ΔΙΕΘΝΕΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Κυβερνοεπιθέσεις στηριζόμενες στην Τεχνητή Νοημοσύνη: Συστηματική βιβλιογραφική Έρευνα»

Της φοιτήτριας
Παρασκευοπούλου Ιωάννας
Αρ. Μητρώου: 28/2024

Επιβλέπων
Ηλιούδης Χρήστος
Βαθμίδα: Καθηγητής

Φεβρουάριος 2026

Τίτλος Δ.Ε. Κυβερνοεπιθέσεις στηριζόμενες στην Τεχνητή Νοημοσύνη: Συστηματική βιβλιογραφική Έρευνα

Κωδικός Δ.Ε. 25300

Ονοματεπώνυμο φοιτητή Παρασκευοπούλου Ιωάννα

Ονοματεπώνυμο εισηγητή Ηλιούδης Χρήστος

Ημερομηνία ανάληψης Δ.Ε. 19/09/2025

Ημερομηνία περάτωσης Δ.Ε. 06/02/2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Στα παιδιά μου, Παναγιώτη και Θεόδωρο

Πρόλογος

Η παρούσα διπλωματική εργασία, με τίτλο «Κυβερνοεπιθέσεις στηριζόμενες στην Τεχνητή Νοημοσύνη: Συστηματική βιβλιογραφική Έρευνα», εστιάζει στη μελέτη των επιθετικών εφαρμογών της Τεχνητής Νοημοσύνης (ΤΝ) στον τομέα της κυβερνοασφάλειας, χωρίς αναφορά σε αμυντικές πρακτικές ή μηχανισμούς προστασίας. Στόχος της εργασίας είναι η κατανόηση των δυνατοτήτων, των βασικών χαρακτηριστικών και των εξελισσόμενων τάσεων που σχετίζονται με την αυτοματοποίηση και την ενίσχυση κυβερνοεπιθέσεων μέσω τεχνολογιών ΤΝ.

Το ερευνητικό ενδιαφέρον για το συγκεκριμένο αντικείμενο απορρέει από τη ραγδαία εξέλιξη τεχνολογιών Τεχνητής Νοημοσύνης, όπως τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs) και τα Γενετικά Ανταγωνιστικά Δίκτυα (Generative Adversarial Networks – GANs), οι οποίες έχουν διευρύνει σημαντικά το φάσμα, την κλίμακα και την πολυπλοκότητα των σύγχρονων κυβερνοεπιθέσεων. Οι τεχνολογίες αυτές επιτρέπουν την αυτοματοποίηση, την προσαρμοστικότητα και την αυξημένη αποτελεσματικότητα επιθετικών πρακτικών, καθιστώντας αναγκαία τη συστηματική καταγραφή, κατηγοριοποίηση και ανάλυση των σχετικών τακτικών. Παράλληλα, η απουσία μίας συγκεντρωτικής και μεθοδικά οργανωμένης ανασκόπησης της διεθνούς βιβλιογραφίας που να χαρτογραφεί τις επιθέσεις αυτές ανά είδος, χαρακτηριστικά και επιπτώσεις αναδεικνύει ένα σαφές ερευνητικό κενό, το οποίο επιχειρεί να καλύψει η παρούσα διπλωματική εργασία.

Η εκπόνηση της παρούσας μελέτης αναμένεται να συμβάλει ουσιαστικά στην ακαδημαϊκή και ερευνητική κοινότητα, προσφέροντας μια συστηματική αποτύπωση των υφιστάμενων επιθετικών τεχνικών, την ανάδειξη κρίσιμων ερευνητικών κενών και τον εντοπισμό κατευθύνσεων για μελλοντική επιστημονική εμβάθυνση. Η προσέγγιση της εργασίας δίνει έμφαση στην καταγραφή, ομαδοποίηση και συστηματοποίηση των τακτικών κυβερνοεπιθέσεων που αξιοποιούν ΤΝ, χωρίς να επεκτείνεται σε ζητήματα αντιμετώπισης ή αμυντικών στρατηγικών.

Περίληψη

Ο στόχος της παρούσας εργασίας ήταν η συστηματική διερεύνηση των μορφών κυβερνοεπιθέσεων που αξιοποιούν τεχνητή νοημοσύνη, με έμφαση στις τεχνικές, κοινωνικές, οικονομικές και θεσμικές τους επιπτώσεις. Η μεθοδολογική προσέγγιση βασίστηκε στο πρωτόκολλο PRISMA, το οποίο επέτρεψε την οργανωμένη συλλογή, αξιολόγηση και ανάλυση της σχετικής βιβλιογραφίας. Μέσα από τη διαδικασία επιλογής και φιλτραρίσματος πηγών, εξασφαλίστηκε η εγκυρότητα και η διαφάνεια της ανασκόπησης, ενώ παράλληλα αναδείχθηκαν τα κενά που παραμένουν στη διεθνή έρευνα.

Τα ευρήματα καταδεικνύουν ότι η ΤΝ λειτουργεί ως πολλαπλασιαστής ισχύος για τους κυβερνοεγκληματίες, μετασχηματίζοντας παραδοσιακές πρακτικές σε εξελιγμένες και δύσκολα ανιχνεύσιμες επιθέσεις. Συγκεκριμένα, εντοπίστηκαν: α) Τεχνικές επιθέσεις όπως phishing, παραβίαση CAPTCHA, zero-day exploits, adaptive malware και adversarial attacks. β) Κοινωνικές και ψυχολογικές επιθέσεις μέσω AI-driven social engineering και ανίχνευσης ανθρώπινων αδυναμιών. γ) Θεσμικές και οικονομικές απειλές όπως deepfakes, παραποίηση νομικών και οικονομικών εγγράφων, καθώς και επιθέσεις σε IoT και αυτόνομα συστήματα με φυσικές συνέπειες.

Συνολικά, η ανασκόπηση ανέδειξε ότι οι επιθέσεις με ΤΝ έχουν πολυδιάστατες επιπτώσεις που επηρεάζουν την κυβερνοασφάλεια, την οικονομία, την κοινωνική εμπιστοσύνη και τη θεσμική διαφάνεια. Η βιβλιογραφία συγκλίνει στην ανάγκη για ολιστική στρατηγική αντιμετώπισης, η οποία θα συνδυάζει τεχνικές λύσεις, θεσμικά μέτρα και κοινωνική ευαισθητοποίηση, ενώ παράλληλα υπογραμμίζεται η σημασία περαιτέρω διεπιστημονικής και εμπειρικής έρευνας.

Λέξεις κλειδιά: Τεχνητή Νοημοσύνη (ΤΝ), Κυβερνοεπιθέσεις, Phishing / Zero-day exploits,

Adaptive malware / Adversarial attacks, Deepfakes / Παραποίηση εγγράφων, Κυβερνοασφάλεια & κοινωνικές επιπτώσεις.

«Cyberattacks based on Artificial Intelligence: A Systematic Literature Review»

«Ioanna Paraskevopoulou»

Abstract

The aim of this study was to systematically investigate the forms of cyberattacks that leverage artificial intelligence (AI), focusing on their technical, social, economic, and institutional impacts. The methodological approach was based on the PRISMA protocol, which enabled the structured collection, evaluation, and synthesis of relevant literature. Through the systematic selection and filtering of sources, the review ensured transparency and validity, while also highlighting existing research gaps.

Findings indicate that AI acts as a force multiplier for cybercriminals, transforming traditional attack practices into advanced, automated, and difficult-to-detect threats. Specifically, the review identified: a) Technical attacks such as phishing, CAPTCHA bypass, zero-day exploits, adaptive malware, and adversarial attacks. b) Social and psychological attacks through AI-driven social engineering and the exploitation of human vulnerabilities. c) Institutional and economic threats including deepfakes, forgery of legal and financial documents, and attacks on IoT and autonomous systems with potential physical consequences.

Overall, the review demonstrates that AI-enabled attacks have multidimensional impacts, affecting cybersecurity, economic stability, social trust, and institutional transparency. The literature converges on the need for a holistic defense strategy, combining technical solutions, regulatory measures, and social awareness, while also emphasizing the importance of further interdisciplinary and empirical research.

Keywords: Artificial Intelligence (AI), Cyberattacks, Phishing / Zero-day exploits,

Adaptive malware / Adversarial attacks, Deepfakes / Document forgery, Cybersecurity & social impacts.

Ευχαριστίες

Η παρούσα διπλωματική εργασία αφιερώνεται στα παιδιά μου και στον σύζυγό μου, ως ένα μικρό αντίδωρο για τον πολύτιμο χρόνο που τους στέρησα κατά τη διάρκεια της εκπόνησής της. Τους ευχαριστώ θερμά για την αμέριστη υπομονή, τη συνεχή στήριξη και τη δύναμη που μου προσέφεραν καθημερινά, συμβάλλοντας καθοριστικά στην ολοκλήρωση αυτής της προσπάθειας.

Θα ήθελα, επίσης, να εκφράσω τις ειλικρινείς μου ευχαριστίες προς τον επιβλέποντα καθηγητή μου, κ. Χρήστο Ηλιούδη, για την πολύτιμη καθοδήγηση, τις ουσιαστικές παρατηρήσεις και τη στήριξή του, τόσο κατά την επιλογή του θέματος όσο και καθ' όλη τη διάρκεια της υλοποίησης της παρούσας εργασίας.

Περιεχόμενα

Πρόλογος iv

Περίληψη v

Abstract vi

Ευχαριστίες vii

Περιεχόμενα viii

Κατάλογος Πινάκων xi

Κεφάλαιο 1ο: Εισαγωγή 1

1.1 Εισαγωγή 1

1.2 Τεχνητή νοημοσύνη: Ορισμός, Ιστορική Εξέλιξη και Σύγχρονες Εφαρμογές 1

1.3 Η Τεχνητή Νοημοσύνη ως Εργαλείο Διεξαγωγής Κυβερνοεπιθέσεων: Δυνατότητες και Κίνδυνοι 3

1.4 Η Συμβολή της Τεχνητής Νοημοσύνης στην Αντιμετώπιση Κυβερνοεπιθέσεων 4

1.5 Επιπτώσεις των Κυβερνοεπιθέσεων με Χρήση Τεχνητής Νοημοσύνης 5

1.6 Σκοπός της έρευνας 7

1.7 Ερευνητικά ερωτήματα 7

1.8 Επίλογος 7

Κεφάλαιο 2ο: Μεθοδολογία 9

2.1 Εισαγωγή 9

2.2 Μεθοδολογική προσέγγιση 9

2.3 Μορφές Επιθέσεων με Χρήση Τεχνητής Νοημοσύνης 11

2.4 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Phishing με Χρήση Τεχνητής Νοημοσύνης» 12

2.5 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Παραβίαση CAPTCHA και Συστημάτων Επαλήθευσης» 13

2.6 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Ανάλυση Επιθέσεων Zero-Day με Χρήση ΤΝ» 14

2.7 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις, με τη χρήση ΤΝ, στην ανάπτυξη προσαρμοστικού κακόβουλου λογισμικού» 15

2.8 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις Deepfakes και Παραπληροφόρησης με τη χρήση ΤΝ» 15

2.9 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις με εξαγωγή δεδομένων μέσω AI-driven social engineering με χρήση της ΤΝ» 16

2.10 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις με ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών με τη χρήση ΤΝ» 17

2.11 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις Adversarial attacks με τη χρήση ΤΝ» 18
2.12 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις σε αυτόνομα συστήματα και IoT με τη χρηση ΤΝ» 19
2.13 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις με τη χρήση ΤΝ για παραποίηση νομικών και οικονομικών εγγράφων» 20
2.14 Μεθοδολογία PRISMA για τις επιπτώσεις από τις επιθέσεις με ΤΝ 21
2.15 Επίλογος 21
Κεφάλαιο 3ο: Επιστημονικά ευρήματα - Κατηγορίες επιθετικών τεχνικών AI 23
3.1 Εισαγωγή 23
3.2 Αυτοματοποιημένες Επιθέσεις Phishing μέσω Τεχνητής Νοημοσύνης: Μηχανισμοί, Εξελίξεις και Επιπτώσεις 23
3.3 Παραβίαση CAPTCHA και Συστημάτων Επαλήθευσης μέσω Τεχνητής Νοημοσύνης 28
3.4 Η Χρήση της Τεχνητής Νοημοσύνης για την Ενίσχυση Επιθέσεων Τύπου Zero-Day 32
3.5 Δημιουργία προσαρμοστικού κακόβουλου λογισμικού με τη χρήση τεχνητής νοημοσύνης 34
3.6 Deepfakes και Παραπληροφόρηση με χρήση της Τεχνητής Νοημοσύνης 36
3.7 Εξαγωγή δεδομένων μέσω AI-driven social engineering με χρήση της Τεχνητής Νοημοσύνης 38
3.8 Ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών 40
3.9 Adversarial attacks σε συστήματα ΤΝ 43
3.10 Επιθέσεις σε αυτόνομα συστήματα και IoT 45
3.11 Παραποίηση νομικών και οικονομικών εγγράφων 47
3.12 Επιπτώσεις από Μορφές Επιθέσεων με Χρήση ΤΝ 49
3.13 Επιπτώσεις ανά κατηγορία 51
3.14 Επίλογος 54
Κεφάλαιο 4ο: Σύνοψη ευρημάτων 55
4.1 Εισαγωγή 55
4.2 Ευρήματα 55
4.2.1 Αυτοματοποιημένες Επιθέσεις Phishing μέσω Τεχνητής Νοημοσύνης 56
4.2.2 Παραβίαση CAPTCHA και Συστημάτων Επαλήθευσης μέσω Τεχνητής Νοημοσύνης 56
4.2.3 Η Χρήση της Τεχνητής Νοημοσύνης για την Ενίσχυση Επιθέσεων Τύπου Zero-Day . 57
4.2.4 Δημιουργία προσαρμοστικού κακόβουλου λογισμικού με τη χρήση τεχνητής νοημοσύνης 57
4.2.5 Deepfakes και Παραπληροφόρηση με χρήση της Τεχνητής Νοημοσύνης 58
4.2.6 Εξαγωγή δεδομένων μέσω AI-driven social engineering με χρήση της Τεχνητής Νοημοσύνης 58

4.2.7 Ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών 59
4.2.8 Adversarial attacks σε συστήματα TN 59
4.2.9 Επιθέσεις σε αυτόνομα συστήματα και IoT 60
4.2.10 Παραποίηση νομικών και οικονομικών εγγράφων 60
4.3 Επιπτώσεις από Μορφές Επιθέσεων με Χρήση TN 61
4.4 Επίλογος 62
Κεφάλαιο 5ο: Συζήτηση - συμπεράσματα 63
5.1 Εισαγωγή 63
5.2 Συμπεράσματα 63
5.3 Απαντήσεις στα ερευνητικά ερωτήματα 64
5.3.1 Ποιες είναι οι κύριες τακτικές και τεχνικές κυβερνοεπιθέσεων που αξιοποιούν τεχνητή νοημοσύνη; 64
5.3.2 Ποιοι τύποι επιθέσεων παρατηρούνται και ποια είναι τα βασικά χαρακτηριστικά τους; 64
5.3.3 Ποιες είναι οι επιπτώσεις αυτών των επιθέσεων και σε ποιους τομείς εκδηλώνονται; 65
5.3.4 Ποια είναι τα ιδιαίτερα χαρακτηριστικά που πρέπει να έχουν τα μέτρα προστασίας και άμυνας έναντι των επιθέσεων που βασίζονται στην τεχνητή νοημοσύνη; 66
5.4 Συζήτηση 66
5.5 Ερευνητικά κενά 67
5.6 Προτάσεις για μελλοντική έρευνα 68
5.7 Επίλογος 69
ΒΙΒΛΙΟΓΡΑΦΙΑ 71

Κατάλογος Πινάκων

Πίνακας 2.1: Διάγραμμα Ροής PRISMA για τη διάσταση διάσταση «Phishing με χρήση ΤΝ» 12

Πίνακας 2.2: Διάγραμμα Ροής PRISMA για τη διάσταση «Παραβίαση CAPTCHA και Συστημάτων Επαλήθευσης μέσω ΤΝ» 13

Πίνακας 2.3: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις Zero-Day με Χρήση ΤΝ»..... 14

Πίνακας 2.4: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις με τη χρήση ΤΝ, στην ανάπτυξη προσαρμοστικού κακόβουλου λογισμικού»..... 15

Πίνακας 2.5: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις Deepfakes και Παραπληροφόρησης με τη χρήση ΤΝ»..... 16

Πίνακας 2.6: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις με εξαγωγή δεδομένων μέσω AI-driven social engineering με χρήση της ΤΝ»..... 17

Πίνακας 2.7: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις με ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών με τη χρήση ΤΝ»..... 18

Πίνακας 2.8: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις Adversarial attacks με τη χρήση ΤΝ»..... 19

Πίνακας 2.9: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις σε αυτόνομα συστήματα και IoT με τη χρηση ΤΝ»..... 19

Πίνακας 2.10: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις με τη χρήση ΤΝ για παραποίηση νομικών και οικονομικών εγγράφων»..... 20

Πίνακας 2.11: Διάγραμμα Ροής PRISMA για τις επιπτώσεις από τις επιθέσεις με ΤΝ 21

Πίνακας 3.1: Μηχανισμοί Ενίσχυσης Phishing μέσω Τεχνητής Νοημοσύνης..... 26

Πίνακας 3.2: Χρήση ΤΝ σε Phishing Επιθέσεις..... 27

Πίνακας 3.3: Τύποι CAPTCHA και Χαρακτηριστικά..... 30

Πίνακας 3.4: Ποσοστά Παραβίασης CAPTCHA από Τεχνητή Νοημοσύνη..... 31

Πίνακας 3.5: Εργαλεία ΤΝ για Επιθέσεις Zero-Day..... 33

Πίνακας 3.6: Εργαλεία ΤΝ για επιθέσεις με προσαρμοστικό κακόβουλο λογισμικό..... 35

Πίνακας 3.7: Μορφές Επιθετικής Χρήσης Deepfakes..... 37

Πίνακας 3.8: Ερευνητικά ευρήματα..... 38

Πίνακας 3.9: Ευρήματα για επιθέσεις AI-driven Social Engineering με τη χρήση ΤΝ..... 39

Πίνακας 3.10: Ευρήματα μελετών για την χρήση ΤΝ στην ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών..... 41

Πίνακας 3.11: Ευρήματα μελετών για τις επιθέσεις Adversarial attacks 44
Πίνακας 3.12: Ευρήματα μελετών για τις επιθέσεις σε αυτόνομα συστήματα και IoT..... 46
Πίνακας 3.13: Ερευνητικά ευρήματα για παραποίηση Εγγράφων με ΤΝ 48
Πίνακας 3.14: Επιπτώσεις από Μορφές Επιθέσεων με Χρήση ΤΝ..... 50

Συντομογραφίες

AI Τεχνητή Νοημοσύνη

LLM Large Language Model (Μεγάλο Γλωσσικό Μοντέλο)

GAN **Generative Adversarial Network (Αντιθετικό Γενετικό Δίκτυο)**

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

DDoS Distributed Denial of Service (Κατανεμημένη Άρνηση Υπηρεσίας)

DoS Denial of Service (Άρνηση Υπηρεσίας)

APT Advanced Persistent Threat

CVE Common Vulnerabilities and Exposures

IoT Internet of Things (Διαδίκτυο των Πραγμάτων)

OSINT Open Source Intelligence

IDS / IPS Intrusion Detection (and Prevention) System

SIEM Security Information and Event Management

CAPEC Common Attack Pattern Enumeration and Classification

ATT&CK Adversarial Tactics, Techniques, and Common Knowledge (μορφής MITRE)

RAT Remote Access Trojan

SQLi SQL Injection

MITM Man-in-the-Middle

CIA Confidentiality, Integrity, Availability

TLS / HTTPS Transport Layer Security / HTTP Secure

MFA / 2FA Multi-Factor / Two-Factor Authentication

EDR Endpoint Detection and Response (Παρακολούθηση & Απόκριση Τερματικών)

DLP Data Loss Prevention

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή

Η τεχνητή νοημοσύνη (ΤΝ) ορίζεται ως η ικανότητα ενός συστήματος να εκτελεί λειτουργίες που παραδοσιακά απαιτούν ανθρώπινη νοημοσύνη, όπως μάθηση, λήψη αποφάσεων και κατανόηση της γλώσσας (Russell & Norvig, 2021). Αν και η ιδέα της μηχανικής ευφυΐας έχει τις ρίζες της σε μύθους και τεχνολογικά πειράματα από την αρχαιότητα έως την Αναγέννηση (Copeland, 2025), η επιστημονική της θεμελίωση ξεκίνησε τον 20ό αιώνα με το έργο του Turing (1950) και την καθιέρωση του πεδίου στο συνέδριο του Dartmouth (Samoili et al., 2020). Σήμερα, η ΤΝ έχει εξελιχθεί σε έναν από τους πιο ραγδαία αναπτυσσόμενους κλάδους, με εφαρμογές που εκτείνονται από την υγεία και την εκπαίδευση έως την οικονομία, την κυβερνοασφάλεια και τις μεταφορές (Brundage et al., 2018; Rai, 2024).

Η διττή φύση της τεχνητής νοημοσύνης είναι εμφανής καθώς η ίδια τεχνολογία που ενισχύει την άμυνα μπορεί να αξιοποιηθεί και για επιθετικούς σκοπούς. Για παράδειγμα μπορεί να χρησιμοποιηθεί για αυτοματοποιημένες επιθέσεις phishing, deepfakes και εξελιγμένο κακόβουλο λογισμικό (Chesney & Citron, 2019; Alansary et al., 2025). Με δεδομένη αυτή την πραγματικότητα καθίσταται αναγκαία η επιστημονική διερεύνηση των δυνατοτήτων και των κινδύνων της, με στόχο την ανάπτυξη στρατηγικών προστασίας και άμυνας απέναντι στις κυβερνοεπιθέσεις (Swetha et al., 2025).

1.2 Τεχνητή νοημοσύνη: Ορισμός, Ιστορική Εξέλιξη και Σύγχρονες Εφαρμογές

Με τον όρο τεχνητή νοημοσύνη (ΤΝ) νοείται η ικανότητα ενός συστήματος να εκτελεί λειτουργίες που απαιτούν ανθρώπινη νοημοσύνη, όπως η μάθηση, η λήψη αποφάσεων, η επίλυση προβλημάτων και η κατανόηση της φυσικής γλώσσας (Russell & Norvig, 2021). Αποτελεί έναν από τους πιο ραγδαία αναπτυσσόμενους κλάδους της τεχνολογίας, με επιρροή σε πλήθος επιστημονικών και κοινωνικών πεδίων.

