύκλος ζωής ενός συστήματος ΤΝ είναι πολύ διαφορετικός από ένα παραδοσιακό πληροφοριακό σύστημα, δημιουργώντας νέες απαιτήσεις κυβερνοασφάλειας. Σημαίνοντα ρόλο στην ποιότητα και την αποτελεσματικότητα ενός συστήματος ΤΝ κατέχει η ποιότητα και το πλήθος των δεδομένων με τα οποία τροφοδοτείται. Το γεγονός αυτό για συστήματα στον τομέα της άμυνας και της ασφάλειας δημιουργεί προκλήσεις σε θέματα ιδιωτικότητας, εθνικής ασφαλείας αλλά και συμμόρφωσης με κανόνες δικαίου. Τα συστήματα ΤΝ απαιτούν συνεχή επιτήρηση, αλλά και επανεκπαίδευση προκειμένου να είναι αποτελεσματικά. Στο πεδίο της μάχης, οι συγκεκριμένες απαιτήσεις δεν είναι πάντα υλοποιήσιμες.

Η επένδυση στην ΤΝ απαιτεί σημαντικό κόστος τόσο σε επίπεδο υποδομών, καθώς απαιτείται η χρήση υλικού υψηλών επιδόσεων, όσο και σε επίπεδο ενέργειας. Η υιοθέτηση των δυνατοτήτων της ΤΝ σε νέους τομείς, όπως είναι οι εξειδικευμένες χρήσεις στην άμυνα και ασφάλεια απαιτεί σημαντική επένδυση τόσο σε υλικούς όσο και σε ανθρώπινους πόρους. Ένα μεγάλο γλωσσικό μοντέλο που έχει εκπαιδευθεί για γενική χρήση, μπορεί να είναι αναποτελεσματικό σε περιβάλλον ασφαλείας ή να περιλαμβάνει προκαταλήψεις, οι οποίες θα το καταστήσουν ανακριβές ή και επικίνδυνο.

Η παραγωγική ΤΝ προσφέρει καινοτόμες ευκαιρίες αποτροπής και άμυνας, ενώ παράλληλα δημιουργεί τρωτά σημεία ασφαλείας και κινδύνους που σχετίζονται με την ενδεχόμενη χρήση της από τους αντιπάλους.

Κατά συνέπεια ο έλεγχος των δυνατοτήτων που αναπτύσσονται στο πλαίσιο της ΤΝ, η συνεργασία των κρατών με τις εταιρείες ανάπτυξης αλλά και η αποκατάσταση μέτρων εμπιστοσύνης μεταξύ των κρατών είναι επιβεβλημένη.

Τα κράτη θα πρέπει να εντοπίσουν τις δυνατότητες που μπορεί να προσφέρει η ΤΝ στο πλαίσιο προώθησης της εθνικής τους στρατηγικής, εξετάζοντας παράλληλα τις απειλές που δημιουργούνται από την χρήση της ΤΝ από ανταγωνιστικά κράτη. Η υιοθέτηση εφαρμογών της ΤΝ σε περιβάλλοντα ασφαλείας και άμυνας θα πρέπει να διακρίνεται από υπευθυνότητα και συμμόρφωση με κανόνες. Η επένδυση τόσο σε υποδομές όσο και σε δυνατότητες ανάπτυξης λογισμικού είναι απαραίτητη προκειμένου να επιτευχθεί αυτονομία στην ανάπτυξη των συγκεκριμένων δυνατοτήτων. Σημαντικότερη όμως είναι η επένδυση στο ανθρώπινο δυναμικό, τόσο σε αυτό το οποίο εμπλέκεται στον κύκλο ζωής ενός συστήματος ΤΝ όσο και σε αυτό το οποίο χρησιμοποιεί ένα αντίστοιχο σύστημα. Όσες δικλείδες ασφαλείας και μηχανισμοί διαφάνειας και επεξηγησιμότητας να τεθούν σε ένα σύστημα, έγκειται στον τελικό χρήστη να είναι ικανός να χρησιμοποιήσει την προσωπική του κρίση προκειμένου να αποδεχθεί την πρόταση ενός συστήματος ΤΝ. Η υπερβολική εμπιστοσύνη σε έναν αλγόριθμο ή ένα μεγάλο γλωσσικό μοντέλο και η ρουτίνα ενός χειριστή ο οποίος ελέγχει ένα σύστημα ΤΝ σε περιβάλλον ασφαλείας ή άμυνας, μπορεί να έχει συνέπειες που κυμαίνονται από τη μεροληπτική μεταχείριση μέχρι την απώλεια ζωών.

Προκειμένου να αξιοποιηθούν οι δυνατότητες της ΤΝ σε άμυνα και ασφάλεια, οι ένοπλες δυνάμεις και οι δυνάμεις ασφαλείας θα πρέπει να έλθουν σε στενότερη συνεργασία με εξειδικευμένες επιχειρήσεις. Οι κρατικοί φορείς δεν διαθέτουν την τεχνογνωσία, την υπολογιστική ισχύ και συχνά, τα δεδομένα για να διεξάγουν από μόνες τους επιχειρήσεις με τεχνητή νοημοσύνη. Το γεγονός αυτό αυξάνει την ανάγκη θέσπισης πλαισίου λειτουργίας και νομοθεσίας η οποία θα καλύπτει την ανάπτυξη συστημάτων ΤΝ σε όλες τις φάσεις του κύκλου ζωής του. Πρότυπα που αναπτύσσονται από τον ISO, το NIST και άλλους φορείς αποτελούν χρήσιμα εργαλεία, ωστόσο η συμμόρφωση στα πρότυπα δεν είναι πάντα εφικτή ή είναι δύσκολο να ελεγχθεί. Κατά συνέπεια εκτός από τη διαφάνεια των ίδιων των συστημάτων ΤΝ η ανάγκη ανάπτυξης διεθνών μηχανισμών ελέγχου των δυνατοτήτων ΤΝ είναι θεμιτή και θα συμβάλει στη διεθνή σταθερότητα.

Για μικρότερα ή αναπτυσσόμενα κράτη η χρήση της ΤΝ σε συστήματα άμυνας και ασφάλειας αποτελεί πρόκληση αλλά και ταυτόχρονα ευκαιρία. Τέτοιου είδους συστήματα μπορούν να συνεισφέρουν σε οικονομία κλίμακας και ανάπτυξη δυνατοτήτων με μικρό σχετικά κόστος. Η ανάπτυξη δυνατοτήτων ΤΝ μπορεί να αποτελέσει επίσης αντίμετρο σε δυνατότητες ΤΝ ανταγωνιστικών κρατών. Ειδικότερα σε ένα πολύπλοκο περιβάλλον ασφαλείας, συστήματα ΤΝ μπορούν συνεισφέρουν στην κυβερνοασφάλεια και των ανθεκτικότητα κρίσιμων υποδομών. Σε κάθε περίπτωση όμως, όπως καταδεικνύεται και από τις πρώτες χρήσεις της ΤΝ στα πεδία των μαχών, ο καθοριστικός παράγοντας παραμένει ο άνθρωπος, οι αποφάσεις που λαμβάνει αλλά και οι συμβατικές τεχνολογίες, οι οποίες δεν αναμένεται να αντικατασταθούν από την τεχνητή νοημοσύνη.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] M. Skilton και F. Hovsepian, *The 4th industrial revolution: Responding to the impact of artificial intelligence on business*. 2017. doi: 10.1007/978-3-319-62479-2.

[2] Κ. Γεωργούλη, *Τεχνητή Νοημοσύνη*. Αθήνα: ΣΕΑΒ, 2015.

[3] Σ. Κ. Κάτσικας, Σ. Γκρίτζαλης, και Κ. Λαμπρινουδάκης, *Ασφάλεια Πληροφοριών & Συστημάτων στον Κυβερνοχώρο*. Αθήνα: Εκδόσεις Νέων Τεχνολογιών, 2021.

[4] R. E. Neapolitan και X. Jiang, *Τεχνητή Νοημοσύνη*, 2η Έκδοση. Αθήνα: Εκδόσεις Φούντας, 2022.

[5] K. Michael, R. Abbas, και G. Roussos, 'AI in Cybersecurity: The Paradox', *IEEE Trans. Technol. Soc.*, τ. 4, τχ. 2, 2023, doi: 10.1109/tts.2023.3280109.

[6] JRC, 'AI Watch - Defining Artificial Intelligence', 2020.