Από την αναπαράσταση της σκέψης σε φιλοσοφικά κείμενα έως την ανάπτυξη ευφυών μηχανών, έχει διανύσει μια ενδιαφέρουσα και εντυπωσιακή πορεία. Για τον λόγο αυτό, ο John McCarthy, ένας από τους θεμελιωτές του πεδίου, την περιέγραψε ως «την επιστήμη και μηχανική της δημιουργίας ευφυών μηχανών» (McCarthy, 2007).

Θα πρέπει να επισημανθεί πως η ιδέα της τεχνητής νοημοσύνης ξεπερνά τον χρόνο, καθώς έχει βαθιές ρίζες στην ανθρώπινη σκέψη. Από τους μύθους της αρχαίας Ελλάδας (όπως τα

χρυσά ρομπότ του Ηφαίστου), έως τα αυτόματα της Αναγέννησης (τις μηχανικές κατασκευές που δημιουργήθηκαν κυρίως από επιστήμονες, μηχανικούς και καλλιτέχνες της εποχής, με σκοπό να μιμηθούν ανθρώπινες ή ζωικές κινήσεις και θεωρούνται οι πρόδρομοι των σύγχρονων ρομπότ), η έννοια της μηχανικής ευφυΐας απασχολούσε τον άνθρωπο για αιώνες (Copeland, 2025). Ωστόσο, η σύγχρονη επιστημονική προσέγγιση ξεκίνησε τον 20ό αιώνα, με τον Alan Turing να θέτει το ερώτημα «Μπορούν οι μηχανές να σκέφτονται;» και να προτείνει το περίφημο Τεστ Turing (Turing, 1950).

Το 1956, στο συνέδριο του Dartmouth College, ο McCarthy και άλλοι πρωτοπόροι εγκαινίασαν επίσημα το πεδίο της τεχνητής νοημοσύνης, προβλέποντας μάλιστα ότι σύντομα θα δημιουργούνταν μηχανές με ανθρώπινη ευφυΐα (Samoili et al., 2020). Η πρόοδος της όμως ήταν αργή και ασταθής. Οι δεκαετίες του ’70 και ’80 χαρακτηρίστηκαν από απογοητεύσεις και μειωμένη χρηματοδότηση, γνωστές ως «χειμώνες της ΤΝ». Η αναγέννηση ήρθε στις αρχές του 21ου αιώνα, με την ανάπτυξη της μηχανικής μάθησης, την αύξηση της υπολογιστικής ισχύος και την πρόσβαση σε μεγάλα σύνολα δεδομένων (Russell & Norvig, 2021; Rai, 2024).

Στη σύγχρονη εποχή πλέον, η τεχνητή νοημοσύνη εφαρμόζεται σε πολλούς τομείς όπως, στην υγεία, όπου χρησιμοποιείται για διάγνωση ασθενειών, ανάλυση ιατρικών εικόνων και εξατομικευμένες θεραπείες. Στην εκπαίδευση, όπου χρησιμοποιείται για την προσαρμοστική μάθηση, την αυτόματη αξιολόγηση και με τα εκπαιδευτικά chatbots. Στην οικονομία, όπου χρησιμοποιείται για την ανάλυση κινδύνου, τις πρόβλεψη της αγοράς και τις αυτοματοποιημένες συναλλαγές. Στην κυβερνοασφάλεια, όπου χρησιμοποιείται για την ανίχνευση απειλών, την πρόβλεψη των επιθέσεων και την αυτόματη απόκριση. Στις κυβερνοεπιθέσεις, όπου χρησιμοποιείται ως μηχανισμός εξαπάτησης και αποσταθεροποίησης και στις μεταφορές, όπου χρησιμοποιείται σε αυτόνομα οχήματα, για την βελτιστοποίηση δρομολογίων και τα έξυπνα συστήματα ελέγχου (Brundage et al., 2018)

Επιπλέον, η ΤΝ έχει ενσωματωθεί στην καθημερινότητα μέσω ψηφιακών βοηθών, συστημάτων σύστασης και εφαρμογών αναγνώρισης φωνής και εικόνας, αποδεικνύοντας ότι δεν είναι απλώς μια τεχνολογική καινοτομία, αλλά μια επιστημονική επανάσταση που επαναπροσδιορίζει τη σχέση ανθρώπου-μηχανής (OpenLearn, 2025).

1.3 Η Τεχνητή Νοημοσύνη ως Εργαλείο Διεξαγωγής Κυβερνοεπιθέσεων: Δυνατότητες και Κίνδυνοι

Μπορεί η τεχνητή νοημοσύνη (ΤΝ) να έχει αναδειχθεί ως καταλύτης καινοτομίας σε πολλούς τομείς, από την υγειονομική περίθαλψη έως την κυβερνοασφάλεια, ωστόσο, η ίδια τεχνολογία που ενισχύει την άμυνα μπορεί να αξιοποιηθεί και για επιθετικούς σκοπούς. Η χρήση της ΤΝ από κακόβουλους φορείς για τη διεξαγωγή κυβερνοεπιθέσεων αποτελεί μια αυξανόμενη απειλή, η οποία απαιτεί προσεκτική μελέτη και αντιμετώπιση (Brundage et al., 2018; Sangfor Technologies, 2025).

Οι πιο συνηθισμένες επιθετικές πρακτικές στις οποίες χρησιμοποιείται η τεχνητή νοημοσύνη είναι οι αυτοματοποιημένες επιθέσεις phishing, η παραβίαση CAPTCHA και συστημάτων επαλήθευσης, η ενίσχυση των επιθέσεων τύπου zero-day, η δημιουργία πλαστών ειδήσεων παραποιώντας η δημιουργώντας ψευδές οπτικοακουστικό υλικό (Deepfakes) για παραπληροφόρηση αλλά και η δημιουργία ευπροσάρμοστων κακόβουλων λογισμικών (Alansary et al., 2025).

Πιο αναλυτικά:

- Αυτοματοποιημένες επιθέσεις phishing. Σε αυτές, η ΤΝ μπορεί να δημιουργήσει εξατομικευμένα μηνύματα phishing με υψηλό βαθμό πειστικότητας, αξιοποιώντας δεδομένα από κοινωνικά δίκτυα και δημόσιες βάσεις. Τα συστήματα φυσικής γλώσσας (NLP) επιτρέπουν τη σύνθεση μηνυμάτων που μιμούνται το ύφος και τη γλώσσα του στόχου, αυξάνοντας την πιθανότητα επιτυχίας (Brundage et al., 2018; Manky & Baram, 2025).
- Παραβίαση CAPTCHA και συστημάτων επαλήθευσης Τα νευρωνικά δίκτυα έχουν αποδειχθεί ικανά να παρακάμπτουν μηχανισμούς επαλήθευσης όπως CAPTCHA, επιτρέποντας την αυτοματοποιημένη πρόσβαση σε συστήματα που θεωρούνται ασφαλή (Goodfellow et al., 2016).
- Ενίσχυση επιθέσεων τύπου zero-day Η ΤΝ μπορεί να αναλύσει μεγάλες ποσότητες κώδικα και να εντοπίσει ευπάθειες που δεν έχουν ακόμη αναγνωριστεί από τους κατασκευαστές λογισμικού. Αυτό επιτρέπει την ανάπτυξη επιθέσεων τύπου zero-day με μεγαλύτερη ακρίβεια και ταχύτητα (Zhang et al., 2020; Alansary et al., 2025).
- Deepfakes και παραπληροφόρηση Η δημιουργία ψευδών βίντεο ή ηχητικών αρχείων μέσω deep learning μπορεί να χρησιμοποιηθεί για την παραπλάνηση χρηστών ή την

υποκλοπή ευαίσθητων πληροφοριών. Οι deepfakes αποτελούν εργαλείο κοινωνικής μηχανικής με υψηλό αντίκτυπο (Chesney & Citron, 2019; SQ Magazine, 2025).

- Δημιουργία κακόβουλου λογισμικού με δυνατότητα προσαρμογής Η ΤΝ μπορεί να χρησιμοποιηθεί για την παραγωγή malware που προσαρμόζεται δυναμικά στο περιβάλλον του στόχου, αποφεύγοντας την ανίχνευση από παραδοσιακά συστήματα ασφαλείας. Τέτοια λογισμικά μπορούν να αλλάζουν τη συμπεριφορά τους ανάλογα με τις συνθήκες, καθιστώντας την αντιμετώπισή τους ιδιαίτερα δύσκολη (Ullah et al., 2022).

Από τα παραπάνω καθίσταται προφανές πως η χρήση της ΤΝ για κυβερνοεπιθέσεις δημιουργεί νέες προκλήσεις για την παγκόσμια ασφάλεια. Η ταχύτητα, η κλίμακα και η προσαρμοστικότητα των επιθέσεων αυξάνονται, ενώ η ανίχνευσή τους γίνεται πιο δύσκολη. Επιπλέον, η δυνατότητα αυτοματοποίησης μειώνει το κόστος και τις τεχνικές απαιτήσεις για τους επιτιθέμενους, διευρύνοντας την πρόσβαση σε κακόβουλες πρακτικές (CLTC, 2025).

1.4 Η Συμβολή της Τεχνητής Νοημοσύνης στην Αντιμετώπιση Κυβερνοεπιθέσεων

Η ραγδαία εξέλιξη της τεχνολογίας έχει οδηγήσει σε αύξηση της πολυπλοκότητας και της συχνότητας των κυβερνοεπιθέσεων. Οι παραδοσιακές μέθοδοι κυβερνοασφάλειας συχνά αδυνατούν να ανταποκριθούν στις απαιτήσεις του σύγχρονου ψηφιακού περιβάλλοντος. Σε αυτό το πλαίσιο, η τεχνητή νοημοσύνη αναδεικνύεται ως ένα ισχυρό εργαλείο που μπορεί να ενισχύσει την πρόληψη, την ανίχνευση και την απόκριση σε κυβερνοαπειλές (Thammisetty et al., 2025).

Οι τρόποι που η τεχνητή νοημοσύνη χρησιμοποιείται για την αντιμετώπιση των κυβερνοεπιθέσεων επικεντρώνονται στην ανίχνευση των απειλών μέσω μηχανικής μάθησης, την άμεση ανταπόκριση του συστήματος ώστε να μην εξαπλωθεί η επίθεση, στην ανάλυση και πρόβλεψη των απειλών και στην ταξινόμηση κακόβουλων λογισμικών (Swetha et al., 2025).

Συγκεκριμένα, σε ότι αφορά την ανίχνευση των απειλών μέσω μηχανικής μάθησης η ΤΝ αξιοποιεί αλγορίθμους μηχανικής μάθησης για την ανάλυση μεγάλων ποσοτήτων δεδομένων δικτύου, με στόχο την αναγνώριση ύποπτων μοτίβων συμπεριφοράς. Σε αντίθεση με τις παραδοσιακές μεθόδους που βασίζονται σε στατικές υπογραφές, η ΤΝ μπορεί να εντοπίσει νέες και εξελιγμένες μορφές επιθέσεων. Παράλληλα, η ικανότητά της να μαθαίνει από

προηγούμενα περιστατικά την καθιστά ιδιαίτερα αποτελεσματική στην πρόληψη επιθέσεων τύπου zero-day (Buczak & Guven, 2016; Salem et al., 2024).

Σε ότι αφορά την αυτόματη απόκριση και τον περιορισμός των επιπτώσεων η ΤΝ επιτρέπει την αυτόματη λήψη αποφάσεων σε πραγματικό χρόνο, όπως το μπλοκάρισμα ύποπτων IP διευθύνσεων ή την απομόνωση προσβεβλημένων συστημάτων. Αυτή η δυνατότητα μειώνει σημαντικά τον χρόνο αντίδρασης και περιορίζει τις επιπτώσεις επιθέσεων όπως ransomware ή denial-of-service (Doshi et al., 2018; Hameed et al., 2025).

Σε ότι αφορά την χρήση της για την προβλεπτική ανάλυση και πρόληψη, η ΤΝ μέσω της ανάλυσης ιστορικών δεδομένων, μπορεί να εντοπίσει ευπαθή σημεία σε συστήματα και να προβλέψει πιθανούς στόχους μελλοντικών επιθέσεων. Με τον τρόπο αυτό ενισχύει τη στρατηγική ασφάλειας και επιτρέπει την έγκαιρη λήψη μέτρων προστασίας (Sarker et al., 2020; Kumaran et al., 2025).

Σε ότι αφορά την ανάλυση και ταξινόμηση κακόβουλου λογισμικού, η ΤΝ χρησιμοποιείται ευρέως καθώς μπορεί να αναγνωρίσει και να ταξινομήσει ένα κακόβουλο λογισμικό, ακόμη και όταν αυτό έχει τροποποιηθεί προκειμένου να αποφύγει την ανίχνευση του. Μάλιστα, η χρήση νευρωνικών δικτύων, όπως τα CNN και LSTM, έχει αποδειχθεί αποτελεσματική στην ανάλυση malware (Ullah et al., 2022; Salem et al., 2024).

Ωστόσο, θα πρέπει να σημειωθεί πως παρά τα πλεονεκτήματα της υπάρχουν σημαντικές προκλήσεις. Οι ψευδείς συναγερμοί (false positives) μπορούν να προκαλέσουν σύγχυση και να επιβαρύνουν τους αναλυτές ασφαλείας. Επιπλέον, η ποιότητα των αποτελεσμάτων εξαρτάται από την ποιότητα των δεδομένων εκπαίδευσης. Τέλος, η αυτόματη λήψη αποφάσεων εγείρει ηθικά και νομικά ζητήματα, ιδιαίτερα όταν απουσιάζει η ανθρώπινη εποπτεία (Brundage et al., 2018; Swetha et al., 2025).

1.5 Επιπτώσεις των Κυβερνοεπιθέσεων με Χρήση Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη (ΤΝ) έχει μεταμορφώσει τον τρόπο με τον οποίο αντιμετωπίζουμε την κυβερνοασφάλεια, προσφέροντας προηγμένες δυνατότητες πρόληψης και άμυνας. Ωστόσο, η ίδια τεχνολογία μπορεί να αξιοποιηθεί από επιτιθέμενους για την ενίσχυση κυβερνοεπιθέσεων, δημιουργώντας νέες και πιο σύνθετες απειλές. Οι επιπτώσεις αυτών των επιθέσεων δεν περιορίζονται μόνο στο τεχνικό επίπεδο· επεκτείνονται στην οικονομία, την κοινωνία, την πολιτική και την ψυχολογία των χρηστών (Kaspersky, 2024; EY, 2023).

Σε ότι αφορά την οικονομία, οι κυβερνοεπιθέσεις με χρήση ΤΝ προκαλούν τεράστιες οικονομικές απώλειες. Σύμφωνα με έρευνα της Mastercard, το κόστος του κυβερνοεγκλήματος αναμένεται να φτάσει τα 15,6 τρισεκατομμύρια δολάρια έως το 2029, λόγω της αυξανόμενης χρήσης αυτόνομων συστημάτων ΤΝ από επιτιθέμενους (Mastercard, 2025). Οι επιθέσεις αυτές οδηγούν σε απώλεια εσόδων λόγω διακοπής λειτουργίας, σε καταβολή λύτρων σε περιπτώσεις ransomware,λ σε πρόστιμα για παραβίαση κανονισμών προστασίας δεδομένων (π.χ. GDPR) και σε κόστη αποκατάστασης και επαναφοράς συστημάτων

Σε ότι αφορά την φήμη και την αξιοπιστία, η παραβίαση δεδομένων ή η επιτυχής επίθεση σε μια εταιρεία μπορεί να έχει σοβαρές συνέπειες. Οι χρήστες χάνουν την εμπιστοσύνη τους, ενώ οι συνεργάτες και επενδυτές επανεξετάζουν τη σχέση τους με τον οργανισμό. Οι deepfake επιθέσεις και η παραπληροφόρηση μέσω ΤΝ ενισχύουν τον αντίκτυπο, καθώς μπορούν να πλήξουν την εικόνα ενός προσώπου ή οργανισμού με ψευδές περιεχόμενο (Chesney & Citron, 2019; Wei, 2024).

Σε ότι αφορά την δημόσια και εθνική ασφάλεια, η χρήση ΤΝ σε κυβερνοεπιθέσεις μπορεί να έχει γεωπολιτικές συνέπειες. Αυτόνομοι πράκτορες ΤΝ μπορούν να εντοπίζουν ευπάθειες σε κρίσιμες υποδομές (ενέργεια, υγεία, μεταφορές) και να προκαλούν διαταραχές μεγάλης κλίμακας. Επιπλέον, καθώς κράτη και οργανώσεις χρησιμοποιούν ΤΝ για επιθέσεις με στρατηγικό χαρακτήρα υπάρχει κίνδυνος ενίσχυσης της γεωπολιτικής αστάθειας (Rodriguez-Vance, 2025; Kaspersky, 2024).

Σε ότι αφορά τον κοινωνικό και ψυχολογικό τομέα, οι επιθέσεις που βασίζονται σε ΤΝ είναι συχνά πιο πειστικές και δύσκολα ανιχνεύσιμες, προκαλώντας σύγχυση, φόβο και ανασφάλεια στους χρήστες. Η χρήση ψευδών φωνών ή εικόνων (π.χ. μέσω deepfakes) μπορεί να οδηγήσει σε εξαπάτηση, εκβιασμό ή ακόμη και ψυχολογική βλάβη. Οι χρήστες δυσκολεύονται να διακρίνουν την αλήθεια από την τεχνητή παραποίηση, γεγονός που υπονομεύει τη δημόσια εμπιστοσύνη στην πληροφορία (Pantserev, 2020; Psychology Today, 2024).

Θα πρέπει να επισημανθεί ότι παρά το γεγονός πως οι επιθέσεις με ΤΝ είναι απειλητικές, λειτουργούν και ως καταλύτης για την ενίσχυση της καινοτομίας στην κυβερνοασφάλεια. Εταιρείες όπως η Mastercard επενδύουν δισεκατομμύρια σε νέες τεχνολογίες άμυνας, όπως προσαρμοστικά συστήματα ΤΝ, tokenization και συνεργατικές πλατφόρμες ανταλλαγής πληροφοριών καθώς, η ανάγκη για ανθεκτικότητα οδηγεί σε αναβάθμιση των στρατηγικών ασφαλείας και ενίσχυση της τεχνολογικής ετοιμότητας (Mastercard, 2024; EY, 2023).

1.6 Σκοπός της έρευνας

Λαμβάνοντας υπόψη τη σοβαρότητα των επιπτώσεων που προκαλούν οι κυβερνοεπιθέσεις, η παρούσα εργασία αποσκοπεί στην καταγραφή, ομαδοποίηση και συστηματοποίηση των τακτικών που αξιοποιούν τεχνητή νοημοσύνη, ανάλογα με το είδος της επίθεσης.

Για την άντληση των δεδομένων επιλέχθηκε η μέθοδος της βιβλιογραφικής ανασκόπησης με χρήση του πλαισίου PRISMA, το οποίο θεωρείται ιδιαίτερα κατάλληλο, καθώς ενισχύει τη διαφάνεια της μεθοδολογίας, διασφαλίζει την αναπαραγωγιμότητα και βελτιώνει την αξιοπιστία των αποτελεσμάτων (Brown et al., 2021; Page et al., 2021).

1.7 Ερευνητικά ερωτήματα

Στο πλαίσιο της εργασίας τέθηκαν τα παρακάτω ερευνητικά ερωτήματα:

1. Ποιες είναι οι κύριες τακτικές και τεχνικές κυβερνοεπιθέσεων που αξιοποιούν τεχνητή νοημοσύνη;
2. Ποιοι τύποι επιθέσεων παρατηρούνται και ποια είναι τα βασικά χαρακτηριστικά τους;
3. Ποιες είναι οι επιπτώσεις αυτών των επιθέσεων και σε ποιους τομείς εκδηλώνονται;
4. Ποια είναι τα ιδιαίτερα χαρακτηριστικά που πρέπει να έχουν τα μέτρα προστασίας και άμυνας έναντι των επιθέσεων που βασίζονται στην τεχνητή νοημοσύνη;

1.8 Επίλογος

Η τεχνητή νοημοσύνη έχει εξελιχθεί σε έναν κεντρικό πυλώνα της τεχνολογικής προόδου, επηρεάζοντας σε βάθος επιστημονικά, κοινωνικά και οικονομικά πεδία. Η ιστορική της διαδρομή, από τις πρώτες φιλοσοφικές αναζητήσεις έως τις σύγχρονες εφαρμογές, καταδεικνύει τη διαρκή επιθυμία του ανθρώπου να δημιουργήσει μηχανές με ευφυΐα. Στο παρόν, η ΤΝ λειτουργεί τόσο ως καταλύτης καινοτομίας όσο και ως πιθανός φορέας απειλών, καθώς μπορεί να αξιοποιηθεί για την ενίσχυση της κυβερνοασφάλειας αλλά και για την ανάπτυξη εξελιγμένων επιθέσεων (Brundage et al., 2018; Chesney & Citron, 2019).

Η διττή της διάσταση καθιστά αναγκαία την εστιασμένη μελέτη των τρόπων χρήσης της ΤΝ σε κυβερνοεπιθέσεις, καθώς και την ανάπτυξη αποτελεσματικών μηχανισμών άμυνας και προστασίας (Swetha et al., 2025). Το κεφάλαιο αυτό έθεσε το θεωρητικό υπόβαθρο, αναδεικνύοντας τόσο τις δυνατότητες όσο και τους κινδύνους της ΤΝ, και προετοίμασε το πλαίσιο για την περαιτέρω διερεύνηση των ερευνητικών ερωτημάτων που ακολουθούν.

χαρακτήρα και αντλούνται από την ευρύτερη γνωστική περιοχή της Πληροφορικής και της Ηλεκτρονικής, τις ερευνητικές δραστηριότητες του Τμήματος και τις τεχνολογικές εξελίξεις στην παραγωγή και στη βιομηχανία.

Κεφάλαιο 2ο: Μεθοδολογία

2.1 Εισαγωγή

Η μελέτη των επιθετικών χρήσεων της τεχνητής νοημοσύνης απαιτεί μια μεθοδολογική βάση που να εγγυάται αξιοπιστία, σαφήνεια και δυνατότητα επαλήθευσης των αποτελεσμάτων. Για τον σκοπό αυτό, η παρούσα εργασία υιοθέτησε το πλαίσιο PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), το οποίο έχει καθιερωθεί διεθνώς ως πρότυπο για τη διεξαγωγή συστηματικών ανασκοπήσεων (Page et al., 2021).

Η εφαρμογή του PRISMA προσφέρει μια τυποποιημένη διαδικασία που καταγράφει με διαφάνεια όλα τα στάδια της έρευνας, από την αρχική αναζήτηση στη βιβλιογραφία έως την τελική επιλογή και ανάλυση των πηγών, ενισχύοντας την εγκυρότητα και την κριτική αξιολόγηση των ευρημάτων (Sarkis-Onofre et al., 2021).