[7] S. Lockey, N. Gillespie, D. Holm, και I. A. Someh, 'A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions', στο *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2021, τ. 2020-January. doi: 10.24251/hicss.2021.664.

[8] European Commission, 'The-AI-Act', *Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. 2021.

[9] G. Allen, 'Understanding AI Technology', *Jt. Artif. Intell. Cent.*, τχ. April, 2020.

[10] K. M. Sayler, 'Artificial Intelligence and National Security – Economic Impacts and Considerations', *Congr. Res. Serv.*, τ. R45178, τχ. Nov 2020, 2020.

[11] F. Corea, 'AI Knowledge Map: How to Classify AI Technologies', στο *Studies in Big Data*, τ. 50, 2019. doi: 10.1007/978-3-030-04468-8_4.

[12] S. Russell και P. Norvig, *Artificial Intelligence A Modern Approach (4th Edition)*. 2021.

[13] L. Pupillo, S. Fantin, A. Stefano, και C. Polito, 'Artificial Intelligence and Cybersecurity: Technology, Governance and Policy Challenges', Bruxelles, 2021.

[14] A. Bandi, P. V. S. R. Adapa, και Y. E. V. P. K. Kuchi, 'The Power of Generative AI: A Review of Requirements, Models, Input–Output Formats, Evaluation Metrics, and Challenges', *Future Internet*, τ. 15, τχ. 8. 2023. doi: 10.3390/fi15080260.

[15] F. Zhuang *κ.ά.*, 'A Comprehensive Survey on Transfer Learning', *Proceedings of the IEEE*, τ. 109, τχ. 1. 2021. doi: 10.1109/JPROC.2020.3004555.

[16] B. Mahesh, 'Machine Learning Algorithms-A Review', *Int. J. Sci. Res.*, τ. 9, τχ. 1, 2018.

[17] F. . Osisanwo, J. E. . Akinsola, J. . Hinmikaiye, O. Awodele, O. Olakanmi, και J. Akinjobi, 'Supervised Machine Learning Algorithms: Classification and Comparison', *Int. J.*

Comput. Trends Technol., τ. 48, τχ. 3, 2017.

[18] S. Ray, 'A Quick Review of Machine Learning Algorithms', στο *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, 2019. doi: 10.1109/COMITCon.2019.8862451.

[19] Q. Ling, 'Machine learning algorithms review', *Appl. Comput. Eng.*, τ. 4, τχ. 1, σσ 91–98, Ιουνίου 2023, doi: 10.54254/2755-2721/4/20230355.

[20] ENISA, 'Securing Machine Learning Algorithms', Athens, 2021.

[21] D. Mehta, 'State-of-the-Art Reinforcement Learning Algorithms', *Int. J. Eng. Res. Technol.*, τ. 08, τχ. 12, 2019.

[22] D. De Silva και D. Alahakoon, 'An artificial intelligence life cycle: From conception to production', *Patterns*, τ. 3, τχ. 6, 2022, doi: 10.1016/j.patter.2022.100489.

[23] ISO, 'Information technology — Artificial intelligence — AI system life cycle processes', ISO/IEC FDIS 5338, 2023

[24] NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. 2023.

[25] M. Comiter, 'Attacking Artificial Intelligence', Cambridge, MA, 2019.

[26] Threatcasting Lab, 'The New Dogs of War: The Future of Weaponized Artificial Intelligence', Tempe, AZ, 2017.

[27] R. V. Yampolskiy, 'Predicting future AI failures from historic examples', *Foresight*, τ. 21, τχ. 1, 2019, doi: 10.1108/FS-04-2018-0034.

[28] Y. Hu *κ.ά.*, 'Artificial Intelligence Security: Threats and Countermeasures', *ACM Computing Surveys*, τ. 55, τχ. 1. 2021. doi: 10.1145/3487890.

[29] M. Altoub, F. AlQurashi, T. Yigitcanlar, J. M. Corchado, και R. Mehmood, 'An Ontological Knowledge Base of Poisoning Attacks on Deep Neural Networks', *Appl. Sci.*, τ. 12, τχ. 21, 2022, doi: 10.3390/app122111053.