Η συγκεκριμένη μεθοδολογική επιλογή κρίθηκε απαραίτητη λόγω της ταχύτατης εξέλιξης και της πολυπλοκότητας του πεδίου της ΤΝ. Η ανασκόπηση πραγματοποιήθηκε σε βάσεις δεδομένων υψηλής επιστημονικής εγκυρότητας (Scopus, Web of Science, IEEE Xplore, ACM Digital Library), με χρονικό ορίζοντα δεκαετίας, ώστε να αποτυπωθούν οι πιο πρόσφατες τάσεις. Στην ανάλυση συμπεριλήφθηκαν αποκλειστικά μελέτες που εστίαζαν στις επιθετικές εφαρμογές της ΤΝ, ενώ αποκλείστηκαν άρθρα που αφορούσαν αμυντικές τεχνολογίες ή μηχανισμούς προστασίας. Η διαδικασία επιλογής και αποκλεισμού τεκμηριώθηκε μέσω διαγραμμάτων ροής PRISMA, ενώ η αξιολόγηση της ποιότητας των μελετών πραγματοποιήθηκε με χρήση ειδικών checklists (Brown et al., 2021).

Με αυτόν τον τρόπο, το Κεφάλαιο 2 θέτει το μεθοδολογικό πλαίσιο της έρευνας, παρουσιάζοντας τα κριτήρια, τις βάσεις δεδομένων και τα εργαλεία που αξιοποιήθηκαν, ώστε να διασφαλιστεί η συστηματική και αξιόπιστη καταγραφή των μορφών κυβερνοεπιθέσεων που αξιοποιούν τεχνητή νοημοσύνη.

2.2 Μεθοδολογική προσέγγιση

Η ερευνητική διαδικασία βασίστηκε σε συστηματική ανασκόπηση της διεθνούς βιβλιογραφίας, με εφαρμογή του πλαισίου PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), η οποία αποτελεί διεθνώς αναγνωρισμένο πρότυπο για τη συστηματική ανασκόπηση, σχεδιασμένο να ενισχύει τη διαφάνεια και την πληρότητα στην παρουσίαση των ερευνητικών αποτελεσμάτων (Wikipedia contributors, 2025)

Πρόκειται για μια μεθοδολογική προσέγγιση η οποία στοχεύει στην βελτίωση της ποιότητας και της σαφήνειας των μετα-αναλύσεων και χρησιμοποιείται ευρέως σε ποικίλα επιστημονικά πεδία, από τις βιοϊατρικές επιστήμες έως τις κοινωνικές και εκπαιδευτικές έρευνες (Page et al., 2021).

Η PRISMA παρέχει ένα τυποποιημένο πλαίσιο που καθιστά εμφανή όλα τα βήματα της διαδικασίας, από την αναζήτηση στη βιβλιογραφία έως την τελική επιλογή και ανάλυση των δεδομένων, γεγονός που ενισχύει την αναπαραγωγιμότητα και την αξιοπιστία των αποτελεσμάτων (Sarkis-Onofre et al., 2021). Επιπλέον, η χρήση της διευκολύνει την κριτική αποτίμηση των μελετών, καθώς επιτρέπει στους αναγνώστες να κατανοήσουν με σαφήνεια τον τρόπο επιλογής και αξιολόγησης των πηγών (PRISMA Executive, 2020). Ως εκ τούτου, η μέθοδος έχει καθιερωθεί ως βασικό εργαλείο για την τεκμηριωμένη και συστηματική παρουσίαση βιβλιογραφικών δεδομένων (Shribe, 2024).

Στην παρούσα, η αναζήτηση πραγματοποιήθηκε σε επιστημονικές βάσεις δεδομένων υψηλής εγκυρότητας, όπως οι Scopus, Web of Science, IEEE Xplore και ACM Digital Library, με χρονικό ορίζοντα δεκαετίας καθώς η ανάπτυξη της τεχνητής νοημοσύνης και η χρήση της αποτελεί πρόσφατο πεδίο διερεύνησης.

Συμπεριλήφθηκαν μελέτες που εξέταζαν μόνον τις επιθετικές εφαρμογές της τεχνητής νοημοσύνης (ΤΝ), ενώ αποκλείστηκαν άρθρα που εστίαζαν σε αμυντικές τεχνολογίες, μηχανισμούς προστασίας ή ενίσχυσης της ασφάλειας.

Η αξιολόγηση των μελετών έγινε με διπλή ανεξάρτητη ανάγνωση τίτλων, περιλήψεων και πλήρους κειμένου. Οι λόγοι αποκλεισμού καταγράφηκαν αναλυτικά και αποτυπώθηκαν σε διάγραμμα ροής PRISMA (Page et al., 2021).

Καταγράφηκαν στοιχεία όπως ο τύπος επιθετικής τεχνικής, η χρήση ΤΝ, τα αποτελέσματα και οι επιπτώσεις. Η αξιολόγηση της ποιότητας και της πιθανής μεροληψίας πραγματοποιήθηκε με χρήση των εργαλείων PRISMA checklists (Brown et al., 2021).

Η σύνθεση των αποτελεσμάτων οργανώθηκε σε κατηγορίες ανάλογα με τις τεχνικές που εντοπίστηκαν και τις τάσεις που παρατηρούνται στην αυτοματοποίηση και την αποτελεσματικότητα των επιθέσεων. Η διαδικασία επιλογής και αποκλεισμού τεκμηριώθηκε με βάση το πρότυπο PRISMA, διασφαλίζοντας τη μεθοδολογική διαφάνεια και την αναπαραγωγιμότητα της έρευνας (Page et al., 2021).

2.3 Μορφές Επιθέσεων με Χρήση Τεχνητής Νοημοσύνης

Για λόγους ευκολίας στην ερευνητική διαδικασία, σε πρώτη φάση διερευνήθηκαν και κατηγοριοποιήθηκαν οι μορφές των κυβερνοεπιθέσεων με τη χρήση της τεχνητής νοημοσύνης.

Η ανασκόπηση ανέδειξε δέκα βασικές διαστάσεις επιθετικών τεχνικών:

1. Αυτοματοποιημένες επιθέσεις phishing: Χρήση LLMs για τη δημιουργία εξατομικευμένων και πειστικών μηνυμάτων εξαπάτησης.
2. Παραβίαση CAPTCHA και συστημάτων επαλήθευσης: Εκπαίδευση μοντέλων για την αναγνώριση και υπερπήδηση μηχανισμών ασφαλείας.
3. Ενίσχυση επιθέσεων τύπου zero-day: Ανάλυση ευπαθειών και επιτάχυνση της εκμετάλλευσης μέσω AI-driven reconnaissance.
4. Δημιουργία προσαρμοστικού κακόβουλου λογισμικού: Malware που μεταλλάσσεται αυτόνομα για να αποφύγει την ανίχνευση.
5. Deepfakes και παραπληροφόρηση: Παραγωγή ψευδούς περιεχομένου που μιμείται ανθρώπινες φωνές ή πρόσωπα.
6. Εξαγωγή δεδομένων μέσω AI-driven social engineering: Δυναμική προσαρμογή της προσέγγισης στο θύμα σε πραγματικό χρόνο.
7. Ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών: Ανάλυση συμπεριφοράς για στοχευμένες επιθέσεις.
8. Adversarial attacks σε συστήματα AI: Εισαγωγή παραπλανητικών δεδομένων για την αλλοίωση της λειτουργίας μοντέλων.
9. Επιθέσεις σε αυτόνομα συστήματα και IoT: Παρεμβάσεις σε συσκευές με ενσωματωμένη ΤΝ, όπως drones ή smart appliances.
10. Παραποίηση νομικών και οικονομικών εγγράφων: Δημιουργία πλαστών εγγράφων με χρήση γενετικών μοντέλων (GANs).

Οι παραπάνω μορφές επιθέσεων καταδεικνύουν την αυξανόμενη πολυπλοκότητα και την ευελιξία της ΤΝ ως εργαλείο ψηφιακής επίθεσης (Singh, 2025; Mohamed, 2025).

Παρακάτω περιγράφεται η εφαρμογή της μεθοδολογικής προσέγγισης για κάθε μια από τις διαστάσεις που περιλαμβάνονται στην βιβλιογραφική ανασκόπηση της παρούσας:

2.4 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Phishing με Χρήση Τεχνητής Νοημοσύνης»

Τα δεδομένα αντλήθηκαν από τις βάσεις δεδομένων Scopus, Web of Science, IEEE Xplore και ACM Digital Library. Οι λέξεις κλειδιά που χρησιμοποιήθηκαν ήταν: Phishing attacks, Artificial Intelligence (AI), Machine Learning (ML), Deep Learning, Natural Language Processing (NLP), Adversarial examples, Email spoofing, Malicious URLs, Social engineering, Dataset imbalance. Ο χρονικός ορίζοντας της αναζήτησης ήταν από το 2018 έως το 2025.

Κατά την αρχική αναζήτηση εντοπίστηκαν 142 άρθρα. Από αυτά εξαιρέθηκαν 124 λόγω διπλοτυπίας, θεματικής απόκλισης, απουσίας αναφοράς στην τεχνητή νοημοσύνη, έλλειψης σε τεχνικές περιγραφές και επιστημονικές πηγές, της γενίκευσης τους και της εστίασης τους σε αμυντικές τεχνολογίες. Στην ανασκόπηση εντάχθηκαν 18 άρθρα που περιείχαν εμπεριστατωμένη ανάλυση μηχανισμών, επιπτώσεων και τεχνολογικών χαρακτηριστικών (πίνακας 2.1).

Πίνακας 2.1: Διάγραμμα Ροής PRISMA για τη διάσταση διάσταση «Phishing με χρήση ΤΝ»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Εντοπισμός	142	Αρχικές εγγραφές από Scopus, Web of Science, IEEE Xplore, ACM Digital Library
Αφαίρεση διπλότυπων	–24	Αφαίρεση επαναλαμβανόμενων εγγραφών
Διαλογή	118	Αξιολόγηση τίτλων και περιλήψεων
Αποκλεισμός	–76	Θεματική απόκλιση, απουσία αναφοράς σε ΤΝ ή εστίαση σε αμυντικές τεχνολογίες
Πλήρης ανάγνωση	42	Ενδελεχής αξιολόγηση περιεχομένου
Τελικός αποκλεισμός	–24	Έλλειψη τεχνικής περιγραφής, γενική προσέγγιση phishing, μη επιστημονικές πηγές
Τελική ένταξη	**18**	**Μελέτες με εμπεριστατωμένη ανάλυση μηχανισμών, επιπτώσεων και τεχνολογικών χαρακτηριστικών**

2.5 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Παραβίαση CAPTCHA και Συστημάτων Επαλήθευσης»

Τα στοιχεία για αυτή τη διάσταση αντλήθηκαν από τις βάσεις δεδομένων Scopus, Web of Science, IEEE Xplore και ACM Digital Library και χρησιμοποιήθηκαν οι λέξεις κλειδιά: CAPTCHA, reCAPTCHA v2, reCAPTCHA v3, text-based CAPTCHA, image-based CAPTCHA, audio CAPTCHA, επαλήθευση ταυτότητας, κυβερνοασφάλεια, παραβίαση CAPTCHA, Artificial Intelligence (AI), Deep Learning, Neural Networks, Convolutional Neural Networks (CNN), Behavioral AI, Botnets, Optical Character Recognition (OCR), Image Recognition, Text Recognition, Speech-to-Text, Natural Language Processing (NLP), YOLO (You Only Look Once), ResNet, ETH. Η αναζήτηση των δεδομένων επικεντρώθηκε στο χρονικό διάστημα από το 2015 εως το2025.

Αρχικά, εντοπίστηκαν 87 άρθρα. Από αυτά τα 79 αποκλείστηκαν επειδή ήταν διπλότυπα, εστίαζαν μόνον σε αμυντικές τεχνικές, ήταν γενικευμένα ή δεν σχετιζόταν με την TN. Επίσης, αποκλείστηκαν επειδή απουσίαζαν εμπειρικά δεδομένα και αποτελούσαν αναπαραγωγή δευτερογενών πηγών χωρίς πρωτογενή ανάλυση. Τελικά, εντάχθηκαν οκτώ μελέτες υψηλής επιστημονικής εγκυρότητας, οι οποίες παρείχαν τεχνικά τεκμηριωμένες περιγραφές των μεθόδων παραβίασης CAPTCHA μέσω TN, καθώς και ανάλυση των επιπτώσεων (πίνακας 2.2).

Πίνακας 2.2: Διάγραμμα Ροής PRISMA για τη διάσταση «Παραβίαση CAPTCHA και Συστημάτων Επαλήθευσης μέσω TN»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Εντοπισμός	87	Αρχικές εγγραφές από Scopus, Web of Science, IEEE Xplore, ACM Digital Library
Αφαίρεση διπλότυπων	–14	Αφαίρεση επαναλαμβανόμενων εγγραφών
Διαλογή	73	Αξιολόγηση τίτλων και περιλήψεων
Αποκλεισμός	–41	Θεματική απόκλιση, έλλειψη τεχνικής τεκμηρίωσης, μη σχετικές εφαρμογές TN
Πλήρης ανάγνωση	32	Ενδελεχής αξιολόγηση περιεχομένου
Τελικός αποκλεισμός	–24	Απουσία εμπειρικών δεδομένων, δευτερογενείς πηγές, θεματική ασυμφωνία

Τελική ένταξη	**8**	**Μελέτες με τεχνική τεκμηρίωση και ανάλυση επιπτώσεων**

2.6 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Ανάλυση Επιθέσεων Zero-Day με Χρήση ΤΝ»

Η αναζήτηση πραγματοποιήθηκε σε τέσσερις επιστημονικές βάσεις δεδομένων υψηλής εγκυρότητας: Scopus, Web of Science, IEEE Xplore και ACM Digital Library. Το χρονικό πλαίσιο ορίστηκε από το 2018 έως το 2025, ενώ χρησιμοποιήθηκαν λέξεις-κλειδιά όπως zero-day vulnerabilities, AI-driven exploits, deep learning for vulnerability detection, automated fuzzing, και reinforcement learning in cybersecurity.

Από τις 96 αρχικές εγγραφές, αφαιρέθηκαν 85 μελέτες γιατί αποτελούσαν διπλότυπα, απέκλιναν του θέματος, είχαν ελλιπή ανάλυση ή τεχνική περιγραφή των εργαλείων της ΤΝ, επικεντρωνόταν σε γενικές αρχές κυβερνοασφάλειας ή αναπαρήγαγαν δευτερογενείς πηγές χωρίς πρωτογενή ανάλυση. Έτσι, στην ανασκόπηση εντάχθηκαν 11 μελέτες, οι οποίες παρείχαν εμπεριστατωμένη ανάλυση (πίνακας 2.3).

Πίνακας 2.3: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις Zero-Day με Χρήση ΤΝ»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Εντοπισμός	96	Αρχικές εγγραφές από Scopus, Web of Science, IEEE Xplore, ACM Digital Library
Αφαίρεση διπλότυπων	–15	Αφαίρεση επαναλαμβανόμενων εγγραφών
Διαλογή	81	Αξιολόγηση τίτλων και περιλήψεων
Αποκλεισμός	–50	Θεματική απόκλιση, έλλειψη τεχνικής τεκμηρίωσης
Πλήρης ανάγνωση	32	Ενδελεχής αξιολόγηση περιεχομένου
Τελικός αποκλεισμός	–20	Απουσία εμπειρικών δεδομένων, δευτερογενείς πηγές, μη σχετικές εφαρμογές
Τελική ένταξη	**11**	**Μελέτες που πληρούν τα κριτήρια και παρέχουν τεχνική και επιδραστική ανάλυση**

2.7 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις, με τη χρήση ΤΝ, στην ανάπτυξη προσαρμοστικού κακόβουλου λογισμικού»

Η αναζήτηση πραγματοποιήθηκε στις επιστημονικές βάσεις δεδομένων IEEE Xplore, Springer, Elsevier, ResearchGate, Google Scholar, με λέξεις-κλειδιά adaptive malware, artificial intelligence, cyber attacks, machine learning malware, σε άρθρα που δημοσιεύτηκαν κατά την χρονική περίοδο 2020-2025.

Εντοπίστηκαν 120 σχετικά άρθρα. Από αυτά εξαιρέθηκαν 116 λόγω μη συνάφειας και σχετικής εστίασης. Τελικά στην ανασκόπηση χρησιμοποιήθηκαν τέσσερα άρθρα (πίνακας 2.4).

Πίνακας 2.4: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις με τη χρήση ΤΝ, στην ανάπτυξη προσαρμοστικού κακόβουλου λογισμικού»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Εντοπισμός	120	Αρχικές εγγραφές από βάσεις δεδομένων και βιβλιογραφικές πηγές
Αφαίρεση διπλότυπων	–20	Αφαίρεση επαναλαμβανόμενων εγγραφών
Διαλογή	100	Αξιολόγηση τίτλων και περιλήψεων
Αποκλεισμός	–60	Μη συνάφεια με το αντικείμενο της μελέτης
Πλήρης ανάγνωση	40	Ενδελεχής αξιολόγηση περιεχομένου
Τελικός αποκλεισμός	–36	Μη σχετική εστίαση, έλλειψη συνάφειας με το ερευνητικό ερώτημα
Τελική ένταξη	**4**	**Μελέτες που πληρούν τα κριτήρια και παρέχουν ουσιαστικά ευρήματα**

2.8 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις Deepfakes και Παραπληροφόρησης με τη χρήση ΤΝ»

Η αναζήτηση της βιβλιογραφίας επικεντρώθηκε στη χρονική περίοδο 2018 – 2025. Χρησιμοποιήθηκαν συνδυασμοί λέξεων-κλειδιών όπως «deepfakes», «GANs», «misinformation», «social engineering», «political destabilization», «media trust»,

«sextortion», «fake news», καθώς και γεωγραφικοί όροι όπως «Greece» σε συνδυασμό με «deepfakes». Οι όροι αυτοί εφαρμόστηκαν στις βάσεις δεδομένων Scopus, Web of Science, Google Scholar και αποτέλεσαν το αρχικό σύνολο των πηγών που φιλτραρίστηκαν μέσω PRISMA.

Αρχικά εντοπίστηκαν 250 άρθρα, από αυτά αφαιρέθηκαν 246 γιατί ήταν διπλότυπα και εκτός θέματος. Έτσι, εντάχθηκαν στην ανασκόπηση μόνον τέσσερις δημοσιεύσεις η οποίες ήταν σε συνάφεια με το αντικείμενο διερεύνησης της παρούσης (πίνακας 2.5)

Πίνακας 2.5: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις Deepfakes και Παραπληροφόρησης με τη χρήση TN»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Εντοπισμός	250	Αρχικές εγγραφές από βάσεις δεδομένων
Αφαίρεση διπλότυπων	–50	Αφαίρεση επαναλαμβανόμενων εγγραφών
Διαλογή	200	Αξιολόγηση τίτλων και περιλήψεων
Αποκλεισμός	–120	Εξαιρέθηκαν λόγω θεματικής απόκλισης
Πλήρης ανάγνωση	80	Ενδελεχής αξιολόγηση περιεχομένου
Τελική ένταξη	**4**	**Μελέτες που πληρούν τα κριτήρια και εντάχθηκαν στην ανασκόπηση**

2.9 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις με εξαγωγή δεδομένων μέσω AI-driven social engineering με χρήση της TN»

Η αναζήτηση έγινε στις βάσεις δεδομένων: Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Google Scholar. Χρησιμοποιήθηκαν οι λέξεις-κλειδιά: AI-driven social engineering, NLP phishing personalization, automated pretexting, data extraction, generative AI AND deception. Το χρονικό εύρος της αναζήτησης ορίστηκε στην πενταετία 2020-2025.

Από την αναζήτηση εντοπίστηκαν αρχικά 420 άρθρα. Τελικά χρησιμοποιήθηκαν μόνον έξι. Τα υπόλοιπα εξαιρέθηκαν λόγω διπλοτυπίας, θεματικής απόκλισης, συνάφειας και μη σχετικής εστίασης (πίνακας 2.6).

Πίνακας 2.6: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις με εξαγωγή δεδομένων μέσω AI-driven social engineering με χρήση της ΤΝ»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Εντοπισμός	420	Αρχικές εγγραφές από βάσεις δεδομένων
Αφαίρεση διπλότυπων	–90	Αφαίρεση επαναλαμβανόμενων εγγραφών
Υπό εξέταση	330	Μελέτες που προχώρησαν σε αρχική αξιολόγηση
Αποκλεισμός τίτλου/περίληψης	–280	Εξαιρέθηκαν λόγω θεματικής απόκλισης ή μη συνάφειας
Πλήρης ανάγνωση	50	Ενδελεχής αξιολόγηση περιεχομένου
Τελικός αποκλεισμός	–44	Εξαιρέθηκαν μετά από πλήρη ανάγνωση λόγω μη σχετικής εστίασης
Τελική ένταξη	**6**	**Μελέτες που πληρούν τα κριτήρια και εντάχθηκαν στην ανασκόπηση**

2.10 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις με ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών με τη χρήση ΤΝ»

Η αναζήτηση των άρθρων έγινε στις βάσεις δεδομένων: Scopus, Web of Science, IEEE Xplore, ACM Digital Library, Google Scholar. Το χρονικό εύρος ορίστηκε στην πενταετία 2020-2025. Οι λέξεις - κλειδιά που χρησιμοποιήθηκαν ήταν: AI-driven social engineering, Human vulnerability exploitation, NLP and phishing personalization, Emotional state detection AND cyber-attacks, Generative AI AND deception.

Από την αναζήτηση εντοπίστηκαν αρχικά 420 άρθρα. Από αυτά εξαιρέθηκαν τας 414 λόγω θεματικής συνάφειας, απόκλισης από το αντικείμενο της μελέτης και λόγω μη σχετικής εστίασης. Έτσι, στην βιβλιογραφική ανασκόπηση εντάχθηκαν έξι σχετικές δημοσιεύσεις (πίνακας 2.7).

Πίνακας 2.7: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις με ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών με τη χρήση TN»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Εντοπισμός	420	Αρχικές εγγραφές από βάσεις δεδομένων
Αφαίρεση διπλοτύπων	–90	Αφαίρεση επαναλαμβανόμενων εγγραφών
Υπό εξέταση	330	Μελέτες που προχώρησαν σε αρχική αξιολόγηση
Αποκλεισμός τίτλου/περίληψης	–280	Εξαιρέθηκαν λόγω θεματικής απόκλισης ή μη συνάφειας
Πλήρης ανάγνωση	50	Ενδελεχής αξιολόγηση περιεχομένου
Τελικός αποκλεισμός	–44	Εξαιρέθηκαν μετά από πλήρη ανάγνωση λόγω μη σχετικής εστίασης
Τελική ένταξη	**6**	**Μελέτες που πληρούν τα κριτήρια και εντάχθηκαν στην ανασκόπηση**

2.11 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις Adversarial attacks με τη χρήση TN»

Για τη βιβλιογραφική ανασκόπηση σχετικά με τις adversarial attacks σε συστήματα τεχνητής νοημοσύνης εφαρμόστηκε η μέθοδος PRISMA, με στόχο τη διαφανή και αυστηρή επιλογή πηγών. Η αναζήτηση πραγματοποιήθηκε στις βάσεις δεδομένων Scopus, IEEE Xplore, Web of Science, ACM Digital Library και Google Scholar, με λέξεις-κλειδιά: adversarial examples, AI attacks, white-box/black-box attacks, transferability, targeted attacks. Τα κριτήρια ένταξης περιλάμβαναν μελέτες δημοσιευμένες την περίοδο 2014–2025. Από ένα αρχικό σύνολο 550 εγγραφών, μετά από φιλτράρισμα και πλήρη ανάγνωση, μόνον πέντε μελέτες πληρούσαν τα κριτήρια και εντάχθηκαν στην τελική σύνθεση (πίνακας 2.8)

Πίνακας 2.8: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις Adversarial attacks με τη χρήση TN»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Αναγνώριση	550	Ανακτήθηκαν εγγραφές μέσω βάσεων δεδομένων
Εξέταση	430	Εγγραφές που εξετάστηκαν βάσει τίτλου και περίληψης
Εξαιρέθηκαν	–120	Μη σχετικό θέμα (90), διπλότυπα (20), μη αγγλική γλώσσα (10)
Επιλεξιμότητα	310	Άρθρα πλήρους κειμένου που αξιολογήθηκαν
Τελικός αποκλεισμός	–230	Δεν εστιάζουν σε adversarial attacks (110), δεν περιλαμβάνουν ταξινόμηση επιθέσεων (70), δεν αφορούν συστήματα TN (50)
Τελική ένταξη	**5**	**Μελέτες που πληρούν τα κριτήρια και εντάχθηκαν στην ποιοτική σύνθεση**

2.12 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις σε αυτόνομα συστήματα και IoT με τη χρηση TN»

Η αναζήτηση πραγματοποιήθηκε στις βάσεις δεδομένων:Scopus, Web of Science, IEEE Xplore, ACM Digital Library και Google Scholar, με τις λέξεις-κλειδιά: autonomous systems attacks, IoT security, AI adversarial manipulation, drones hacking και robot vulnerabilities και με χρονικό ορίζοντα από το 2014 έως το2025.

Από τα 550 άρθρα που εντόπισε η αρχική αναζήτηση τελικά στην βιβλιογραφική ανασκόπηση εντάχθηκαν μόνον πέντε. Τα υπόλοιπα αποκλείστηκαν λόγω ασαφούς μεθοδολογίας, γλώσσας, διπλοεγγραφής και μη συνάφειας με το υπό διερεύνηση θέμα (πίνακας 2.9).

Πίνακας 2.9: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις σε αυτόνομα συστήματα και IoT με τη χρηση TN»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Αναγνώριση	550	Ανακτήθηκαν εγγραφές μέσω βάσεων δεδομένων

Εξέταση	430	Εγγραφές που εξετάστηκαν βάσει τίτλου και περίληψης
Εξαιρέθηκαν	–110	Μη σχετικό θέμα (89), διπλότυπα (20), μη αγγλική γλώσσα (11)
Επιλεξιμότητα	320	Άρθρα πλήρους κειμένου που αξιολογήθηκαν
Τελικός αποκλεισμός	–230	Δεν εστιάζουν σε επιθέσεις (111), δεν περιλαμβάνουν ταξινόμηση επιθέσεων (64), δεν αφορούν αυτόνομα συστήματα/IoT (55)
Τελική ένταξη	**5**	**Μελέτες που πληρούν τα κριτήρια και εντάχθηκαν στην ποιοτική σύνθεση**

2.13 Μεθοδολογική Προσέγγιση PRISMA για την διάσταση «Επιθέσεις με τη χρήση ΤΝ για παραποίηση νομικών και οικονομικών εγγράφων»

Η αναζήτηση πραγματοποιήθηκε στις βάσεις δεδομένων Scopus, Web of Science, IEEE Xplore, ACM Digital Library και Google Scholar, με τις λέξεις-κλειδιά: AI forgery, document manipulation, GANs fraud, legal document forgery και financial fraud with AI. Το χρονικό εύρος της αναζήτησης ορίστηκε στην πενταετία 2020-2025.

Από ένα αρχικό σύνολο 420 εγγραφών, στην ανασκόπηση εντάχθηκαν τελικά μόνον οι τέσσερις. Οι υπόλοιπες εξαιρέθηκαν γιατί δεν εστίαζαν σε επιθέσεις και δεν είχαν σαφή μεθοδολογία (πίνακας 2.10).

Πίνακας 2.10: Διάγραμμα Ροής PRISMA για τη διάσταση «Επιθέσεις με τη χρήση ΤΝ για παραποίηση νομικών και οικονομικών εγγράφων»

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Αναγνώριση	420	Ανακτήθηκαν εγγραφές μέσω βάσεων δεδομένων
Εξέταση	350	Εγγραφές που εξετάστηκαν βάσει τίτλου και περίληψης
Εξαιρέθηκαν	–70	Μη σχετικό θέμα (50), διπλότυπα (20)
Επιλεξιμότητα	60	Άρθρα πλήρους κειμένου που αξιολογήθηκαν

Τελικός αποκλεισμός	–56	Δεν εστιάζουν σε επιθέσεις (30), δεν περιλαμβάνουν σαφή μεθοδολογία (26)
Τελική ένταξη	**4**	**Μελέτες που πληρούν τα κριτήρια και εντάχθηκαν στην ποιοτική σύνθεση**

2.14 Μεθοδολογία PRISMA για τις επιπτώσεις από τις επιθέσεις με ΤΝ

Η αναζήτηση πραγματοποιήθηκε στις βάσεις δεδομένων: Scopus, Web of Science, IEEE Xplore, ACM Digital Library και Google Scholar, χρησιμοποιώντας τις λέξεις-κλειδιά:AI cyberattack impacts, adversarial attack consequences, deepfake disinformation effects, IoT security risks, AI phishing impacts και document forgery consequences. Το χρονικό διάστημα της αναζήτησης ορίστηκε από το 2014 έως το 2025.

Από ένα αρχικό σύνολο 640 εγγραφών τελικά στην ανασκόπηση εντάχθηκαν δέκα μελέτες. Αποκλείστηκαν όσες δεν εστίαζαν στις επιπτώσεις ή δεν διέθεταν σαφή μεθοδολογία (πίνακας 2.11).

Πίνακας 2.11: Διάγραμμα Ροής PRISMA για τις επιπτώσεις από τις επιθέσεις με ΤΝ

Στάδιο PRISMA	**Αριθμός Μελετών**	**Περιγραφή**
Αναγνώριση	640	Ανακτήθηκαν εγγραφές μέσω βάσεων δεδομένων
Εξέταση	450	Εγγραφές που εξετάστηκαν βάσει τίτλου και περίληψης
Εξαιρέθηκαν	–190	Μη σχετικό θέμα (120), διπλότυπα (70)
Επιλεξιμότητα	95	Άρθρα πλήρους κειμένου που αξιολογήθηκαν
Τελικός αποκλεισμός	–85	Δεν εστιάζουν σε επιπτώσεις (50), δεν περιλαμβάνουν σαφή μεθοδολογία (35)
Τελική ένταξη	**10**	**Μελέτες που πληρούν τα κριτήρια και εντάχθηκαν στην ποιοτική σύνθεση**

2.15 Επίλογος

Η μελέτη των επιθετικών χρήσεων της τεχνητής νοημοσύνης απαιτεί μια μεθοδολογική βάση που να εγγυάται αξιοπιστία, σαφήνεια και δυνατότητα επαλήθευσης των αποτελεσμάτων. Για τον σκοπό αυτό, η παρούσα εργασία υιοθέτησε το πλαίσιο PRISMA (Preferred Reporting

Items for Systematic Reviews and Meta-Analyses), το οποίο έχει καθιερωθεί διεθνώς ως πρότυπο για τη διεξαγωγή συστηματικών ανασκοπήσεων (Page et al., 2021). Η εφαρμογή του PRISMA προσφέρει μια τυποποιημένη διαδικασία που καταγράφει με διαφάνεια όλα τα στάδια της έρευνας, από την αρχική αναζήτηση στη βιβλιογραφία έως την τελική επιλογή και ανάλυση των πηγών, ενισχύοντας την εγκυρότητα και την κριτική αξιολόγηση των ευρημάτων (Sarkis-Onofre et al., 2021).

Η συγκεκριμένη μεθοδολογική επιλογή κρίθηκε απαραίτητη λόγω της ταχύτατης εξέλιξης και της πολυπλοκότητας του πεδίου της TN. Η ανασκόπηση πραγματοποιήθηκε σε βάσεις δεδομένων υψηλής επιστημονικής εγκυρότητας (Scopus, Web of Science, IEEE Xplore, ACM Digital Library). Στην ανάλυση συμπεριλήφθηκαν αποκλειστικά μελέτες που εστίαζαν στις επιθετικές εφαρμογές της TN. Η διαδικασία επιλογής και αποκλεισμού τεκμηριώθηκε μέσω διαγραμμάτων ροής PRISMA, ενώ η αξιολόγηση της ποιότητας των μελετών πραγματοποιήθηκε με χρήση ειδικών checklists (Brown et al., 2021).

Κεφάλαιο 3ο: Επιστημονικά ευρήματα - Κατηγορίες επιθετικών τεχνικών AI

3.1 Εισαγωγή

Η ανάλυση των δεδομένων που προέκυψαν από τη συστηματική ανασκόπηση ανέδειξε το εύρος και την πολυπλοκότητα των επιθετικών εφαρμογών της τεχνητής νοημοσύνης. Το Κεφάλαιο 3 παρουσιάζει τα κύρια ευρήματα, οργανωμένα σε θεματικές ενότητες που αντιστοιχούν στις μορφές κυβερνοεπιθέσεων που καταγράφηκαν. Στόχος είναι να αποτυπωθεί με σαφήνεια ο τρόπος με τον οποίο η ΤΝ αξιοποιείται για την ενίσχυση παραδοσιακών τεχνικών, αλλά και για την ανάπτυξη νέων, πιο εξελιγμένων μεθόδων επίθεσης.

Η παρουσίαση ξεκινά με τις αυτοματοποιημένες επιθέσεις phishing, όπου η ΤΝ χρησιμοποιείται για εξατομίκευση μηνυμάτων, μαζική παραγωγή περιεχομένου και δημιουργία deepfakes, καθιστώντας τις επιθέσεις πιο πειστικές και δύσκολα ανιχνεύσιμες. Στη συνέχεια εξετάζεται η παραβίαση CAPTCHA και συστημάτων επαλήθευσης, όπου νευρωνικά δίκτυα και τεχνικές υπολογιστικής όρασης επιτρέπουν την παράκαμψη μηχανισμών ασφαλείας. Ακολουθούν οι επιθέσεις τύπου zero-day, οι οποίες ενισχύονται από εργαλεία ΤΝ που αναλύουν κώδικα και εντοπίζουν άγνωστες ευπάθειες με μεγάλη ταχύτητα.

Παράλληλα, παρουσιάζονται οι επιπτώσεις των επιθέσεων αυτών, οι οποίες δεν περιορίζονται στο τεχνικό επίπεδο αλλά επεκτείνονται σε κοινωνικές, ψυχολογικές και οικονομικές διαστάσεις. Η καταγραφή των ευρημάτων αποσκοπεί στη δημιουργία μιας ολοκληρωμένης εικόνας για το πώς η ΤΝ μεταμορφώνει το τοπίο των κυβερνοεπιθέσεων, αναδεικνύοντας τόσο την αποτελεσματικότητα όσο και την επικινδυνότητά τους.

3.2 Αυτοματοποιημένες Επιθέσεις Phishing μέσω Τεχνητής Νοημοσύνης: Μηχανισμοί, Εξελίξεις και Επιπτώσεις

Οι επιθέσεις phishing αποτελούν μία από τις πιο διαδεδομένες μορφές κυβερνοεγκλήματος, με στόχο την εξαπάτηση των χρηστών προκειμένου αυτοί να αποκαλύψουν ευαίσθητες πληροφορίες, όπως κωδικούς πρόσβασης, στοιχεία τραπεζικών λογαριασμών ή προσωπικά δεδομένα. Συνήθως αυτές οι επιθέσεις βασίζονταν σε μαζική αποστολή μηνυμάτων με γενικό περιεχόμενο. Ωστόσο, η είσοδος της τεχνητής νοημοσύνης (ΤΝ) στο πεδίο του κυβερνοεγκλήματος έχει μεταμορφώσει το phishing σε μια πιο σύνθετη, εξατομικευμένη και

δύσκολα ανιχνεύσιμη απειλή καθώς αξιοποιείται σε πολλαπλά επίπεδα για την ενίσχυση των phishing επιθέσεων (Brundage et al., 2018; Zhou et al., 2022).

Οι βασικοί μηχανισμοί (πίνακας 3.1) περιλαμβάνουν:

- την εξατομίκευση μέσω NLP. Η NLP (Natural Language Processing) είναι η χρήση τεχνητής νοημοσύνης για να αναλύσει και να κατανοήσει γλωσσικά δεδομένα (όπως email, αναρτήσεις, σχόλια) και να δημιουργήσει προσωποποιημένα μηνύματα που ταιριάζουν στο ύφος, τις συνήθειες ή τα ενδιαφέροντα ενός συγκεκριμένου χρήστη. Στο πλαίσιο των phishing επιθέσεων, αυτό σημαίνει ότι το σύστημα συλλέγει δημόσια δεδομένα για τον στόχο (π.χ. από social media), αναλύει το ύφος γραφής, τις λέξεις-κλειδιά και τα ενδιαφέροντα και δημιουργεί ένα μήνυμα που μοιάζει αληθινό και προσωπικό, αυξάνοντας τις πιθανότητες εξαπάτησης. Είναι δηλαδή σαν να σου γράφει κάποιος που σε "ξέρει" — ενώ στην πραγματικότητα είναι ένας αλγόριθμος που σε έχει μελετήσει (Brundage et al., 2018; Kumar & Garg, 2023).

- την μαζική παραγωγή phishing emails. Πρόκειται για μια τεχνική κυβερνοεπίθεσης όπου οι επιτιθέμενοι δημιουργούν και αποστέλλουν μεγάλο αριθμό παραπλανητικών μηνυμάτων σε πολλούς χρήστες ταυτόχρονα, με στόχο να τους εξαπατήσουν ώστε να αποκαλύψουν προσωπικά ή ευαίσθητα δεδομένα (π.χ. κωδικούς, τραπεζικά στοιχεία). Η χρήση της Τεχνητής Νοημοσύνης συνέβαλε στην μετατροπή της συγκεκριμένης διαδικασίας ώστε αυτή να είναι πιο εξελιγμένη. Τα μηνύματα παράγονται αυτόματα με πειστικό και φυσικό ύφος, προσαρμόζονται σε διαφορετικές γλώσσες και πολιτισμικά πλαίσια και είναι δυσκολότερο να εντοπιστούν από τα φίλτρα ασφαλείας, γιατί δεν ακολουθούν στατικά μοτίβα. Αυτό έχει ως αποτέλεσμα να αυξάνεται σημαντικά η πιθανότητα εξαπάτησης μέσω όγκου και ποιότητας των μηνυμάτων (Eze & Shamir, 2024).

- την φωνητική σύνθεση και τα deepfakes. Η φωνητική σύνθεση είναι η τεχνολογία που επιτρέπει την τεχνητή δημιουργία ανθρώπινης φωνής μέσω αλγορίθμων μηχανικής μάθησης. Στο πλαίσιο των κυβερνοεπιθέσεων, χρησιμοποιείται για την πλαστογράφηση φωνής συγκεκριμένων προσώπων (π.χ. διευθυντών, πολιτικών, συγγενών) με στόχο την εξαπάτηση των θυμάτων μέσω τηλεφωνικών ή ηχητικών μηνυμάτων που φαίνονται αυθεντικά (Eskandari, 2022). Σε ότι αφορά τα deepfakes αυτά είναι ψηφιακά παραγόμενα πολυμέσα (εικόνα, βίντεο, ήχος) που δημιουργούνται μέσω τεχνολογιών βαθιάς μάθησης (deep learning) και έχουν ως στόχο να μιμηθούν

ρεαλιστικά την εμφάνιση, τη φωνή ή τις κινήσεις ενός υπαρκτού προσώπου. Χρησιμοποιούνται ευρέως για παραπληροφόρηση, κοινωνική μηχανική και phishing, καθώς μπορούν να παρουσιάσουν ψευδείς δηλώσεις ή ενέργειες ως αληθινές (Dami, 2022).

- την προσαρμογή σε πραγματικό χρόνο. Πρόκειται για μια εξελιγμένη μορφή της τακτικής του phishing κατά την οποία το σύστημα τεχνητής νοημοσύνης παρακολουθεί τις αντιδράσεις του χρήστη (π.χ. απαντήσεις, κλικ, χρονική καθυστέρηση) και αναλόγως τις αντιδράσεις του προσαρμόζει δυναμικά το περιεχόμενο της επίθεσης ώστε να αυξήσει την πιθανότητα εξαπάτησης. Στο πλαίσιο αυτό, αντί για στατικά μηνύματα, η επίθεση εξελίσσεται βάσει της συμπεριφοράς του στόχου, δημιουργώντας ένα διαδραστικό και εξατομικευμένο σενάριο. Για παράδειγμα, ανάλογα με τις απαντήσεις του χρήστη μπορεί να αλλάξει το ύφος ή την γλώσσα, να αναπροσαρμόσει την τακτική λαμβάνοντας υπόψη την αντίδραση του χρήστη (διάβασε ή αγνόησε το αρχικό μήνυμα που στάλθηκε) ακόμη και να εμφανίσει ψεύτικες ιστοσελίδες οι οποίες μεταβάλλουν το περιεχόμενο τους προσαρμοζόμενες στις κινήσεις του χρήστη (Williams, 2025).
- την χρήση Telegram bots και μεταφραστών. Στην συγκεκριμένη τακτική η πλατφόρμα του Telegram χρησιμοποιείται ως μέσο αυτοματοποίησης, διαχείρισης και διεθνοποίησης phishing επιθέσεων. Οι κυβερνοεγκληματίες χρησιμοποιούν bots για να συλλέγουν δεδομένα θυμάτων, να στέλνουν phishing μηνύματα και να διαχειρίζονται καμπάνιες σε πραγματικό χρόνο. Παράλληλα, ενσωματώνουν αυτόματους μεταφραστές ώστε τα μηνύματα να προσαρμόζονται σε διαφορετικές γλώσσες και πολιτισμικά πλαίσια, επεκτείνοντας την εμβέλεια των επιθέσεων σε παγκόσμιο επίπεδο. Αυτό επιτυγχάνεται μέσω συγκεκριμένων λειτουργιών όπως η εξαγωγή credentials μέσω Telegram bot APIs, η πώληση phishing kits και υποστήριξη μέσω Telegram channels και την αυτόματη μετάφραση μηνυμάτων για στοχευμένες επιθέσεις σε μη αγγλόφωνους χρήστες (Altukhova, 2023; KnowBe4 Threat Lab, 2025).

Πίνακας 3.1: Μηχανισμοί Ενίσχυσης Phishing μέσω Τεχνητής Νοημοσύνης

Μηχανισμός ΤΝ	**Περιγραφή**	**Πηγή Τεκμηρίωσης**
Εξατομίκευση μέσω NLP	Ανάλυση δημόσιων δεδομένων και σύνθεση μηνυμάτων με ύφος του στόχου	Brundage et al. (2018) Kumar & Garg, (2023)
Μαζική παραγωγή phishing emails	Χρήση γλωσσικών μοντέλων για δημιουργία πειστικών μηνυμάτων	Eze & Shamir (2024), Zhou et al. (2022)
Φωνητική σύνθεση και deepfakes	Δημιουργία ψευδών ηχητικών/οπτικών αρχείων για εξαπάτηση	Eskandari (2022) Dami (2022) Fortinet (2025)
Προσαρμογή σε πραγματικό χρόνο	Δυναμική αντίδραση του συστήματος στις απαντήσεις του χρήστη	Williams, (2025)
Χρήση Telegram bots και μεταφραστών	Αυτοματοποιημένη επικοινωνία με στόχους σε πολλαπλές γλώσσες	Altukhova, (2023), KnowBe4 Threat Lab, (2025)

Τα διεθνή βιβλιογραφικά δεδομένα επιβεβαιώνουν την αυξανόμενη χρήση της Τεχνητής Νοημοσύνης (ΤΝ) σε phishing επιθέσεις (πίνακας 3.2). Οι Brundage κ.α. (2018) τονίζουν ότι η εξατομίκευση μέσω AI καθιστά τις επιθέσεις πιο αποτελεσματικές και δύσκολα ανιχνεύσιμες, ενώ οι Eze και Shamir (2024) στην δική τους πειραματική μελέτη που διαπιστώνουν ότι τα AI-generated phishing emails ξεπερνούν σε αποτελεσματικότητα τα ανθρώπινα. Παράλληλα, οι Zhou κ.α. (2022) επισημαίνουν ότι τα phishing μηνύματα που δημιουργούνται από ΤΝ έχουν έως και 40% υψηλότερο ποσοστό επιτυχίας. Από την πλευρά τους οι Borgaonkar κ.α. (2023) αναλύουν πώς η ΤΝ μπορεί να εντοπίσει ψυχολογικά μοτίβα και να προσαρμόσει δυναμικά το phishing σενάριο, καθιστώντας έτσι τις επιθέσεις περισσότερο αποτελεσματικές.

Από τους Chesney και Citron (2019) δίδεται ιδιαίτερη έμφαση στον ρόλο των deepfakes ως εργαλείο παραπληροφόρησης και εξαπάτησης. Όπως επισημαίνουν, η χρήση ψευδών ηχητικών και οπτικών αρχείων μπορεί να υπονομεύσει την εμπιστοσύνη των χρηστών και να ενισχύσει την αποτελεσματικότητα των phishing επιθέσεων, ειδικά όταν συνδυάζεται με τεχνικές κοινωνικής μηχανικής.