[30] H. Huang, Z. Zhao, M. Backes, Y. Shen, και Y. Zhang, 'Composite Backdoor Attacks Against Large Language Model', 2023. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://arxiv.org/abs/2310.07676

[31] M. Mozes, X. He, B. Kleinberg, και L. D. Griffin, 'Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities', 2023. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://arxiv.org/abs/2308.12833

[32] H. Li *κ.ά.*, 'Multi-step Jailbreaking Privacy Attacks on ChatGPT', 2023. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://arxiv.org/abs/2304.05197

[33] S. Ozdemir, *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Addison-Wesley Professional, 2023.

[34] S. Bhatnagar *κ.ά.*, 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation Authors are listed in order of contribution Design Direction', *arXiv Prepr. arXiv1802.07228*, τχ. February 2018, 2018.

[35] M. M. Yamin, M. Ullah, H. Ullah, και B. Katt, 'Weaponized AI for cyber attacks', *J. Inf. Secur. Appl.*, τ. 57, 2021, doi: 10.1016/j.jisa.2020.102722.

[36] P. V. Falade, 'Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks', *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, τ. 9, τχ. 5, σσ 185–198, 2023.

[37] E. Hubinger *κ.ά.*, 'Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training', 2024. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://arxiv.org/abs/2401.05566

[38] L. Columbus, 'How FraudGPT presages the future of weaponized AI', *VentureBeat*, 2023. https://venturebeat.com/security/how-fraudgpt-presages-the-future-of-weaponized-ai/ (Ημερομηνία πρόσβασης: 20 Ιανουαρίου 2024).

[39] C. da Costa-Luis, *State of Open Source AI*. 2023. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://book.premai.io/state-of-open-source-ai

[40] D. Huynh και J. Hardouin, 'PoisonGPT: How We Hid a Lobotomized LLM on Hugging Face to Spread Fake News', *Mithril Securoty*. https://blog.mithrilsecurity.io/poisongpt-how-we-hid-a-lobotomized-llm-on-hugging-face-to-spread-fake-news/ (Ημερομηνία πρόσβασης: 21 Ιανουαρίου 2024).

[41] T. Stevens, 'Knowledge in the grey zone: AI and cybersecurity', *Digit. War*, τ. 1, τχ. 1–3, 2020, doi: 10.1057/s42984-020-00007-w.

[42] I. Azhar και M. Sr, 'Artificial Intelligence For Cybersecurity: A Systematic Mapping Of Literature', *Int. J. Innov. Eng. Res. Technol. [IJIERT]*, τ. 7, 2020.

[43] G. Varshney και B. Bhushan, 'Cybersecurity Solutions and Communication Technologies for Internet of Things Applications', στο *Artificial Intelligence and Cybersecurity*, 2021. doi: 10.1201/9781003097518-3.

[44] I. H. Sarker, M. H. Furhad, και R. Nowrozy, 'AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions', *SN Computer Science*, τ. 2, τχ. 3. 2021. doi: 10.1007/s42979-021-00557-0.

[45] M. Nachaat, 'Current trends in AI and ML for cybersecurity: A state-of-the-art survey', *Cogent Eng.*, τ. 10, 2023.

[46] S. Zobaed *κ.ά.*, 'DeepFakes: Detecting Forged and Synthetic Media Content Using Machine Learning', στο *Advanced Sciences and Technologies for Security Applications*, 2021. doi: 10.1007/978-3-030-88040-8_7.

[47] M. Taddeo, T. McCutcheon, και L. Floridi, 'Trusting artificial intelligence in cybersecurity is a double-edged sword', *Nat. Mach. Intell.*, τ. 1, τχ. 12, 2019, doi: 10.1038/s42256-019-0109-1.

[48] M. A. Manjramkar και K. C. Jondhale, 'Cyber Security Using Machine Learning Techniques', 2023. doi: 10.2991/978-94-6463-136-4_59.

[49] P. Dixit και S. Silakari, 'Deep Learning Algorithms for Cybersecurity Applications: A Technological and Status Review', *Computer Science Review*, τ. 39. 2021. doi: 10.1016/j.cosrev.2020.100317.

[50] M. Macas, C. Wu, και W. Fuertes, 'A survey on deep learning for cybersecurity: Progress, challenges, and opportunities', *Computer Networks*, τ. 212. 2022. doi: 10.1016/j.comnet.2022.109032.