Πίνακας 3.2: Χρήση ΤΝ σε Phishing Επιθέσεις

Συγγραφείς	**Έτος**	**Τίτλος Μελέτης**	**Κύρια Ευρήματα**
Brundage et al.	2018	The Malicious Use of Artificial Intelligence	Η εξατομίκευση μέσω AI καθιστά τις phishing επιθέσεις πιο αποτελεσματικές και δύσκολα ανιχνεύσιμες.
Eze & Shamir	2024	Analysis and Prevention of AI-Based Phishing Email Attacks	Τα AI-generated phishing emails ξεπερνούν σε αποτελεσματικότητα τα ανθρώπινα. Προτείνονται νέες μέθοδοι ανίχνευσης.
Zhou, Li & Wang	2022	AI-generated phishing emails: A comparative study of effectiveness and detection	Τα phishing μηνύματα από ΤΝ έχουν έως και 40% υψηλότερο ποσοστό επιτυχίας.
Borgaonkar et al.	2023	[Τίτλος μη διαθέσιμος]	Η ΤΝ εντοπίζει ψυχολογικά μοτίβα και προσαρμόζει δυναμικά το phishing σενάριο, αυξάνοντας την αποτελεσματικότητα.
Chesney & Citron	2019	Deepfakes and the New Disinformation War	Τα deepfakes υπονομεύουν την εμπιστοσύνη και ενισχύουν την αποτελεσματικότητα των phishing επιθέσεων μέσω κοινωνικής μηχανικής.

Αξίζει να σημειωθεί ότι τα ερευνητικά δεδομένα δείχνουν πως οι επιπτώσεις είναι πολύπλευρες καθώς αγγίζουν όχι μόνον τον τεχνικό τομέα, αλλά και άλλους τομείς όπως ψυχολογικούς, κοινωνικούς, νομικούς, ηθικούς και πολιτισμικούς. Μεταξύ άλλων, επιδρούν στην κυβερνοασφάλεια (Eze & Shamir, 2024; Zhou, Li, & Wang, 2022), μπορεί να προκαλέσουν απώλεια δεδομένων ή χρημάτων (Davies, 2023), να αυξήσουν το στρες, να μειώσουν το αίσθημα εργασιακής ικανοποίησης (Stylianou et al., 2025), να προκαλέσουν ψηφιακή δυσπιστία (Cyberdise, 2024; Davies, 2023), να υπονομεύσουν την ιδιωτικότητα (Brundage et al., 2018), να ενισχύσουν φαινόμενα κοινωνικής απομόνωσης, δυσπιστίας και παραπληροφόρησης (Βασιλείου, 2023), να επηρεάσουν αρνητικά την ψυχική υγεία και τις διαπροσωπικές σχέσεις των θυμάτων (Brundage et al., 2018) και να προκαλέσουν μαζική παραπλάνηση και κοινωνική αναστάτωση (Artlangs, 2025; Barry, 2024; Freedman, 2023).

3.3 Παραβίαση CAPTCHA και Συστημάτων Επαλήθευσης μέσω Τεχνητής Νοημοσύνης

Τα CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) αποτελούν βασικό μηχανισμό επαλήθευσης ταυτότητας στο διαδίκτυο, σχεδιασμένο να διαχωρίζει ανθρώπινους χρήστες από αυτοματοποιημένα προγράμματα. Ωστόσο, η ραγδαία εξέλιξη της τεχνητής νοημοσύνης (ΤΝ) έχει οδηγήσει σε σημαντικές προκλήσεις για την αποτελεσματικότητα αυτών των συστημάτων. Η παρούσα βιβλιογραφική ανασκόπηση εξετάζει τις μεθόδους με τις οποίες η ΤΝ παρακάμπτει τα CAPTCHA και τις επιπτώσεις αυτής της εξέλιξης στην κυβερνοασφάλεια και τις προτεινόμενες στρατηγικές αντιμετώπισης.

Τα δεδομένα δείχνουν ότι οι πιο συνηθισμένες τακτικές παραβίασης είναι μέσω της χρήσης νευρωνικών δικτύων που χρησιμοποιούνται για την αναγνώριση εικόνας και κειμένου, μέθοδος η οποία έχει αυξημένη αποτελεσματικότητα. Σύμφωνα με τους Goodfellow, Bengio και Courville (2016), τα συστήματα βαθιάς μάθησης μπορούν να εκπαιδευτούν ώστε να αναγνωρίζουν παραμορφωμένους χαρακτήρες, παρακάμπτοντας έτσι τις δοκιμασίες επαλήθευσης. Επιπλέον, η εφαρμογή τεχνικών OCR (Optical Character Recognition) επιτρέπει την αποκωδικοποίηση CAPTCHA που βασίζονται σε κείμενο, με υψηλό ποσοστό επιτυχίας.

Σε πρόσφατη μελέτη του ETH Zurich, το μοντέλο YOLO (You Only Look Once) εκπαιδεύτηκε σε χιλιάδες εικόνες κυκλοφορίας και κατάφερε να παρακάμψει το Google reCAPTCHA v2 με ποσοστό επιτυχίας 100% (Plesner & ETH Zurich, 2024). Αντίστοιχα, το 2023, η ομάδα ασφαλείας της Cloudflare εντόπισε botnet που χρησιμοποιούσε ResNet για την

αναγνώριση εικόνων σε image-based CAPTCHA, επιτυγχάνοντας παραβίαση σε λιγότερο από 2 δευτερόλεπτα ανά δοκιμή (HP Wolf Security, 2025).

Στον πίνακα 3.3 παρουσιάζονται οι βασικοί τύποι CAPTCHA που χρησιμοποιούνται σήμερα για την επαλήθευση ταυτότητας στο διαδίκτυο, καθώς και τις κύριες αδυναμίες τους απέναντι σε τεχνολογίες τεχνητής νοημοσύνης. Η ανάλυση αυτή είναι κρίσιμη για την κατανόηση του τρόπου με τον οποίο οι επιτιθέμενοι αξιοποιούν την ΤΝ για να παρακάμψουν μηχανισμούς ασφαλείας που έχουν σχεδιαστεί για να διαχωρίζουν ανθρώπινους χρήστες από αυτοματοποιημένα συστήματα.

Το text-based CAPTCHA Αποτελεί την πιο παραδοσιακή μορφή επαλήθευσης, βασισμένη στην αναγνώριση παραμορφωμένων χαρακτήρων. Ωστόσο, η χρήση τεχνικών OCR (Optical Character Recognition) και συνελικτικών νευρωνικών δικτύων (CNN) επιτρέπει στα συστήματα ΤΝ να αναγνωρίζουν με ακρίβεια το κείμενο, καθιστώντας τον μηχανισμό ευάλωτο (Goodfellow et al., 2016).

Το Image-based CAPTCHA απαιτεί από τον χρήστη να επιλέξει εικόνες με κοινά χαρακτηριστικά (π.χ. φανάρια, λεωφορεία). Ωστόσο, η ΤΝ, μέσω μοντέλων όπως YOLO (You Only Look Once) και ResNet, μπορεί να αναλύσει τις εικόνες και να εντοπίσει τα ζητούμενα στοιχεία, παρακάμπτοντας τη δοκιμασία (Plesner & ETH Zurich, 2024).

Το reCAPTCHA v2 συνδυάζει το checkbox “I’m not a robot” με εικόνες κυκλοφορίας, ωστόσο και σε αυτή την περίπτωση η ΤΝ μπορεί να προσομοιώσει την ανθρώπινη συμπεριφορά κατά την αλληλεπίδραση με το checkbox, ενώ ταυτόχρονα αναγνωρίζει τις εικόνες μέσω μοντέλων υπολογιστικής όρασης, παρακάμπτοντας τον μηχανισμό (HP Wolf Security, 2025).

Το reCAPTCHA v3 αν και βασίζεται αποκλειστικά στην ανάλυση της συμπεριφοράς του χρήστη (π.χ. κίνηση ποντικιού, χρόνος απόκρισης) και θεωρείται πιο εξελιγμένο, εντούτοις η ΤΝ μπορεί να μιμηθεί την ανθρώπινη αλληλεπίδραση με υψηλό βαθμό ακρίβειας, καθιστώντας τον μηχανισμό ευάλωτο σε bots με behavioral AI (Shib Daily, 2024).

Τέλος, το Audio CAPTCHA το οποίο απαιτεί την αναγνώριση λέξεων από ηχητικό αρχείο, μπορεί επίσης να παραβιαστεί καθώς η ΤΝ αξιοποιεί τεχνικές NLP και speech-to-text για να μετατρέπει τον ήχο σε κείμενο, παρακάμπτοντας τη δοκιμασία με υψηλό ποσοστό επιτυχίας (HP Wolf Security, 2025).

Πίνακας 3.3: Τύποι CAPTCHA και Χαρακτηριστικά

Τύπος CAPTCHA	**Περιγραφή**	**Βασική Αδυναμία απέναντι στην ΤΝ**
Text-based CAPTCHA	Αναγνώριση παραμορφωμένων χαρακτήρων	OCR και CNN μπορούν να αναγνωρίσουν το κείμενο
Image-based CAPTCHA	Επιλογή εικόνων με κοινό χαρακτηριστικό	Ανάλυση εικόνας μέσω YOLO και ResNet
reCAPTCHA v2	Checkbox “I’m not a robot” + εικόνες κυκλοφορίας	Προσομοίωση συμπεριφοράς και αναγνώριση εικόνας
reCAPTCHA v3	Ανάλυση συμπεριφοράς χρήστη χωρίς οπτική δοκιμασία	Μίμηση ανθρώπινης αλληλεπίδρασης από bots
Audio CAPTCHA	Αναγνώριση λέξεων από ηχητικό αρχείο	Φωνητική ανάλυση μέσω NLP και speech-to-text

Αξίζει να σημειωθεί πως τα πιο πρόσφατα συστήματα επαλήθευσης, όπως το reCAPTCHA v3, βασίζονται σε ανάλυση συμπεριφοράς αντί για οπτικές δοκιμασίες. Ωστόσο, η ΤΝ μπορεί να προσομοιώσει την ανθρώπινη αλληλεπίδραση με ιστοσελίδες, όπως η κίνηση του ποντικιού, ο χρόνος απόκρισης και η επιλογή στοιχείων. Αυτό επιτρέπει στα bots να ξεγελούν τα συστήματα, αποκτώντας πρόσβαση χωρίς να εντοπίζονται (Shib Daily, 2024).

Χαρακτηριστικό παράδειγμα αποτελεί η επίθεση σε πλατφόρμα ηλεκτρονικών κρατήσεων το 2022, όπου bot με ενσωματωμένο behavioral AI κατάφερε να παρακάμψει το reCAPTCHA v3, πραγματοποιώντας μαζικές κρατήσεις εισιτηρίων χωρίς ανθρώπινη παρέμβαση (Dinh & Ogiela, 2022).

Επιπλέον, η χρήση VPN και proxy servers σε συνδυασμό με τα εργαλεία της ΤΝ ενισχύουν την ανωνυμία των επιτιθέμενων. Τα bots μπορούν να παρακάμπτουν γεωγραφικούς περιορισμούς και να αποφεύγουν την ανίχνευση, αυξάνοντας την αποτελεσματικότητα των

επιθέσεων. Για τον λόγο αυτό η Dinh και Ogiela (2022) προτείνουν την ανάπτυξη γνωσιακών CAPTCHA που βασίζονται σε φυσική αλληλεπίδραση, ως μέσο ενίσχυσης της ασφάλειας.

Παράλληλα, σε μια άλλη έρευνα που έγινε από την HP Wolf Security (2025) εντοπίστηκαν κακόβουλες εκστρατείες όπου επιτιθέμενοι δημιουργούσαν ψεύτικα CAPTCHA για να παραπλανήσουν χρήστες και να εγκαταστήσουν κακόβουλο λογισμικό όπως το Lumma Stealer, μέσω αυτοματοποιημένων εντολών PowerShell. Το περιστατικό καταγράφηκε σε ελληνική ιστοσελίδα τεχνολογίας, όπου χρήστες παραπλανήθηκαν από ψευδές CAPTCHA που οδηγούσε σε λήψη αρχείου .exe (HP Wolf Security, 2025).

Αξίζει να σημειωθεί ότι τα ποσοστά επιτυχίας των παραβιάσεων, όπως αποτυπώνονται και στον πίνακα 3.4 είναι σημαντικά υψηλά.

Πίνακας 3.4: Ποσοστά Παραβίασης CAPTCHA από Τεχνητή Νοημοσύνη

CAPTCHA Σύστημα	**Πηγή Μελέτης**	**Ποσοστό Επιτυχούς Παραβίασης από AI**
Google reCAPTCHA v2	ETH Zurich (Plesner & ETH Zurich, 2024)	100%
Text-based CAPTCHA	Goodfellow et al. (2016)	96%
Image-based CAPTCHA	Shib Daily (2024)	85%
reCAPTCHA v3	Dinh & Ogiela (2022)	74%
Audio CAPTCHA	Internal NLP benchmarks	88%

Η ελληνική επιστημονική κοινότητα παρουσιάζει έντονη δραστηριότητα στον τομέα της τεχνητής νοημοσύνης καθώς σύμφωνα με το Εθνικό Κέντρο Τεκμηρίωσης (2023), έχουν καταγραφεί πάνω από 4.000 ελληνικές επιστημονικές δημοσιεύσεις σχετικές με την ΤΝ, πολλές από τις οποίες αφορούν εφαρμογές ασφαλείας και επαλήθευσης ταυτότητας. Αν και δεν εστιάζουν αποκλειστικά στα CAPTCHA, αναδεικνύουν τη δυναμική της χώρας στον τομέα της κυβερνοασφάλειας. Ενδεικτικά, το ΕΚΠΑ και το ΑΠΘ έχουν δημοσιεύσει μελέτες για την ανάπτυξη CAPTCHA με βάση την ελληνική γλώσσα και τη χρήση NLP για την ενίσχυση της ασφάλειας (Εθνικό Κέντρο Τεκμηρίωσης, 2023).

Σε ότι αφορά τις επιπτώσεις η παραβίαση CAPTCHA υπονομεύει την αποτελεσματικότητα των συστημάτων ασφαλείας, (Dinh & Ogiela, 2022), οδηγεί σε απώλεια εμπιστοσύνης των χρηστών προς τις ψηφιακές πλατφόρμες (Davies, 2023), διαταράσσει την κοινωνική συνοχή (HP Wolf Security, 2025) και εγείρει ζητήματα λογοδοσίας και προστασίας προσωπικών δεδομένων (Chesney & Citron, 2019).

3.4 Η Χρήση της Τεχνητής Νοημοσύνης για την Ενίσχυση Επιθέσεων Τύπου Zero-Day

Οι επιθέσεις τύπου zero-day αποτελούν μία από τις πιο επικίνδυνες μορφές κυβερνοαπειλών, καθώς εκμεταλλεύονται ευπάθειες που δεν έχουν ακόμη εντοπιστεί ή διορθωθεί από τους κατασκευαστές λογισμικού. Οι επιθέσεις zero-day βασίζονται στην εκμετάλλευση ευπαθειών που δεν έχουν ακόμη δημοσιοποιηθεί ή διορθωθεί. Αυτό καθιστά τις επιθέσεις ιδιαίτερα επικίνδυνες, καθώς δεν υπάρχουν διαθέσιμες ενημερώσεις ή μέτρα προστασίας κατά την εκδήλωσή τους (Roumani, 2021).

Η τεχνητή νοημοσύνη περιλαμβάνει τεχνικές όπως η μηχανική μάθηση (ML), η βαθιά μάθηση (DL) και η επεξεργασία φυσικής γλώσσας (NLP), οι οποίες επιτρέπουν την ανάλυση μεγάλων ποσοτήτων δεδομένων και την εξαγωγή μοτίβων (Brundage et al., 2018). Επιπλέον, η ΤΝ μπορεί να αναλύσει πηγαίο κώδικα και να εντοπίσει ευπάθειες που δεν είναι ακόμη γνωστές, ενώ η χρήση deep learning επιτρέπει την εκπαίδευση μοντέλων που προβλέπουν ευπάθειες με βάση ιστορικά δεδομένα και χαρακτηριστικά του κώδικα (Zhang, Wang & Chen 2020). Επίσης η ΤΝ μπορεί να συμβάλει στην αυτοματοποιημένη δημιουργία κώδικα exploit. Οι Sadeghian, Zhang και Amin (2019) παρουσιάζουν τεχνικές semantic analysis που επιτρέπουν την κατανόηση της λειτουργίας του λογισμικού και την παραγωγή επιθέσεων που εκμεταλλεύονται συγκεκριμένες αδυναμίες.

Η ανάπτυξη εργαλείων βασισμένων στην ΤΝ έχει επιταχύνει τη διαδικασία εντοπισμού και εκμετάλλευσης ευπαθειών. Ο πίνακας 3.5 παρουσιάζει μια σειρά από εξειδικευμένα εργαλεία τεχνητής νοημοσύνης που έχουν αναπτυχθεί ή αξιοποιηθεί για την εκτέλεση επιθέσεων τύπου zero-day έτσι όπως αυτά καταγράφηκαν από τους μελετητές. Οι επιθέσεις αυτές χαρακτηρίζονται από την εκμετάλλευση άγνωστων ή μη διορθωμένων ευπαθειών, γεγονός που τις καθιστά εξαιρετικά επικίνδυνες, καθώς δεν υπάρχουν διαθέσιμα μέτρα προστασίας κατά την εκδήλωσή τους (Roumani, 2021).

- Το εργαλείο VulDeePecker αξιοποιεί Bi-LSTM (Bidirectional Long Short-Term Memory), μια μορφή βαθιάς μάθησης, για την ανίχνευση ευπαθειών σε πηγαίο

κώδικα. Το εργαλείο μπορεί να εντοπίσει μοτίβα που υποδηλώνουν πιθανές αδυναμίες, ακόμα και σε μεγάλα και πολύπλοκα αποθετήρια λογισμικού (Li et al., 2018).

- Το εργαλείο DeepExploit ενσωματώνει τεχνικές ενισχυτικής μάθησης (Reinforcement Learning) για την αυτοματοποιημένη εκμετάλλευση ευπαθειών. Το σύστημα μαθαίνει πώς να επιλέγει και να εκτελεί επιθέσεις με βάση την ανταπόκριση του στόχου, μειώνοντας την ανάγκη για ανθρώπινη παρέμβαση (Yamato et al., 2019).
- Το εργαλείο SemFuzz συνδυάζει σημασιολογική ανάλυση με τεχνικές fuzzing για την παραγωγή εισόδων που μπορούν να αποκαλύψουν σφάλματα ή ευπάθειες. Η χρήση ΤΝ επιτρέπει την πιο στοχευμένη και αποδοτική δημιουργία δοκιμαστικών δεδομένων (Sadeghian et al., 2019).
- Το εργαλείο CodeXray βασίζεται σε Graph Neural Networks (GNNs) για την ανάλυση της δομής του λογισμικού. Το εργαλείο μπορεί να εντοπίσει σύνθετες σχέσεις μεταξύ στοιχείων του κώδικα, αποκαλύπτοντας zero-day ευπάθειες που δεν είναι άμεσα ορατές μέσω συμβατικών μεθόδων (Zhang et al., 2020).
- Τέλος το εργαλείο GenAI Phishing αν και δεν στοχεύει άμεσα σε τεχνικές ευπάθειες, το εργαλείο αυτό χρησιμοποιεί Generative AI για τη δημιουργία εξατομικευμένων phishing επιθέσεων. Αυτές μπορούν να λειτουργήσουν ως προθάλαμος για την εκμετάλλευση zero-day ευπαθειών, ειδικά όταν συνδυάζονται με κοινωνική μηχανική και παραπλανητικά σενάρια (Euronews Next, 2024).

Σύμφωνα με τους μελετητές η παρουσία αυτών των εργαλείων καταδεικνύει τη μετατόπιση των επιθέσεων zero-day από χειροκίνητες και χρονοβόρες διαδικασίες σε πλήρως αυτοματοποιημένα και ευφυή συστήματα. Η τεχνητή νοημοσύνη δεν λειτουργεί απλώς ως επιταχυντής, αλλά ως στρατηγικός πολλαπλασιαστής της αποτελεσματικότητας και της εμβέλειας των επιθέσεων.

Πίνακας 3.5: Εργαλεία ΤΝ για Επιθέσεις Zero-Day

Εργαλείο	**Τεχνολογία ΤΝ**	**Σκοπός Χρήσης**	**Πηγή**
VulDeePecker	Bi-LSTM (Deep	Ανίχνευση ευπαθειών σε	Li et al., 2018

	Learning)	πηγαίο κώδικα	
DeepExploit	Reinforcement Learning	Αυτοματοποιημένη εκμετάλλευση ευπαθειών	Yamato et al., 2019
SemFuzz	Semantic Analysis + Fuzzing	Παραγωγή εισόδων για εντοπισμό σφαλμάτων	Sadeghian et al., 2019
CodeXray	Graph Neural Networks	Ανάλυση δομής λογισμικού για zero-day ευπάθειες	Zhang et al., 2020
GenAI Phishing	Generative AI	Δημιουργία εξατομικευμένων επιθέσεων phishing	Euronews Next, 2024

Οι επιθέσεις τύπου zero-day που αξιοποιούν τεχνητή νοημοσύνη έχουν πολυδιάστατες επιπτώσεις. Οι πιο σημαντικές είναι οι αιφνίδιες παραβιάσεις κρίσιμων συστημάτων (Roumani, 2021; Yamato et al., 2019) που μπορεί να προκαλέσουν οικονομικές απώλειες λόγω διακοπής λειτουργίας, κόστους αποκατάστασης και απώλειας πελατών, (Zhang et al., 2020; Sadeghian et al., 2019), η καλλιέργεια φόβου και άγχους στους χρήστες και διαχειριστές των συστημάτων (Davies, 2023; Brundage et al., 2018), να οδηγήσουν τους χρήστες να είναι επιφυλακτικοί στις τεχνολογικές υποδομές (Chesney & Citron, 2019) αλλά και στην απονομή δικαιοσύνης καθώς ο εντοπισμός των υπευθύνων είναι δύσκολος (Li et al., 2018; Euronews Next, 2024).

3.5 Δημιουργία προσαρμοστικού κακόβουλου λογισμικού με τη χρήση τεχνητής νοημοσύνης

Η τεχνητή νοημοσύνη (ΤΝ) έχει μετατραπεί σε ισχυρό εργαλείο για τους κυβερνοεγκληματίες, επιτρέποντας την ανάπτυξη επιθέσεων που είναι πιο ευέλικτες, δυναμικές και δύσκολες στην ανίχνευση. Το προσαρμοστικό κακόβουλο λογισμικό (adaptive malware) αποτελεί χαρακτηριστικό παράδειγμα αυτής της εξέλιξης (πίνακας 3.6).

Σύμφωνα με τους Ullah, Ahmad και Khan (2022), το adaptive malware χρησιμοποιεί αλγορίθμους μηχανικής μάθησης για να μεταβάλλει τη συμπεριφορά του σε πραγματικό

χρόνο, ανάλογα με το περιβάλλον του στόχου. Μπορεί να αποκρύπτει τον κώδικά του, να αλλάζει τα μοτίβα επικοινωνίας και να παρακάμπτει μηχανισμούς ασφαλείας. Αυτό το καθιστά ιδιαίτερα αποτελεσματικό σε επιθέσεις που απαιτούν παρατεταμένη παραμονή στο σύστημα χωρίς να εντοπίζονται.

Οι Fritsch, Jaber και Yazidi (2023) επισημαίνουν ότι η ΤΝ επιτρέπει την ανάπτυξη malware με δυνατότητα αυτο-εξέλιξης, το οποίο μπορεί να προσαρμόζεται στις άμυνες του στόχου και να χρησιμοποιεί τεχνικές ευφυούς απόκρυψης. Έτσι, οι επιθέσεις γίνονται πιο επίμονες και ανθεκτικές απέναντι σε παραδοσιακά εργαλεία ανίχνευσης.

Επιπλέον, οι Song et al. (2025) τονίζουν ότι οι πολυμορφικές και προσαρμοστικές απειλές που δημιουργούνται με ΤΝ αποτελούν σοβαρή πρόκληση, καθώς μπορούν να μεταβάλλουν συνεχώς την υπογραφή τους και να ξεφεύγουν από τα συστήματα που βασίζονται σε στατικά μοτίβα. Αυτό σημαίνει ότι οι επιθέσεις γίνονται πιο απρόβλεπτες και δύσκολα αντιμετωπίσιμες.

Τέλος, η μελέτη του Ramaswamy (2024) δείχνει ότι οι επιθέσεις που αξιοποιούν ΤΝ δεν περιορίζονται σε απλή απόκρυψη, αλλά μπορούν να χρησιμοποιούν σύνθετες στρατηγικές πολλαπλών επιπέδων, συνδυάζοντας τεχνικές μηχανικής μάθησης, βαθιάς μάθησης και αλγορίθμους γραφημάτων για να ενισχύσουν την αποτελεσματικότητα τους.

Πίνακας 3.6: Εργαλεία ΤΝ για επιθέσεις με προσαρμοστικό κακόβουλο λογισμικό

Συγγραφείς / Έτος	**Εστίαση**	**Ευρήματα για Επιθέσεις**
Ullah, Ahmad & Khan (2022)	Adaptive malware	Δυναμική τροποποίηση συμπεριφοράς, απόκρυψη κώδικα, αλλαγή επικοινωνίας, αποφυγή ανίχνευσης
Fritsch, Jaber & Yazidi (2023)	AI-driven malware	Αυτο-εξέλιξη, ευφυής απόκρυψη, προσαρμογή στις άμυνες του στόχου
Song et al. (2025)	Πολυμορφικές απειλές	Συνεχής μεταβολή υπογραφών, δυσκολία εντοπισμού από στατικά συστήματα

Ramaswamy (2024)	Σύνθετες επιθέσεις	Χρήση ML, DL και γραφημάτων για πολυεπίπεδες στρατηγικές επίθεσης

Σε ότι αφορά τις επιπτώσεις, η εφαρμογή της Τεχνητής Νοημοσύνης (ΤΝ) στο προσαρμοστικό κακόβουλο λογισμικό έχει μετασχηματίσει τις κυβερνοεπιθέσεις, καθιστώντας τες πιο περίπλοκες, ευέλικτες και δύσκολα ανιχνεύσιμες (Ullah, Ahmad, & Khan, 2022; Fritsch, Jaber, & Yazidi, 2023).

Η δυναμική αλλαγή υπογραφής και η χρήση βαθιάς μάθησης ενισχύουν την ανθεκτικότητα των απειλών (Song et al., 2025), ενώ οι πολυεπίπεδες στρατηγικές που προκύπτουν οδηγούν σε σοβαρές λειτουργικές και οικονομικές συνέπειες για τους οργανισμούς, εγείροντας παράλληλα κρίσιμα νομικά και ηθικά ζητήματα επεκτείνοντας τις επιπτώσεις πέρα από το τεχνικό πεδίο (Ramaswamy, 2024).

3.6 Deepfakes και Παραπληροφόρηση με χρήση της Τεχνητής Νοημοσύνης

Η τεχνολογία των deepfakes, βασισμένη σε αλγορίθμους βαθιάς μάθησης (deep learning), επιτρέπει την παραγωγή εξαιρετικά ρεαλιστικών ψευδών βίντεο και ηχητικών αρχείων. Αν και μπορεί να αξιοποιηθεί για δημιουργικούς σκοπούς, η επιθετική χρήση της έχει προκαλέσει ανησυχία σε παγκόσμιο επίπεδο, καθώς οι έρευνες δείχνουν πως αποτελούν ισχυρό εργαλείο κοινωνικής μηχανικής, ικανό να προκαλέσει παραπληροφόρηση, εκβιασμό και πολιτική αποσταθεροποίηση (Chesney & Citron, 2019).

Τα deepfakes είναι ψηφιακά παραποιημένα αρχεία εικόνας, βίντεο ή ήχου, που δημιουργούνται μέσω νευρωνικών δικτύων, όπως τα GANs (Generative Adversarial Networks). Η τεχνολογία αυτή επιτρέπει την αναπαραγωγή προσώπων, φωνών και κινήσεων με τέτοια ακρίβεια, ώστε να καθίσταται δύσκολη η διάκριση από το αυθεντικό υλικό (Chesney & Citron, 2019).

Η επιθετική χρήση των deepfakes (πίνακας 3.7) περιλαμβάνει την εσκεμμένη δημιουργία ψευδών περιεχομένων με στόχο την εξαπάτηση, τη χειραγώγηση της κοινής γνώμης ή την πρόκληση κοινωνικής αναταραχής, καθώς η δυνατότητα τους να αποδοθεί ψευδής συμπεριφορά ή λόγος σε δημόσια πρόσωπα καθιστά τα deepfakes εργαλείο πολιτικής και ψυχολογικής επίθεσης (Citron & Chesney, 2019; Kietzmann et al., 2020; Paris & Donovan, 2019; Westerlund, 2019).

Πίνακας 3.7: Μορφές Επιθετικής Χρήσης Deepfakes

Εφαρμογή	**Σκοπός Επίθεσης**	**Παράδειγμα**	**Πηγή**
Πολιτική Παραπληροφόρηση	Αποσταθεροποίηση, χειραγώγηση εκλογών	Ψευδές βίντεο πολιτικού σε σκάνδαλο	Chesney & Citron, 2019
Εκβιασμός και δυσφήμιση	Ψυχολογική πίεση, οικονομική απειλή	Deepfake πορνογραφία δημοσιογράφων	ERTNews, 2024
Εταιρική κατασκοπεία	Παραπλάνηση στελεχών, οικονομική ζημία	Ψευδές ηχητικό μήνυμα CEO	Kietzmann et al., 2020
Προπαγάνδα και τρομοκρατία	Υποκίνηση βίας, κοινωνική αναταραχή	Βίντεο στρατιωτικών σε «ιεροσυλία»	Paris & Donovan, 2019

Η εσκεμμένη χρήση deepfakes για παραπληροφόρηση, εκβιασμό και πολιτική αποσταθεροποίηση έχει καταγραφεί σε πολλές μελέτες που αποδεικνύουν την επικινδυνότητα τους (Citron & Chesney, 2019; Kietzmann et al., 2020; Paris & Donovan, 2019; Vaccari & Chadwick, 2020; Westerlund, 2019), αναδεικνύοντας τον ρόλο τους στη διάβρωση της εμπιστοσύνης στα μέσα ενημέρωσης και στην ενίσχυση γεωπολιτικών εντάσεων (Fernández Gambín et al., 2024; Shoaib et al., 2023).

Στην Ελλάδα, περιστατικά deepfakes έχουν ήδη προκαλέσει ανησυχία. Δημοσιογράφοι εμφανίστηκαν σε παραποιημένα βίντεο που προωθούσαν προϊόντα ή τζόγο (ΕΡΤ, 2024). Η Δίωξη Ηλεκτρονικού Εγκλήματος έχει καταγράψει περιπτώσεις sextortion, ενώ η ΕΕΤΤ (2023) αναφέρει αύξηση 37% στη χρήση deepfakes σε fake news και πολιτικές εκστρατείες. Επιπλέον, μελέτη του Πανεπιστημίου Πειραιώς (2024) έδειξε ότι ψευδές βίντεο επηρέασε την πρόθεση ψήφου σε τοπικές εκλογές (πίνακας 3.8).

Πίνακας 3.8: Ερευνητικά ευρήματα

Πηγή / Έρευνα	**Εστίαση**	**Κύρια Ευρήματα**
Vaccari & Chadwick (2020)	Πολιτική παραπληροφόρηση	Επηρεασμός εκλογών μέσω ψευδών βίντεο
Paris & Donovan (2019)	Γεωπολιτική χρήση deepfakes	Στρατηγική παραπλάνηση σε διεθνές επίπεδο
Kietzmann et al. (2020)	Επιχειρησιακή εξαπάτηση	Οικονομικές απώλειες μέσω ψευδών ηχητικών
ΕΕΤΤ (2023)	Εθνική παραπληροφόρηση	Αύξηση χρήσης deepfakes σε fake news στην Ελλάδα
Δαμαλίτης (2024)	Εκλογική χειραγώγηση	Μεταβολή πρόθεσης ψήφου λόγω ψευδών δηλώσεων

Σε ότι αφορά τις επιπτώσεις που επιφέρουν αυτές είναι οικονομικές, καθώς από τις έρευνες έχει διαπιστωθεί ότι έχουν την δυνατότητα να βλάψουν το κύρος και την ευρωστία μιας επιχείρησης, ηθικές και ψυχολογικές καθώς υπάρχει κίνδυνος να χαθεί η εμπιστοσύνη των πολιτών στην ψηφιακή τεχνολογία, να αλλάξουν ακόμη και εκλογικά αποτελέσματα χειραγωγώντας τους ψηφοφόρους, αλλά και να προκαλέσουν εθνικές και διεθνείς εντάσεις γεωπολιτικής φύσεως (Citron & Chesney, 2019; Kietzmann et al., 2020; Paris & Donovan, 2019; Vaccari & Chadwick, 2020; Westerlund, 2019).

3.7 Εξαγωγή δεδομένων μέσω AI-driven social engineering με χρήση της Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη (ΤΝ) έχει αναδειχθεί σε καταλύτη για την εξέλιξη της κοινωνικής μηχανικής, καθώς επιτρέπει την αυτοματοποιημένη ανάλυση δημόσιων προφίλ και επικοινωνιών με στόχο την εξαγωγή προσωπικών δεδομένων. Μέσω τεχνικών επεξεργασίας φυσικής γλώσσας (NLP) και προγνωστικής ανάλυσης, η ΤΝ μπορεί να εντοπίσει ενδιαφέροντα, σχέσεις και ευάλωτα σημεία, διευκολύνοντας την κατασκευή πειστικών σεναρίων εξαπάτησης (Borges et al., 2022; Tsiatsos & Karyda, 2023).

Τα ευρήματα των μελετητών (πίνακας 3.9) δείχνουν ότι η χρήση γενετικών μοντέλων περιεχομένου (π.χ. LLMs, GANs) επιτρέπει την παραγωγή εξατομικευμένων μηνυμάτων, email, φωνητικών κλήσεων ή ακόμη και οπτικού υλικού που μιμείται το ύφος και τις συνήθειες του στόχου (Kietzmann et al., 2020; Vaccari & Chadwick, 2020). Αυτό οδηγεί σε νέες μορφές spear-phishing, smishing και vishing, όπου η εξαπάτηση αποκτά υψηλή πειστικότητα και δυσκολία ανίχνευσης (Shoaib et al., 2023).

Άλλες, σύγχρονες μελέτες δείχνουν ότι η ΤΝ μπορεί να συνδυάσει δεδομένα από κοινωνικά δίκτυα, εταιρικές ιστοσελίδες και δημόσιες βάσεις, δημιουργώντας ολοκληρωμένα προφίλ που χρησιμοποιούνται για εξατομικευμένες επιθέσεις. Η κλιμάκωση και η αυτοματοποίηση αυτών των πρακτικών καθιστούν την άμυνα πιο δύσκολη, καθώς οι οργανισμοί καλούνται να αντιμετωπίσουν επιθέσεις που μοιάζουν «χειροποίητες» αλλά παράγονται μαζικά (Fernández Gambín et al., 2024).

Πίνακας 3.9: Ευρήματα για επιθέσεις AI-driven Social Engineering με τη χρήση ΤΝ

Πηγή (Συγγραφέας & Έτος)	**Εστίαση**	**Ευρήματα**
Borges, Silva & Almeida (2022)	AI-driven social engineering και ανάλυση δημόσιων προφίλ	Η ΤΝ αυτοματοποιεί την εξαγωγή προσωπικών δεδομένων και διευκολύνει την κατασκευή πειστικών σεναρίων εξαπάτησης
Tsiatsos & Karyda (2023)	ΤΝ και κοινωνική μηχανική	NLP και προγνωστική ανάλυση εντοπίζουν ευάλωτα σημεία και σχέσεις, αυξάνοντας την αποτελεσματικότητα επιθέσεων
Kietzmann et al. (2020)	Χρήση γενετικών μοντέλων (GANs, LLMs)	Παραγωγή εξατομικευμένων μηνυμάτων και περιεχομένου που μιμείται το ύφος του στόχου
Vaccari & Chadwick (2020)	Παραπληροφόρηση μέσω AI	Επίδραση στην εμπιστοσύνη στα ΜΜΕ και πολιτική σταθερότητα μέσω εξατομικευμένων ψευδών περιεχομένων

Shoaib et al. (2023)	Frontier AI και εξαπάτηση	Νέες μορφές spear-phishing, smishing και vishing με υψηλή πειστικότητα και δυσκολία ανίχνευσης
Fernández Gambín et al. (2024)	Τάσεις και μελλοντική πορεία AI deception	Η ΤΝ συνδυάζει δεδομένα από κοινωνικά δίκτυα και δημόσιες βάσεις, δημιουργώντας ολοκληρωμένα προφίλ για εξατομικευμένες επιθέσεις

Οι επιθέσεις κοινωνικής μηχανικής που υποστηρίζονται από τεχνητή νοημοσύνη έχουν σημαντικές επιπτώσεις σε πολλαπλά επίπεδα. Σε οργανωσιακό επίπεδο, αυξάνουν την πιθανότητα παραβίασης δεδομένων και οικονομικών απωλειών, καθώς η αυτοματοποιημένη ανάλυση δημόσιων προφίλ και επικοινωνιών οδηγεί σε πιο πειστικά και εξατομικευμένα σενάρια εξαπάτησης (Borges et al., 2022; Tsiatsos & Karyda, 2023).

Σε κοινωνικό επίπεδο, ενισχύουν την παραπληροφόρηση και τη διάβρωση της εμπιστοσύνης στα μέσα ενημέρωσης, καθώς τα γενετικά μοντέλα περιεχομένου μπορούν να παράγουν ρεαλιστικά ψευδή μηνύματα και οπτικοακουστικό υλικό (Kietzmann et al., 2020; Vaccari & Chadwick, 2020).

Επιπλέον, η κλιμάκωση και η αυτοματοποίηση αυτών των επιθέσεων καθιστούν την ανίχνευση δυσκολότερη, οδηγώντας σε νέες μορφές spear-phishing, smishing και vishing που απειλούν τόσο την κυβερνοασφάλεια όσο και την πολιτική σταθερότητα (Shoaib et al., 2023; Fernández Gambín et al., 2024).

3.8 Ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών

Η αξιοποίηση της τεχνητής νοημοσύνης (ΤΝ) στις επιθέσεις κοινωνικής μηχανικής δεν περιορίζεται πλέον στη δημιουργία ψευδών περιεχομένων, επεκτείνεται και στην εντοπισμένη εκμετάλλευση ψυχολογικών και συναισθηματικών αδυναμιών. Μέσα από προηγμένα εργαλεία ανάλυσης φυσικής γλώσσας και πολυτροπικά μοντέλα, οι επιτιθέμενοι μπορούν να ανιχνεύσουν σημάδια κόπωσης, άγχους ή γνωστικής υπερφόρτωσης, επιλέγοντας το

κατάλληλο χρονικό σημείο για να εξαπολύσουν την επίθεση (Borges et al., 2022; Tsiatsos & Karyda, 2023).

Η χρήση γενετικών μοντέλων, όπως τα LLMs και τα GANs, επιτρέπει την παραγωγή εξατομικευμένων μηνυμάτων και πολυμεσικού υλικού που προσαρμόζονται στο ύφος και τις συνήθειες του στόχου. Έτσι, οι επιθέσεις αποκτούν μεγαλύτερη πειστικότητα και οδηγούν σε εξελιγμένες μορφές spear-phishing, smishing και vishing, οι οποίες είναι δύσκολο να ανιχνευθούν με παραδοσιακά μέσα (Kietzmann et al., 2020; Vaccari & Chadwick, 2020; Shoaib et al., 2023).

Πρόσφατες μελέτες (πίνακας 3.10) δείχνουν ότι η ΤΝ μπορεί να συνδυάζει δεδομένα από κοινωνικά δίκτυα, εταιρικές ιστοσελίδες και δημόσιες βάσεις, δημιουργώντας ολιστικά προφίλ στόχων. Αυτά τα προφίλ χρησιμοποιούνται για εξατομικευμένες επιθέσεις μεγάλης κλίμακας, οι οποίες μοιάζουν «χειροποίητες» αλλά παράγονται μαζικά, καθιστώντας την άμυνα των οργανισμών ιδιαίτερα δύσκολη (Fernández Gambín et al., 2024).

Συνολικά, όπως δείχνουν τα ευρήματα των μελετητών η ΤΝ λειτουργεί ως καταλύτης για την ενίσχυση της κοινωνικής μηχανικής, μετατρέποντας τις ανθρώπινες αδυναμίες σε σημεία εισόδου για επιθέσεις υψηλής αποτελεσματικότητας και δυσκολίας ανίχνευσης (Borges et al., 2022; Fernández Gambín et al., 2024; Kietzmann et al., 2020; Shoaib et al., 2023; Tsiatsos & Karyda, 2023; Vaccari & Chadwick, 2020).

Πίνακας 3.10: Ευρήματα μελετών για την χρήση ΤΝ στην ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών

Πηγή (Συγγραφέας & Έτος)	**Εστίαση**	**Ευρήματα**
Borges, Silva & Almeida (2022)	AI-driven social engineering και ανάλυση δημόσιων προφίλ	Η ΤΝ αυτοματοποιεί την εξαγωγή προσωπικών δεδομένων και εντοπίζει στιγμές ευαλωτότητας (κόπωση, άγχος) για εξατομικευμένες επιθέσεις
Tsiatsos & Karyda (2023)	ΤΝ και κοινωνική μηχανική	NLP και προγνωστική ανάλυση εντοπίζουν σχέσεις και αδυναμίες, αυξάνοντας την αποτελεσματικότητα των

		επιθέσεων
Kietzmann et al. (2020)	Χρήση γενετικών μοντέλων (GANs, LLMs)	Παραγωγή εξατομικευμένων μηνυμάτων και πολυμεσικού υλικού που μιμούνται το ύφος του στόχου
Vaccari & Chadwick (2020)	Παραπληροφόρηση μέσω AI	Ενίσχυση της διάβρωσης εμπιστοσύνης στα ΜΜΕ και πολιτική αποσταθεροποίηση μέσω ψευδών περιεχομένων
Shoaib et al. (2023)	Frontier AI και εξαπάτηση	Νέες μορφές spear-phishing, smishing και vishing με υψηλή πειστικότητα και δυσκολία ανίχνευσης
Fernández Gambín et al. (2024)	Τάσεις και μελλοντική πορεία AI deception	Η ΤΝ συνδυάζει δεδομένα από κοινωνικά δίκτυα και δημόσιες βάσεις, δημιουργώντας ολοκληρωμένα προφίλ για εξατομικευμένες επιθέσεις μεγάλης κλίμακας

Οι επιθέσεις που αξιοποιούν την τεχνητή νοημοσύνη για να εντοπίσουν στιγμές ευαλωτότητας του χρήστη έχουν σοβαρές συνέπειες σε πολλαπλά επίπεδα. Σε οργανωσιακό πλαίσιο, αυξάνουν την πιθανότητα παραβίασης δεδομένων και οικονομικών απωλειών, καθώς οι επιτιθέμενοι εκμεταλλεύονται καταστάσεις κόπωσης ή άγχους για να επιτύχουν μεγαλύτερη συμμόρφωση σε αιτήματα αποκάλυψης πληροφοριών (Borges et al., 2022; Tsiatsos & Karyda, 2023).

Σε κοινωνικό επίπεδο, ενισχύουν την παραπληροφόρηση και τη διάβρωση της εμπιστοσύνης στα μέσα ενημέρωσης, καθώς τα γενετικά μοντέλα περιεχομένου παράγουν πειστικά και εξατομικευμένα ψευδή μηνύματα (Kietzmann et al., 2020; Vaccari & Chadwick, 2020).

Επιπλέον, η κλιμάκωση και η αυτοματοποίηση αυτών των επιθέσεων καθιστούν την ανίχνευση δυσκολότερη, οδηγώντας σε νέες μορφές spear-phishing, smishing και vishing που απειλούν τόσο την κυβερνοασφάλεια όσο και την πολιτική σταθερότητα (Shoaib et al., 2023; Fernández Gambín et al., 2024).

3.9 Adversarial attacks σε συστήματα ΤΝ

Αν και τα ερευνητικά δεδομένα δείχνουν ότι η τεχνητή νοημοσύνη χρησιμοποιείται ευρέως για τις κυβερνοεπιθέσεις, ωστόσο υπόκειται και η ίδια σε επιθέσεις. Οι adversarial attacks αποτελούν μια από τις πιο κρίσιμες απειλές για τα συστήματα τεχνητής νοημοσύνης, καθώς βασίζονται στη δημιουργία παραπλανητικών δεδομένων εισόδου που οδηγούν τα μοντέλα σε λανθασμένες αποφάσεις. Η θεμελιώδης εργασία των Goodfellow, Shlens και Szegedy (2014) ανέδειξε ότι ακόμη και μικρές, αόρατες για τον άνθρωπο διαταραχές μπορούν να προκαλέσουν σοβαρά σφάλματα ταξινόμησης, εγκαινιάζοντας την έννοια των adversarial examples.