[51] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, και M. Marchetti, 'On the effectiveness of machine and deep learning for cyber security', στο *International Conference on Cyber Conflict, CYCON*, 2018, τ. 2018-May. doi: 10.23919/CYCON.2018.8405026.

[52] G. Abdiyeva-Aliyeva, J. Aliyev, και U. Sadigov, 'Application of classification algorithms of Machine learning in cybersecurity', στο *Procedia Computer Science*, 2022, τ. 215. doi: 10.1016/j.procs.2022.12.093.

[53] M. F. Ansari, B. Dash, P. Sharma, και N. Yathiraju, 'The Impact and Limitations of Artificial Intelligence in Cybersecurity: A Literature Review', *IJARCCE*, τ. 11, τχ. 9, 2022, doi: 10.17148/ijarcce.2022.11912.

[54] L. Lazic, 'Benefit From AI in Cybersecurity', *11th Int. Conf. Bus. Inf. Secur. Belgrade, Serbia*, τχ. October, 2019.

[55] R. Csernatoni, 'Weaponizing Innovation? Mapping Artificial Intelligence-Enabled Security and Defence in the EU', 2023.

[56] E. Schmidt *κ.ά.*, 'Final report of the National Security Commission on Artificial Intelligence', 2021.

[57] EPRS, 'Artificial Intelligence at EU borders - Overview of applications and key issues', 2021.

[58] C.-H. Huang, T.-C. Chou, και S.-H. Wu, 'Towards Convergence of AI and IoT for Smart Policing', *J. Glob. Inf. Manag.*, τ. 29, τχ. 6, 2022, doi: 10.4018/jgim.296260.

[59] M. Afzal και P. Panagiotopoulos, 'Smart Policing: A Critical Review of the Literature', στο *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020, τ. 12219 LNCS. doi: 10.1007/978-3-030-57599-1_5.

[60] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, και J. E. Gilbert, 'A

review of predictive policing from the perspective of fairness', *Artificial Intelligence and Law*, τ. 30, τχ. 1. 2022. doi: 10.1007/s10506-021-09286-4.

[61] S. Sieveneck και C. Sutter, 'Predictive policing in the context of road traffic safety: A systematic review and theoretical considerations', *Transportation Research Interdisciplinary Perspectives*, τ. 11. 2021. doi: 10.1016/j.trip.2021.100429.

[62] P. Cochrane και M. P. Pfeiffer, 'Patterns in Policing', *Adv. Sci. Technol. Secur. Appl.*, 2020, doi: 10.1007/978-3-030-50613-1_10.

[63] D. Kavallieros, D. Myttas, E. Kermitsis, E. Lissaris, G. Giataganas, και E. Darra, 'Understanding the Dark Web', 2021. doi: 10.1007/978-3-030-55343-2_1.

[64] B. Staley και R. Montasari, 'A Survey of Challenges Posed by the DarkWeb', στο *Artificial Intelligence in Cyber Security: Impact and Implications*, Springer, σσ 203–214.

[65] K. Foy, 'Artificial intelligence is helping investigators fight crime on the dark web', *MIT Lincoln Laboratory*, 2019. https://www.ll.mit.edu/news/artificial-intelligence-helping-investigators-fight-crime-dark-web (Ημερομηνία πρόσβασης: 2 Φεβρουαρίου 2024).

[66] R. Thornton και M. Miron, 'Towards the 'Third Revolution in Military Affairs'', *RUSI J.*, τ. 165, τχ. 3, 2020, doi: 10.1080/03071847.2020.1765514.

[67] G. Yan, 'The impact of Artificial Intelligence on hybrid warfare', *Small Wars Insur.*, τ. 31, τχ. 4, 2020, doi: 10.1080/09592318.2019.1682908.

[68] B. Cartwright, R. Frank, G. Weir, και K. Padda, 'Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks', *Neural Comput. Appl.*, τ. 34, σσ 15141–15163, 2022.

[69] Zachary Davis, 'Artificial Intelligence on the Battlefield: Implications for Deterrence and Surprise', *Prism (Washington, D.C.)*, τ. 8, τχ. 2, 2019.

[70] DAIC, 'The Defence AI Playbook', 2024.