Στη συνέχεια, η έρευνα ανέδειξε διαφορετικούς τύπους επιθέσεων που διακρίνονται σε white-box, όπου ο αντίπαλος έχει πλήρη γνώση του μοντέλου, και σε black-box, όπου η επίθεση βασίζεται σε μεταφερσιμότητα παραδειγμάτων από surrogate μοντέλα (Qiu et al., 2019). Οι επιθέσεις μπορεί να είναι στοχευμένες, οδηγώντας το μοντέλο σε συγκεκριμένη εσφαλμένη κλάση, ή μη-στοχευμένες, όπου οποιοδήποτε λάθος είναι αποδεκτό (Wang et al., 2023).

Συνολικά, οι adversarial attacks αποδεικνύουν ότι η ΤΝ μπορεί να χειραγωγηθεί με ελάχιστες αλλά στρατηγικές παρεμβάσεις στα δεδομένα εισόδου, γεγονός που καθιστά την αξιοπιστία των συστημάτων ιδιαίτερα ευάλωτη σε κακόβουλους αντιπάλους. Οι μελέτες γύρω από τις adversarial attacks καταδεικνύουν ότι ακόμη και ελάχιστες διαταραχές στα δεδομένα εισόδου μπορούν να οδηγήσουν μοντέλα τεχνητής νοημοσύνης σε σοβαρά σφάλματα ταξινόμησης, γεγονός που αποδεικνύει τη δομική τους ευαλωτότητα (Goodfellow et al., 2014).

Η έρευνα έχει δείξει ότι οι επιθέσεις μπορούν να πραγματοποιηθούν τόσο σε white-box περιβάλλοντα, όπου ο αντίπαλος έχει πλήρη γνώση του μοντέλου, όσο και σε black-box σενάρια, αξιοποιώντας τη μεταφερσιμότητα adversarial παραδειγμάτων μεταξύ διαφορετικών αρχιτεκτονικών (Qiu et al., 2019).

Επιπλέον, οι μελέτες (πίνακας 3.11) ανέδειξαν ότι οι επιθέσεις δεν περιορίζονται στην όραση υπολογιστών αλλά επεκτείνονται σε φωνή, κείμενο και δίκτυα, αποδεικνύοντας την πολυτροπική φύση του φαινομένου (Abomakhelb et al., 2023). Σημαντικό εύρημα είναι ότι οι

επιθέσεις μπορούν να είναι στοχευμένες, οδηγώντας το μοντέλο σε συγκεκριμένη εσφαλμένη κλάση, ή μη-στοχευμένες, όπου οποιοδήποτε λάθος είναι αποδεκτό, με διαφορετικά ποσοστά επιτυχίας και κόστους (Wang et al., 2023).

Τέλος, πρόσφατες ανασκοπήσεις επιβεβαιώνουν ότι η μεταφερσιμότητα των adversarial παραδειγμάτων αποτελεί κρίσιμο χαρακτηριστικό, επιτρέποντας σε αντιπάλους να παρακάμπτουν πολλαπλά συστήματα με κοινά μοτίβα ευαλωτότητας (Ahmed et al., 2024).

Πίνακας 3.11: Ευρήματα μελετών για τις επιθέσεις Adversarial attacks

Πηγή (Συγγραφέας & Έτος)	**Εστίαση**	**Κυριότερα Ευρήματα**
Goodfellow, Shlens & Szegedy (2014)	Θεμελίωση adversarial examples	Μικρές διαταραχές στα δεδομένα εισόδου μπορούν να προκαλέσουν σοβαρά σφάλματα ταξινόμησης σε νευρωνικά δίκτυα
Qiu et al. (2019)	Ταξινόμηση επιθέσεων	Διαχωρισμός σε white-box και black-box επιθέσεις· ανάδειξη της μεταφερσιμότητας παραδειγμάτων μεταξύ μοντέλων
Wang et al. (2023)	Επιθέσεις σε δίκτυα και ML-powered συστήματα	Στοχευμένες και μη-στοχευμένες επιθέσεις επηρεάζουν την αξιοπιστία συστημάτων επικοινωνίας και κατανομής πόρων
Abomakhelb et al. (2023)	Πολυτροπικές επιθέσεις	Adversarial παραδείγματα σε εικόνα, φωνή και κείμενο· επέκταση πέρα από την όραση υπολογιστών
Ahmed et al. (2024)	Μεταφερσιμότητα και πρακτική αξιοποίηση	Adversarial παραδείγματα μπορούν να παρακάμπτουν διαφορετικές αρχιτεκτονικές, επιτρέποντας black-box επιθέσεις με surrogate μοντέλα

Οι adversarial attacks έχουν σημαντικές επιπτώσεις στην αξιοπιστία και ασφάλεια των συστημάτων τεχνητής νοημοσύνης. Η δυνατότητα των επιτιθέμενων να δημιουργούν παραπλανητικά δεδομένα εισόδου οδηγεί σε λανθασμένες αποφάσεις ταξινόμησης, γεγονός που υπονομεύει την εμπιστοσύνη στη χρήση της ΤΝ σε κρίσιμες εφαρμογές (Goodfellow et al., 2014).

Οι μελέτες δείχνουν ότι οι επιθέσεις μπορούν να εφαρμοστούν σε διαφορετικά πεδία, από την αναγνώριση εικόνας και φωνής έως τα δίκτυα επικοινωνίας, επηρεάζοντας την ακρίβεια και τη λειτουργικότητα των συστημάτων (Qiu et al., 2019; Abomakhelb et al., 2023). Επιπλέον, η μεταφερσιμότητα των adversarial παραδειγμάτων καθιστά δυνατή την εκμετάλλευση πολλαπλών μοντέλων με κοινά μοτίβα ευαλωτότητας, αυξάνοντας τον κίνδυνο μαζικών επιθέσεων (Ahmed et al., 2024). Σε περιβάλλοντα υψηλής κρισιμότητας, όπως η υγεία, η αυτόνομη οδήγηση και οι τηλεπικοινωνίες, οι επιθέσεις αυτές μπορούν να οδηγήσουν σε σοβαρές λειτουργικές και κοινωνικές συνέπειες, καθιστώντας την αντιμετώπισή τους ζήτημα στρατηγικής σημασίας (Wang et al., 2023).

3.10 Επιθέσεις σε αυτόνομα συστήματα και IoT

Η ενσωμάτωση τεχνητής νοημοσύνης σε αυτόνομα συστήματα και δίκτυα IoT έχει διευρύνει σημαντικά την επιφάνεια απειλής, δημιουργώντας νέες ευκαιρίες για επιθέσεις που μπορούν να προκαλέσουν τόσο ψηφιακή όσο και φυσική ζημιά. Σύμφωνα με τους Radanliev et al. (2024), η ετερογένεια των IoT συσκευών και η περιορισμένη υπολογιστική ισχύς τους καθιστούν τα συστήματα αυτά ιδιαίτερα ευάλωτα σε επιθέσεις που στοχεύουν την ιδιωτικότητα και την ακεραιότητα των αισθητήρων.

Από την πλευρά τους οι Khazane et al. (2024) επισημαίνουν ότι οι επιθέσεις αυτές μπορούν να επηρεάσουν κρίσιμες λειτουργίες όπως η δρομολόγηση, η αντίληψη και η ανίχνευση ανωμαλιών, οδηγώντας σε λειτουργική αστάθεια.

Σε αυτόνομα οχήματα και drones, οι επιθέσεις συχνά εκδηλώνονται μέσω παραποίησης σημάτων GPS και IMU, καθώς και μέσω χειραγώγησης των μοντέλων αντίληψης, με αποτέλεσμα εσφαλμένες αποφάσεις πλοήγησης (IEEE, 2025). Οι Xing et al. (2025) αναλύοντας τις ευπάθειες των embodied AI, υπογραμμίζοντας ότι η παραπλάνηση της αντίληψης μπορεί να οδηγήσει σε επικίνδυνες ενέργειες, όπως συγκρούσεις ή απώλεια ελέγχου. Παράλληλα, οι Radanliev et al. (2024) τονίζουν ότι η συλλογή και συσχέτιση δεδομένων από IoT συσκευές μπορεί να οδηγήσει σε παραβίαση της ιδιωτικότητας, με δυνατότητα ανασύνθεσης ευαίσθητων πληροφοριών και επιτήρησης σε μαζική κλίμακα.

Συνολικά, η βιβλιογραφία (πίνακας 3.12) συγκλίνει στο ότι οι επιθέσεις σε αυτόνομα συστήματα και IoT δεν είναι απλώς τεχνικές προκλήσεις, αλλά ενσαρκώνουν μια πολυδιάστατη απειλή που αγγίζει τη φυσική ασφάλεια, τη λειτουργική διαθεσιμότητα και την κοινωνική δεοντολογία. Η ανάγκη για ολιστική προσέγγιση στην ασφάλεια αυτών των συστημάτων είναι πλέον επιτακτική (Xing et al., 2025)

Πίνακας 3.12: Ευρήματα μελετών για τις επιθέσεις σε αυτόνομα συστήματα και IoT

Πηγή (Συγγραφέας & Έτος)	**Εστίαση**	**Κυριότερα Ευρήματα**
Radanliev et al. (2024)	Ασφάλεια AI και κυβερνοκίνδυνοι σε IoT	Η ετερογένεια και οι περιορισμένοι πόροι των IoT συσκευών αυξάνουν τον κίνδυνο επιθέσεων σε αισθητήρες και ιδιωτικότητα
Khazane et al. (2024)	Adversarial επιθέσεις σε IoT δίκτυα	Οι επιθέσεις μπορούν να επηρεάσουν κρίσιμες λειτουργίες όπως δρομολόγηση, αντίληψη και ανίχνευση ανωμαλιών
IEEE (2025)	Προκλήσεις ασφάλειας σε AVs και drones	Ευπάθειες σε GPS/IMU και κανάλια επικοινωνίας· χειραγώγηση perception μοντέλων οδηγεί σε εσφαλμένες αποφάσεις πλοήγησης
Xing et al. (2025)	Ευπάθειες σε embodied AI	Sensor spoofing και adversarial perception μπορούν να προκαλέσουν απώλεια ελέγχου ή συγκρούσεις σε ρομπότ και αυτόνομα οχήματα
Radanliev et al. (2024, arXiv)	AI security σε IoT (προδημοσίευση)	Συλλογή και συσχέτιση δεδομένων IoT οδηγεί σε παραβίαση ιδιωτικότητας και κλιμακούμενη επιτήρηση

Οι επιθέσεις σε αυτόνομα συστήματα και δίκτυα IoT έχουν πολυδιάστατες συνέπειες που επηρεάζουν τόσο τη φυσική ασφάλεια όσο και την κοινωνική εμπιστοσύνη. Η παραπλάνηση αισθητήρων μέσω μπορεί να οδηγήσει σε εσφαλμένες αποφάσεις πλοήγησης σε drones και αυτόνομα οχήματα, προκαλώντας συγκρούσεις ή απώλεια ελέγχου (IEEE, 2025; Xing et al., 2025).

Παράλληλα, η χειραγώγηση δικτυακών καναλιών και η εισαγωγή κακόβουλων εντολών σε IoT συσκευές μπορεί να διαταράξει κρίσιμες υποδομές, μειώνοντας τη λειτουργική διαθεσιμότητα και την αξιοπιστία υπηρεσιών σε έξυπνες πόλεις και βιομηχανικά περιβάλλοντα (Khazane et al., 2024).

Επιπλέον, η εκτεταμένη συλλογή και συσχέτιση δεδομένων από IoT συσκευές ενισχύει τον κίνδυνο παραβίασης ιδιωτικότητας και επιτήρησης σε μαζική κλίμακα, με κοινωνικές και δεοντολογικές προεκτάσεις (Radanliev et al., 2024).

3.11 Παραποίηση νομικών και οικονομικών εγγράφων

Η χρήση γενετικών μοντέλων (π.χ. GANs, LLMs) έχει αναδείξει νέες μορφές επιθέσεων που στοχεύουν στην παραποίηση νομικών και οικονομικών εγγράφων. Οι επιθέσεις αυτές δεν περιορίζονται σε απλή πλαστογραφία, αλλά αξιοποιούν την ικανότητα των μοντέλων να παράγουν κείμενα και μορφοποιήσεις που μοιάζουν αυθεντικά, καθιστώντας δύσκολη την ανίχνευσή τους από ανθρώπους ή παραδοσιακά συστήματα ελέγχου (Kumar et al., 2023).

Σύμφωνα με τους Kietzmann et al. (2023), η ΤΝ μπορεί να δημιουργήσει ψευδή συμβόλαια, οικονομικές καταστάσεις ή φορολογικά έγγραφα, τα οποία ενσωματώνουν ρεαλιστικά στοιχεία και γλωσσικά μοτίβα, διευκολύνοντας επιθέσεις απάτης. Οι επιθέσεις αυτές εντάσσονται στο πλαίσιο της AI-enabled forgery, όπου η παραπλάνηση βασίζεται στη γλωσσική και μορφολογική πιστότητα των παραγόμενων εγγράφων.

Επιπλέον, οι επιθέσεις σε οικονομικά συστήματα μέσω παραποιημένων τιμολογίων ή τραπεζικών εγγράφων μπορούν να οδηγήσουν σε οικονομικές απώλειες μεγάλης κλίμακας, καθώς οι παραγόμενες πλαστογραφίες είναι δύσκολο να διακριθούν από τα γνήσια έγγραφα (Zhang et al., 2024). Σε νομικό επίπεδο, η δημιουργία ψευδών συμβολαίων ή δικαστικών εγγράφων μπορεί να υπονομεύσει την αξιοπιστία θεσμών και να προκαλέσει παραβίαση εμπιστοσύνης (Floridi, 2023).

Η βιβλιογραφία (πίνακας 3.13) συγκλίνει ότι οι επιθέσεις αυτές αποτελούν μια νέα κατηγορία γενετικών απειλών, όπου η ΤΝ χρησιμοποιείται όχι μόνο για κυβερνοεπιθέσεις αλλά και για

κοινωνικο-οικονομική χειραγώγηση, με άμεσες επιπτώσεις στην ασφάλεια, την οικονομία και τη δικαιοσύνη.

Συγκεκριμένα, η μελέτη των Kumar, Singh και Patel (2023) η οποία επικεντρώθηκε στη χρήση Generative Adversarial Networks (GANs) για την πλαστογραφία εγγράφων, έδειξε ότι τα παραγόμενα κείμενα και μορφοποιήσεις έχουν τόσο υψηλή πιστότητα που καθιστούν δύσκολη την ανίχνευση πλαστογραφίας από παραδοσιακά συστήματα ελέγχου.

Οι Zhang, Li και Chen (2024) εξετάζοντας την εφαρμογή της ΤΝ σε οικονομικές απάτες, διαπίστωσαν ότι υπάρχει κίνδυνος παραποίησης τιμολογίων και τραπεζικών εγγράφων τα οποία μπορούν να οδηγήσουν σε απώλειες μεγάλης κλίμακας, καθώς οι παραγόμενες πλαστογραφίες είναι δύσκολο να διακριθούν από τα γνήσια.

Στην δική του μελέτη, ο Floridi (2023) εστίασε στις νομικές συνέπειες, υπογραμμίζοντας ότι η δημιουργία ψευδών συμβολαίων και δικαστικών εγγράφων υπονομεύει την εμπιστοσύνη στους θεσμούς και θέτει σε κίνδυνο τη διαφάνεια του νομικού συστήματος.

Από την πλευρά τους, στην δική τους μελέτη οι Kietzmann, Paschen και Treen (2023) αναλύοντας την έννοια της AI-enabled forgery, επισημαίνουν ότι η ΤΝ μπορεί να χρησιμοποιηθεί για κοινωνικο-οικονομική χειραγώγηση, ενισχύοντας την απάτη και την παραπληροφόρηση σε κρίσιμους τομείς, με σοβαρές δεοντολογικές προεκτάσεις.

Πίνακας 3.13: Ερευνητικά ευρήματα για παραποίηση Εγγράφων με ΤΝ

Πηγή (Συγγραφέας & Έτος)	**Εστίαση**	**Κυριότερα Ευρήματα**
Kumar et al. (2023)	Χρήση GANs για πλαστογραφία εγγράφων	Τα παραγόμενα έγγραφα έχουν υψηλή γλωσσική και μορφολογική πιστότητα, δυσκολεύοντας την ανίχνευση πλαστογραφίας
Zhang et al. (2024)	Οικονομική απάτη μέσω παραποιημένων εγγράφων	Παραποιημένα τιμολόγια και τραπεζικά έγγραφα μπορούν να οδηγήσουν σε οικονομικές απώλειες μεγάλης κλίμακας

Floridi (2023)	Επιπτώσεις στην εμπιστοσύνη νομικών θεσμών	Ψευδή συμβόλαια και δικαστικά έγγραφα υπονομεύουν την αξιοπιστία θεσμών και τη διαφάνεια
Kietzmann et al. (2023)	AI-enabled forgery και κοινωνικο-οικονομικές συνέπειες	Η ΤΝ μπορεί να χρησιμοποιηθεί για ενίσχυση απάτης και παραπληροφόρησης, με κοινωνικές και δεοντολογικές προεκτάσεις

Οι επιθέσεις που βασίζονται στην παραποίηση νομικών και οικονομικών εγγράφων με χρήση γενετικών μοντέλων έχουν σοβαρές συνέπειες σε πολλαπλά επίπεδα. Σε οικονομικό πλαίσιο, η δημιουργία ψευδών τιμολογίων, τραπεζικών εγγράφων ή οικονομικών καταστάσεων μπορεί να οδηγήσει σε απάτες μεγάλης κλίμακας και σημαντικές οικονομικές απώλειες για οργανισμούς και ιδιώτες (Zhang et al., 2024).

Στο νομικό πεδίο, η παραγωγή πλαστών συμβολαίων ή δικαστικών εγγράφων υπονομεύει την αξιοπιστία των θεσμών και απειλεί τη διαφάνεια του δικαστικού συστήματος (Floridi, 2023). Παράλληλα, η δυσκολία ανίχνευσης τέτοιων πλαστογραφιών, λόγω της υψηλής γλωσσικής και μορφολογικής πιστότητας που προσφέρουν τα γενετικά μοντέλα, καθιστά τις επιθέσεις αυτές ιδιαίτερα επικίνδυνες (Kumar et al., 2023).

Τέλος, σε κοινωνικό επίπεδο, η χρήση της ΤΝ για παραποίηση εγγράφων ενισχύει την απάτη και την παραπληροφόρηση, οδηγώντας σε απώλεια εμπιστοσύνης και σε κίνδυνο κοινωνικο-οικονομικής χειραγώγησης (Kietzmann et al., 2023).

3.12 Επιπτώσεις από Μορφές Επιθέσεων με Χρήση ΤΝ

Η συστηματική μελέτη των μορφών επιθέσεων με χρήση Τεχνητής Νοημοσύνης αναδεικνύει ότι κάθε κατηγορία διαθέτει διαφορετικά χαρακτηριστικά και προκαλεί ποικίλες επιπτώσεις σε τεχνικό, οικονομικό, νομικό και κοινωνικό επίπεδο. Η πολυπλοκότητα και η ποικιλία των συνεπειών, παρουσιάζονται στον πίνακα 3.14 στον οποίον συνοψίζονται οι βασικές επιπτώσεις κάθε μορφής επίθεσης.

Πίνακας 3.14: Επιπτώσεις από Μορφές Επιθέσεων με Χρήση ΤΝ

Μορφή Επίθεσης	**Επιπτώσεις**	**Ενδεικτικές Πηγές**
1. Αυτοματοποιημένες επιθέσεις phishing	Απώλεια οικονομικών πόρων, παραβίαση προσωπικών δεδομένων, ενίσχυση κυβερνοεγκλήματος	Bose & Leung (2023)
2.Παραβίαση CAPTCHA και συστημάτων επαλήθευσης	Μη εξουσιοδοτημένη πρόσβαση σε πλατφόρμες και υπηρεσίες, υπονόμευση μηχανισμών ασφαλείας	Yan et al. (2022)
3. Ενίσχυση επιθέσεων zero-day	Ταχύτερη ανακάλυψη και εκμετάλλευση ευπαθειών, αύξηση κλίμακας κυβερνοεπιθέσεων	Shen et al. (2023)
4.Προσαρμοστικό κακόβουλο λογισμικό	Δυσκολία ανίχνευσης, αυξημένη ανθεκτικότητα επιθέσεων, παρατεταμένη παραμονή σε συστήματα	Demetrio et al. (2021)
5.Deepfakes και παραπληροφόρηση	Υπονόμευση εμπιστοσύνης στην πληροφόρηση, πολιτικές/κοινωνικές συνέπειες, διάδοση ψευδών ειδήσεων	Chesney & Citron (2019)
6.AI-driven social engineering	Παραβίαση ιδιωτικότητας, εξαπάτηση χρηστών, απώλεια εμπιστοσύνης σε ψηφιακές υπηρεσίες	Hadnagy (2021)
7.Εκμετάλλευση ανθρώπινων αδυναμιών	Ψυχολογική χειραγώγηση, κοινωνική εκμετάλλευση, αύξηση κινδύνων εξαπάτησης	West (2020)

8.Adversarial attacks σε AI	Παραπλάνηση μοντέλων αντίληψης, κίνδυνοι σε αυτόνομα οχήματα και ιατρική διάγνωση	Goodfellow et al. (2014); Qiu et al. (2019)
9.Επιθέσεις σε αυτόνομα συστήματα και IoT	Απώλεια ελέγχου, φυσικές ζημιές, παραβίαση ιδιωτικότητας, λειτουργική αστάθεια	Radanliev et al. (2024); Xing et al. (2025)
10.Παραποίηση νομικών και οικονομικών εγγράφων	Οικονομικές απώλειες, υπονόμευση θεσμών, κοινωνικο-οικονομική χειραγώγηση	Kumar et al. (2023); Zhang et al. (2024); Floridi (2023)

3.13 Επιπτώσεις ανά κατηγορία

1. Αυτοματοποιημένες επιθέσεις phishing

Η αξιοποίηση τεχνητής νοημοσύνης για την αυτοματοποίηση επιθέσεων phishing ενισχύει σημαντικά την αποτελεσματικότητα αυτών των πρακτικών, καθώς επιτρέπει τη δημιουργία εξατομικευμένων και ιδιαίτερα πειστικών μηνυμάτων. Η στοχευμένη φύση των επιθέσεων αυξάνει την πιθανότητα εξαπάτησης των χρηστών, οδηγώντας σε απώλεια οικονομικών πόρων και παραβίαση ευαίσθητων προσωπικών δεδομένων. Παράλληλα, η συστηματική χρήση τέτοιων τεχνικών συμβάλλει στην ενίσχυση του κυβερνοεγκλήματος, καθώς διευκολύνει την κλιμάκωση των επιθέσεων και υπονομεύει την εμπιστοσύνη στις ψηφιακές επικοινωνίες (Bose & Leung, 2023).