[71] S. Duan *κ.ά.*, 'Distributed Artificial Intelligence Empowered by End-Edge-Cloud Computing: A Survey', *IEEE Commun. Surv. Tutorials*, τ. 25, τχ. 1, 2023, doi: 10.1109/COMST.2022.3218527.

[72] D. Papakostas, T. Kasidakis, E. Fragkou, και D. Katsaros, 'Backbones for Internet of Battlefield Things', στο *16th Conference on Wireless On-Demand Network Systems and Services, WONS 2021*, 2021. doi: 10.23919/WONS51326.2021.9415560.

[73] L. Zhu, S. Majumdar, και C. Ekenna, 'An invisible warfare with the internet of battlefield things: A literature review', *Human Behavior and Emerging Technologies*, τ. 3, τχ. 2. 2021. doi: 10.1002/hbe2.231.

[74] S. E. Kase, C. P. Hung, T. Krayzman, J. Z. Hare, B. C. Rinderspacher, και S. M. Su, 'The Future of Collaborative Human-Artificial Intelligence Decision-Making for Mission

Planning', *Front. Psychol.*, τ. 13, 2022, doi: 10.3389/fpsyg.2022.850628.

[75] D. C. Tarraf *κ.ά.*, 'An experiment in tactical wargaming with platforms enabled by artificial intelligence', *J. Def. Model. Simul.*, 2022, doi: 10.1177/15485129221097103.

[76] P. K. Davis και P. Bracken, 'Artificial intelligence for wargaming and modeling', *J. Def. Model. Simul.*, 2022, doi: 10.1177/15485129211073126.

[77] T. F. Blauth, O. J. Gstrein, και A. Zwitter, 'Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI', *IEEE Access*, τ. 10, 2022, doi: 10.1109/ACCESS.2022.3191790.

[78] Lockheed Martin, 'Cyber Kill Chain® | Lockheed Martin', *Lockheed Martin*. 2023. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[79] T. Yadav και A. M. Rao, 'Technical aspects of cyber kill chain', στο *Communications in Computer and Information Science*, 2015, τ. 536. doi: 10.1007/978-3-319-22915-7_40.

[80] N. Kaloudi και L. I. Jingyue, 'The AI-based cyber threat landscape: A survey', *ACM Computing Surveys*, τ. 53, τχ. 1. 2020. doi: 10.1145/3372823.

[81] J. Sánchez-Monedero και L. Dencik, 'The politics of deceptive borders: 'biomarkers of deceit' and the case of iBorderCtrl', *Inf. Commun. Soc.*, τ. 25, τχ. 3, 2022, doi: 10.1080/1369118X.2020.1792530.

[82] A. Andreou, 'e-Securing the EU Borders: AI in European Integrated Border Management', *J. Polit. Ethics New Technol. AI*, τ. 2, τχ. 1, 2023, doi: 10.12681/jpentai.34287.

[83] Z. Campbell, 'Swarms of Drones, Piloted by Artificial Intelligence, May Soon Patrol Europe's Borders', *The Intercept*, 2019. https://theintercept.com/2019/05/11/drones-artificial-intelligence-europe-roborder/ (Ημερομηνία πρόσβασης: 21 Ιανουαρίου 2024).

[84] H. Nasu, 'The Kargu-2 Autonomous Attack Drone: Legal & Ethical Dimensions', *Articles of War*, 2021. https://lieber.westpoint.edu/kargu-2-autonomous-attack-drone-legal-ethical/ (Ημερομηνία πρόσβασης: 16 Ιανουαρίου 2024).

[85] A. Dangwal, '1st Kill By 'Artificial Intelligence'! Russia Confirms Its S-350 Vityaz System 'Shot Down' Ukrainian Aircraft In Auto Mode', *Eurasian Times*, 2023. https://www.eurasiantimes.com/1st-kill-by-artificial-intelligence-russia-says-its-s-350-vityaz-system/ (Ημερομηνία πρόσβασης: 15 Ιανουαρίου 2024).

[86] D. Hambling, 'Drones killing without oversight?', *New Scientist*, σ 8, Οκτωβρίου 2023.

[87] Reuters, 'Israel's Shin Bet spy service uses generative AI to thwart threats', 2023. https://www.reuters.com/technology/israels-shin-bet-spy-service-uses-generative-ai-thwart-threats-2023-06-27/ (Ημερομηνία πρόσβασης: 16 Ιανουαρίου 2024).

[88] The Soufan Center, 'IntelBrief: AI-Powered Disinformation in the Israel-Hamas War and Beyond', 2023. https://thesoufancenter.org/intelbrief-2023-october-26/ (Ημερομηνία πρόσβασης 16 Ιανουαρίου 2024).

[89] H. Davies, B. McKernan, και D. Sabbagh, ''The Gospel': how Israel uses AI to select bombing targets in Gaza', *The Guardian*. https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets (Ημερομηνία πρόσβασης: 16 Ιανουαρίου 2024).

[90] A. Ahronheim, 'Israel's operation against Hamas was the world's first AI war', *The Jerusalem Post*, 2021. https://www.jpost.com/arab-israeli-conflict/gaza-news/guardian-of-the-walls-the-first-ai-war-669371 (Ημερομηνία πρόσβασης: 16 Ιανουαρίου 2024).

[91] M. Krishnan, 'Indian army ramps up AI, but how effective will it be?', *Deutsche Welle*, 2023. https://www.dw.com/en/indian-army-ramps-up-ai-but-how-effective-will-it-be (Ημερομηνία πρόσβασης: 16 Ιανουαρίου 2024).

[92] P. Svenmarck, L. Luotsinen, M. Nilsson, και J. Schubert, 'Possibilities and Challenges for Artificial Intelligence in Military Applications', *Proc. NATO Big Data Artif. Intell. Mil. Decis. Mak. Spec. Meet.*, 2018.

[93] P. Feldman, A. Dant, και A. Massey, 'Integrating Artificial Intelligence into Weapon Systems', 2019.

[94] Σ. Γκρίτζαλης, 'Αλγοριθμικές Προκαταλήψεις', *Το Βήμα*, Αθήνα, 2024.

[95] L. Floridi *κ.ά.*, 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', *Minds Mach.*, τ. 28, τχ. 4, 2018, doi: 10.1007/s11023-018-9482-5.

[96] M. Taddeo, 'Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity', *Minds and Machines*, τ. 29, τχ. 2. 2019. doi: 10.1007/s11023-019-09504-8.

[97] K. E. Busch, 'Generative Artificial Intelligence and Data Privacy: A Primer', Washington DC, 2023.

[98] The White House, 'FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence', 2023. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/ (Ημερομηνία πρόσβασης: 16 Ιανουαρίου 2024).

[99] Λ. Μήτρου, 'Το Κανονιστικό Πλαίσιο της (Κυβερνο)Ασφάλειας', στο *Ασφάλεια Πληροφοριών & Συστημάτων στον Κυβερνοχώρο*, Σ. Γκρίτζαλης, Σ. Κάτσικας, και Κ. Λαμπρινουδάκης, Επιμ. Αθήνα: Εκδόσεις Νέων Τεχνολογιών, 2021, σ 812.

[100] The European Parliament, 'Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI', 2023. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai (Ημερομηνία πρόσβασης: 16 Ιανουαρίου 2024).

[101] AI HLEG, 'Ethics guidelines for trustworthy AI', 2019.

[102] AI HLEG, 'Assessment List for Trustworthy AI (ALTAI)', 2020.

[103] C. Radclyffe, M. Ribeiro, και R. H. Wortham, 'The assessment list for trustworthy artificial intelligence: A review and recommendations', *Front. Artif. Intell.*, τ. 6, 2023, doi: 10.3389/frai.2023.1020592.

[104] Google, 'Secure AI Framework', 2023.

[105] NATO, 'Summary of the NATO Artificial Intelligence Strategy', 2021. https://www.nato.int/cps/en/natohq/official_texts_187617.htm (Ημερομηνία πρόσβασης: 10 Μαρτίου 2024).

[106] P. Bezombes, S. Brunessaux, και S. Cadzow, 'Cybersecurity of AI and Standarisation', Athens, 2023.

[107] D. Araya και M. King, 'The Impact of Artificial Intelligence on Military Defence and Security', Waterloo, Canada, 2022.

[108] U. S. Gaire, 'Application of Artificial Intelligence in the Military: An Overview', *Unity J.*, τ. 4, τχ. 01, 2023, doi: 10.3126/unityj.v4i01.52237.

[109] A. Goldfarb και J. R. Lindsay, 'Prediction and Judgment Prediction and Judgment: Why Artiacial Intelligence Increases the Importance of Humans in War', *Int. Secur.*, τ. 46, τχ. 3, 2022, doi: 10.1162/isec_a_00425.

[110] C. Hunter και B. E. Bowen, 'We'll never have a model of an AI major-general: Artificial Intelligence, command decisions, and kitsch visions of war', *J. Strateg. Stud.*, 2023, doi: 10.1080/01402390.2023.2241648.

[111] B. Buchanan, 'A National Security Research Agenda for Cybersecurity and Artificial Intelligence CSET Issue Brief', *Cent. Secur. Emerg. Technol.*, τ. May, τχ. May, 2020.

[112] P. Davies, 'This AI malware worm can steal private data and send spam emails without you ever having to click', *Euronews*, 2024. https://www.euronews.com/next/2024/03/07/this-ai-worm-can-steal-private-data-and-send-spam-emails (Ημερομηνία πρόσβασης: 29 Μαρτίου 2024).

[113] F. Ding, G. Zhu, Y. Li, X. Zhang, P. K. Atrey, και S. Lyu, 'Anti-Forensics for Face Swapping Videos via Adversarial Training', *IEEE Trans. Multimed.*, τ. 24, 2022, doi: 10.1109/TMM.2021.3098422.

[114] Σ. Κάτσικας, 'Ασφάλεια Κυβερνο-Φυσικών Συστημάτων', στο *Ασφάλεια Πληροφοριών*

& Συστημάτων στον Κυβερνοχώρο, Σ. Γκρίτζαλης, Σ. Κάτσικας, και Κ. Λαμπρινουδάκης, Επιμ. Αθήνα: Εκδόσεις Νέων Τεχνολογιών, 2021, σ 812.

[115] STO, 'Science & Technology Trends 2023-2043', Brussels, 2023.

[116] T. Paterson και L. Hanley, 'Political warfare in the digital age: cyber subversion, information operations and 'deep fakes'', *Aust. J. Int. Aff.*, τ. 74, τχ. 4, 2020, doi: 10.1080/10357718.2020.1734772.

[117] P. Timmers, 'Ethics of AI and Cybersecurity When Sovereignty is at Stake', *Minds and Machines*, τ. 29, τχ. 4. 2019. doi: 10.1007/s11023-019-09508-4.

[118] A. N. Guiora, *Κυβερνοασφάλεια: Γεωπολιτική, Δίκαιο και Στρατηγική*. Αθήνα: Πεδίο, 2023.

[119] J. Burton και S. R. Soare, 'Understanding the Strategic Implications of the Weaponization of Artificial Intelligence', στο *International Conference on Cyber Conflict, CYCON*, 2019, τ. 2019-May. doi: 10.23919/CYCON.2019.8756866.

[120] M. O'Connor, 'Replicator: A Bold New Path for DoD', *Center for Security and Emerging Technology (CSET)*, 2023. https://cset.georgetown.edu/article/replicator-a-bold-new-path-for-dod/ (Ημερομηνία πρόσβασης: 16 Ιανουαρίου 2024).

[121] D. Garcia, *The AI Military Race*. 2023. doi: 10.1093/oso/9780192864604.001.0001.

[122] H. A. Kissinger και G. Allison, 'The Path to AI Arms Control', *Foreign Affairs*, 2023. [Έκδοση σε ψηφιακή μορφή]. Διαθέσιμο στο: https://www.foreignaffairs.com/issues/2023/102/6#

[123] M. Mittelsteadt, 'AI Verification: Mechanisms to Ensure AI Arms Control Compliance', Washington DC, 2021.

[124] J. Cohen και G. Lee, 'The generative world order: AI, geopolitics, and power', *Goldman Sachs*, 2023. https://www.goldmansachs.com/intelligence/pages/the-generative-world-order-ai-geopolitics-and-power.html (Ημερομηνία πρόσβασης: 15 Φεβρουαρίου 2024).

[125] A. Nadibaidze και N. Miotto, 'The Impact of AI on Strategic Stability is What States Make of It: Comparing US and Russian Discourses', *J. Peace Nucl. Disarm.*, τ. 6, τχ. 1, 2023, doi: 10.1080/25751654.2023.2205552.