2. Παραβίαση CAPTCHA και συστημάτων επαλήθευσης

Η αξιοποίηση προηγμένων μοντέλων αναγνώρισης εικόνας και κειμένου έχει καταστήσει δυνατή την παράκαμψη συστημάτων CAPTCHA και άλλων μηχανισμών επαλήθευσης. Η πρακτική αυτή εντάσσεται στις σύγχρονες μορφές επιθέσεων, καθώς επιτρέπει σε κακόβουλους χρήστες να αποκτούν πρόσβαση σε διαδικτυακές πλατφόρμες χωρίς εξουσιοδότηση. Οι συνέπειες περιλαμβάνουν την ενίσχυση αυτοματοποιημένων επιθέσεων, την αύξηση περιστατικών phishing και την ανεξέλεγκτη συλλογή δεδομένων, γεγονός που υπονομεύει την ασφάλεια και την εμπιστοσύνη των χρηστών στις ψηφιακές υπηρεσίες (Yan et al., 2022).

3. Ενίσχυση επιθέσεων τύπου zero-day

Η αξιοποίηση τεχνικών τεχνητής νοημοσύνης για την ανάλυση μεγάλων συνόλων δεδομένων ενισχύει σημαντικά τις επιθέσεις τύπου zero-day. Μέσω της ταχείας αναγνώρισης άγνωστων ευπαθειών, οι επιτιθέμενοι αποκτούν τη δυνατότητα να επιταχύνουν την εκμετάλλευση νέων κενών ασφαλείας πριν υπάρξουν διαθέσιμες διορθώσεις. Οι συνέπειες αυτής της πρακτικής είναι η αύξηση της κλίμακας και της ταχύτητας των κυβερνοεπιθέσεων, γεγονός που καθιστά δυσκολότερη την έγκαιρη ανίχνευση και αντιμετώπισή τους, ενώ παράλληλα εντείνει τον κίνδυνο για κρίσιμες υποδομές και υπηρεσίες (Shen et al., 2023).

4. Δημιουργία προσαρμοστικού κακόβουλου λογισμικού

Η αξιοποίηση γενετικών μοντέλων για τη δημιουργία προσαρμοστικού κακόβουλου λογισμικού ενισχύει σημαντικά την αποτελεσματικότητα των επιθέσεων. Μέσω της δυναμικής τροποποίησης του κώδικα, το λογισμικό αποκτά την ικανότητα να μεταβάλλεται συνεχώς ώστε να παρακάμπτει μηχανισμούς ανίχνευσης και άμυνας. Αυτό έχει ως αποτέλεσμα την αυξημένη ανθεκτικότητα των επιθέσεων, τη δυσκολία εντοπισμού τους από παραδοσιακά συστήματα ασφαλείας και την επιμήκυνση του χρόνου παραμονής τους σε κρίσιμες υποδομές. Συνεπώς, η προσαρμοστικότητα αυτή καθιστά τις επιθέσεις πιο επίμονες και επικίνδυνες, εντείνοντας τον κίνδυνο για την ακεραιότητα των ψηφιακών υπηρεσιών (Demetrio et al., 2021).

5. Deepfakes και παραπληροφόρηση

Η παραγωγή παραπλανητικών οπτικοακουστικών δεδομένων με τη χρήση τεχνητής νοημοσύνης συνιστά μια ιδιαίτερα επικίνδυνη μορφή επίθεσης, καθώς υπονομεύει την αξιοπιστία της πληροφόρησης και ενισχύει την εξάπλωση της παραπληροφόρησης. Οι συνέπειες δεν περιορίζονται μόνο στην παραπλάνηση των χρηστών, αλλά επεκτείνονται σε πολιτικό και κοινωνικό επίπεδο, δημιουργώντας συνθήκες χειραγώγησης της κοινής γνώμης και αποσταθεροποίησης θεσμών. Η δυνατότητα δημιουργίας ρεαλιστικών αλλά ψευδών περιεχομένων καθιστά δυσχερή την ανίχνευση και την αντιμετώπιση τέτοιων επιθέσεων, αυξάνοντας τον κίνδυνο για την ασφάλεια της πληροφορίας και την εμπιστοσύνη στις ψηφιακές πλατφόρμες (Chesney & Citron, 2019).

6. Εξαγωγή δεδομένων μέσω AI-driven social engineering

Η ενσωμάτωση τεχνητής νοημοσύνης σε τεχνικές κοινωνικής μηχανικής ενισχύει σημαντικά την αποτελεσματικότητα των επιθέσεων, καθώς επιτρέπει τη δημιουργία

εξατομικευμένων σεναρίων εξαπάτησης που προσαρμόζονται στα χαρακτηριστικά και τις συνήθειες των χρηστών. Η στοχευμένη αυτή προσέγγιση αυξάνει την πιθανότητα επιτυχίας των επιθέσεων, οδηγώντας σε παραβίαση της ιδιωτικότητας και σε μη εξουσιοδοτημένη πρόσβαση σε ευαίσθητα δεδομένα. Παράλληλα, η συστηματική χρήση τέτοιων μεθόδων υπονομεύει την εμπιστοσύνη των χρηστών στις ψηφιακές υπηρεσίες και εντείνει τον κίνδυνο ευρείας κλίμακας παραπλάνησης και εκμετάλλευσης (Hadnagy, 2021).

7. Ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών

Η εφαρμογή τεχνητής νοημοσύνης στην ανάλυση συμπεριφορικών δεδομένων δημιουργεί νέες δυνατότητες για την ανίχνευση ψυχολογικών ευπαθειών των χρηστών. Η στοχευμένη αξιοποίηση αυτών των πληροφοριών μπορεί να οδηγήσει σε χειραγώγηση και εκμετάλλευση, καθώς οι επιτιθέμενοι αποκτούν πρόσβαση σε λεπτομερή προφίλ που αποκαλύπτουν αδυναμίες και μοτίβα συμπεριφοράς. Οι επιπτώσεις περιλαμβάνουν την ενίσχυση επιθέσεων κοινωνικής μηχανικής, την αύξηση της αποτελεσματικότητας εξατομικευμένων στρατηγικών παραπλάνησης και την υπονόμευση της αυτονομίας των χρηστών. Συνεπώς, η χρήση της ΤΝ σε αυτό το πλαίσιο εντείνει τον κίνδυνο για την ιδιωτικότητα και την ασφάλεια των ψηφιακών αλληλεπιδράσεων (West, 2020).

8. Adversarial attacks σε συστήματα AI

Οι επιθέσεις τύπου adversarial αξιοποιούν μικρές αλλά στρατηγικά σχεδιασμένες διαταραχές στα δεδομένα εισόδου, οι οποίες έχουν τη δυνατότητα να παραπλανήσουν ακόμη και εξελιγμένα μοντέλα τεχνητής νοημοσύνης. Η πρακτική αυτή εντάσσεται στις πιο κρίσιμες μορφές επιθέσεων, καθώς μπορεί να οδηγήσει σε λανθασμένες προβλέψεις ή αποφάσεις σε εφαρμογές υψηλής σημασίας, όπως η ιατρική διάγνωση και τα αυτόνομα οχήματα. Οι συνέπειες περιλαμβάνουν την αύξηση του κινδύνου για την ασφάλεια των χρηστών, την υπονόμευση της αξιοπιστίας των συστημάτων και τη δημιουργία σοβαρών προκλήσεων για την ανίχνευση και την αντιμετώπιση τέτοιων επιθέσεων (Goodfellow et al., 2014; Qiu et al., 2019).

9. Επιθέσεις σε αυτόνομα συστήματα και IoT

Η παραποίηση δεδομένων αισθητήρων και η χειραγώγηση δικτυακών καναλιών σε αυτόνομα συστήματα και συσκευές IoT συνιστούν ιδιαίτερα επικίνδυνες μορφές επιθέσεων. Μέσα από τέτοιες πρακτικές, οι επιτιθέμενοι μπορούν να προκαλέσουν απώλεια ελέγχου των συστημάτων, οδηγώντας σε δυσλειτουργίες που ενδέχεται να έχουν

άμεσες φυσικές συνέπειες, όπως ζημιές σε υποδομές ή κινδύνους για την ασφάλεια των χρηστών. Παράλληλα, η εκμετάλλευση των δικτυακών καναλιών επιτρέπει την παραβίαση της ιδιωτικότητας και την υποκλοπή ευαίσθητων δεδομένων, γεγονός που υπονομεύει την αξιοπιστία των ψηφιακών υπηρεσιών και εντείνει την τρωτότητα των δικτυωμένων περιβαλλόντων (Radanliev et al., 2024; Xing et al., 2025).

10. Παραποίηση νομικών και οικονομικών εγγράφων

Η χρήση γενετικών μοντέλων για τη δημιουργία ψευδών νομικών και οικονομικών εγγράφων συνιστά μια ιδιαίτερα επικίνδυνη μορφή επίθεσης, καθώς υπονομεύει την αξιοπιστία θεσμικών διαδικασιών και δημιουργεί σοβαρούς κινδύνους για την κοινωνικο-οικονομική σταθερότητα. Μέσα από την παραγωγή πλαστών εγγράφων, οι επιτιθέμενοι μπορούν να προκαλέσουν άμεσες οικονομικές απώλειες, να χειραγωγήσουν αγορές ή να αλλοιώσουν νομικές διαδικασίες. Οι συνέπειες εκτείνονται πέρα από το οικονομικό πεδίο, καθώς η συστηματική παραποίηση εγγράφων μειώνει την εμπιστοσύνη των πολιτών στους θεσμούς και ενισχύει φαινόμενα κοινωνικής αποσταθεροποίησης. Συνεπώς, η πρακτική αυτή εντείνει τον κίνδυνο θεσμικής διάβρωσης και ευρείας κλίμακας χειραγώγησης (Kumar et al., 2023; Zhang et al., 2024; Floridi, 2023).

3.14 Επίλογος

Η παρουσίαση των ευρημάτων ανέδειξε ότι η τεχνητή νοημοσύνη έχει μετασχηματίσει ριζικά το τοπίο των κυβερνοεπιθέσεων, προσδίδοντάς τους μεγαλύτερη πολυπλοκότητα, εξατομίκευση και αποτελεσματικότητα. Από τις αυτοματοποιημένες επιθέσεις phishing και την παραβίαση CAPTCHA έως την ενίσχυση των zero-day exploits, η ΤΝ αποδείχθηκε ότι λειτουργεί ως καταλύτης για την ανάπτυξη νέων και πιο εξελιγμένων μορφών απειλής (Brundage et al., 2018; Zhou et al., 2022; Goodfellow et al., 2016).

Τα αποτελέσματα καταδεικνύουν ότι οι επιπτώσεις δεν περιορίζονται στο τεχνικό επίπεδο, αλλά επεκτείνονται στον κοινωνικό, ψυχολογικό και οικονομικό τομέα, επηρεάζοντας την εμπιστοσύνη, την ιδιωτικότητα και τη συνοχή των ψηφιακών κοινοτήτων (Chesney & Citron, 2019; Davies, 2023). Επιπλέον, διαφάνηκε πως η πολυπλοκότητα των επιθέσεων και η δυνατότητα προσαρμογής τους σε πραγματικό χρόνο απαιτούν νέες εφαρμογές και ότι η αντιμετώπισή τους απαιτεί νέες στρατηγικές και καινοτόμες προσεγγίσεις.

Συνολικά, το Κεφάλαιο 3 προσφέρει μια ολοκληρωμένη εικόνα για το πώς η ΤΝ αξιοποιείται σε επιθετικά σενάρια, θέτοντας τις βάσεις για την επόμενη ενότητα, όπου θα εξεταστούν οι τρόποι αντιμετώπισης και οι προοπτικές ενίσχυσης της κυβερνοασφάλειας.

Κεφάλαιο 4ο: Σύνοψη ευρημάτων

4.1 Εισαγωγή

Η συστηματική ανάλυση που προηγήθηκε ανέδειξε το εύρος και την πολυπλοκότητα των επιθετικών εφαρμογών της τεχνητής νοημοσύνης. Στο Κεφάλαιο 4 παρουσιάζεται μια συνοπτική αποτύπωση των βασικών ευρημάτων, με στόχο να αναδειχθούν οι κύριες μορφές επιθέσεων και οι επιπτώσεις τους σε τεχνικό, κοινωνικό, οικονομικό και θεσμικό επίπεδο.

Η ενότητα αυτή συγκεντρώνει τα αποτελέσματα που καταγράφηκαν στα προηγούμενα κεφάλαια, προσφέροντας μια ολοκληρωμένη εικόνα για το πώς η ΤΝ μετασχηματίζει το τοπίο της κυβερνοασφάλειας. Από τις αυτοματοποιημένες επιθέσεις phishing και την παραβίαση CAPTCHA έως τις zero-day επιθέσεις, το προσαρμοστικό κακόβουλο λογισμικό και τα deepfakes, η ΤΝ αποδεικνύεται καταλύτης για την ανάπτυξη εξελιγμένων και πολυδιάστατων απειλών.

Παράλληλα, η σύνοψη αναδεικνύει ότι οι επιπτώσεις δεν περιορίζονται στο τεχνικό επίπεδο, αλλά επεκτείνονται σε κοινωνικές και ψυχολογικές διαστάσεις, επηρεάζοντας την εμπιστοσύνη, την ιδιωτικότητα και τη σταθερότητα των θεσμών. Με αυτόν τον τρόπο, το Κεφάλαιο 4 λειτουργεί ως γέφυρα ανάμεσα στην παρουσίαση των ευρημάτων και στη συζήτηση που θα ακολουθήσει, θέτοντας το πλαίσιο για την ερμηνεία και την αξιολόγηση των αποτελεσμάτων.

4.2 Ευρήματα

Η ραγδαία εξέλιξη της τεχνητής νοημοσύνης έχει δημιουργήσει νέες προοπτικές αλλά και σοβαρές προκλήσεις στον τομέα της κυβερνοασφάλειας. Ενώ η τεχνητή νοημοσύνη (ΤΝ) προσφέρει σημαντικά πλεονεκτήματα σε επίπεδο αυτοματοποίησης, ανάλυσης δεδομένων και βελτίωσης υπηρεσιών, παράλληλα έχει καταστεί εργαλείο στα χέρια κυβερνοεγκληματιών, οι οποίοι αξιοποιούν τις δυνατότητές της για την ανάπτυξη πιο εξελιγμένων και δύσκολα ανιχνεύσιμων επιθέσεων. Η βιβλιογραφία αναδεικνύει ότι οι επιθέσεις που βασίζονται στην ΤΝ δεν περιορίζονται σε τεχνικές παραβιάσεις, αλλά επεκτείνονται σε κοινωνικές, οικονομικές και θεσμικές διαστάσεις, υπονομεύοντας την εμπιστοσύνη και τη σταθερότητα σε πολλαπλά επίπεδα (Stylianou et al., 2025).

4.2.1 Αυτοματοποιημένες Επιθέσεις Phishing μέσω Τεχνητής Νοημοσύνης

Οι επιθέσεις phishing έχουν εξελιχθεί σημαντικά με την αξιοποίηση της τεχνητής νοημοσύνης, η οποία τις καθιστά πιο εξατομικευμένες, πειστικές και δύσκολα ανιχνεύσιμες. Μέσω NLP, τα μηνύματα προσαρμόζονται στο ύφος και τα ενδιαφέροντα του στόχου (Brundage et al., 2018; Kumar & Garg, 2023), ενώ η μαζική παραγωγή emails με φυσικό χαρακτήρα και πολυγλωσσική προσαρμογή αυξάνει την αποτελεσματικότητα (Eze & Shamir, 2024). Επιπλέον, η χρήση φωνητικής σύνθεσης και deepfakes ενισχύει την εξαπάτηση (Eskandari, 2022; Dami, 2022), ενώ η προσαρμογή σε πραγματικό χρόνο δημιουργεί δυναμικά σενάρια εξαπάτησης (Williams, 2025). Τέλος, Telegram bots και μεταφραστές διευκολύνουν την αυτοματοποίηση και διεθνοποίηση των επιθέσεων (Altukhova, 2023).

Μελέτες δείχνουν ότι τα AI-generated phishing μηνύματα έχουν έως και 40% υψηλότερο ποσοστό επιτυχίας από τα ανθρώπινα (Zhou et al., 2022), ενώ η TN μπορεί να εντοπίσει ψυχολογικά μοτίβα και να προσαρμόσει δυναμικά το περιεχόμενο (Borgaonkar et al., 2023). Οι επιπτώσεις είναι πολυδιάστατες. Οδηγούν σε απώλεια δεδομένων και χρημάτων, αύξηση άγχους και μείωση εργασιακής ικανοποίησης, ενίσχυση παραπληροφόρησης και κοινωνικής δυσπιστίας, καθώς και υπονόμευση της ιδιωτικότητας (Davies, 2023; Stylianou et al., 2025; Brundage et al., 2018).

4.2.2 Παραβίαση CAPTCHA και Συστημάτων Επαλήθευσης μέσω Τεχνητής Νοημοσύνης

Τα CAPTCHA έχουν σχεδιαστεί για να διαχωρίζουν ανθρώπινους χρήστες από αυτοματοποιημένα προγράμματα, όμως η εξέλιξη της τεχνητής νοημοσύνης έχει καταστήσει τα συστήματα αυτά ιδιαίτερα ευάλωτα. Νευρωνικά δίκτυα και τεχνικές OCR μπορούν να αναγνωρίζουν παραμορφωμένους χαρακτήρες (Goodfellow, Bengio, & Courville, 2016), ενώ μοντέλα όπως YOLO και ResNet παρακάμπτουν image-based CAPTCHA με υψηλή ακρίβεια (Plesner & ETH Zurich, 2024; HP Wolf Security, 2025). Ακόμη και πιο εξελιγμένες εκδοχές, όπως το reCAPTCHA v3 και τα Audio CAPTCHA, μπορούν να παραβιαστούν μέσω behavioral AI και speech-to-text τεχνικών (Shib Daily, 2024; HP Wolf Security, 2025).

Επιπλέον, επιθέσεις με VPN και proxy servers ενισχύουν την ανωνυμία των δραστών, ενώ έχουν καταγραφεί κακόβουλες εκστρατείες με ψεύτικα CAPTCHA που οδηγούν σε εγκατάσταση malware (HP Wolf Security, 2025). Οι επιπτώσεις είναι πολυδιάστατες: υπονομεύεται η αποτελεσματικότητα των μηχανισμών ασφαλείας (Dinh & Ogiela, 2022), μειώνεται η εμπιστοσύνη των χρηστών στις ψηφιακές πλατφόρμες (Davies, 2023),

διαταράσσεται η κοινωνική συνοχή και εγείρονται ζητήματα λογοδοσίας και προστασίας προσωπικών δεδομένων (Chesney & Citron, 2019).

Συνολικά, η ΤΝ έχει καταστήσει τα CAPTCHA ευάλωτα σε πολλαπλές μορφές παραβίασης, με συνέπειες που επεκτείνονται πέρα από την κυβερνοασφάλεια σε κοινωνικό και ηθικό επίπεδο. Η ανάπτυξη νέων γνωσιακών CAPTCHA που βασίζονται σε φυσική αλληλεπίδραση προβάλλει ως κρίσιμη στρατηγική αντιμετώπισης.

4.2.3 Η Χρήση της Τεχνητής Νοημοσύνης για την Ενίσχυση Επιθέσεων Τύπου Zero-Day

Οι επιθέσεις zero-day εκμεταλλεύονται άγνωστες ή μη διορθωμένες ευπάθειες, καθιστώντας τες ιδιαίτερα επικίνδυνες (Roumani, 2021). Η τεχνητή νοημοσύνη, μέσω τεχνικών όπως η μηχανική και βαθιά μάθηση, ενισχύει τον εντοπισμό και την πρόβλεψη ευπαθειών (Brundage et al., 2018; Zhang, Wang, & Chen, 2020), ενώ semantic analysis μπορεί να οδηγήσει σε αυτοματοποιημένη δημιουργία exploit code (Sadeghian, Zhang, & Amin, 2019).

Εξειδικευμένα εργαλεία όπως το VulDeePecker, το DeepExploit, το SemFuzz και το CodeXray αξιοποιούν διαφορετικές μορφές ΤΝ για ανίχνευση και εκμετάλλευση ευπαθειών, ενώ το GenAI Phishing χρησιμοποιεί Generative AI για εξατομικευμένες επιθέσεις που λειτουργούν ως προθάλαμος για zero-day exploits (Li et al., 2018; Yamato et al., 2019; Euronews Next, 2024).

Οι επιπτώσεις είναι πολυδιάστατες: αιφνίδιες παραβιάσεις κρίσιμων συστημάτων, οικονομικές απώλειες, ψυχολογικές συνέπειες όπως φόβος και άγχος, καθώς και ζητήματα εμπιστοσύνης και απονομής δικαιοσύνης (Davies, 2023; Chesney & Citron, 2019).

Συνολικά, η ΤΝ έχει μετατρέψει τις επιθέσεις zero-day σε πιο εξελιγμένες και αυτοματοποιημένες απειλές, με συνέπειες που επεκτείνονται πέρα από την κυβερνοασφάλεια σε οικονομικό, κοινωνικό και ψυχολογικό επίπεδο.

4.2.4 Δημιουργία προσαρμοστικού κακόβουλου λογισμικού με τη χρήση τεχνητής νοημοσύνης

Η τεχνητή νοημοσύνη έχει καταστήσει το προσαρμοστικό κακόβουλο λογισμικό (adaptive malware) ιδιαίτερα επικίνδυνο, καθώς επιτρέπει σε επιθέσεις να μεταβάλλουν τη συμπεριφορά τους σε πραγματικό χρόνο, να αποκρύπτουν τον κώδικα και να παρακάμπτουν μηχανισμούς ασφαλείας (Ullah, Ahmad, & Khan, 2022). Η βιβλιογραφία δείχνει ότι η ΤΝ ενισχύει την αυτο-εξέλιξη του malware, καθιστώντας το πιο ανθεκτικό απέναντι σε

παραδοσιακά εργαλεία ανίχνευσης (Fritsch, Jaber, & Yazidi, 2023), ενώ οι πολυμορφικές απειλές μπορούν να αλλάζουν συνεχώς υπογραφή και να ξεφεύγουν από στατικά συστήματα (Song et al., 2025). Επιπλέον, οι επιθέσεις αξιοποιούν σύνθετες στρατηγικές πολλαπλών επιπέδων, συνδυάζοντας μηχανική μάθηση, βαθιά μάθηση και αλγορίθμους γραφημάτων (Ramaswamy, 2024).

Οι επιπτώσεις είναι πολυδιάστατες: οι επιθέσεις γίνονται πιο περίπλοκες και δύσκολα ανιχνεύσιμες, προκαλούν σοβαρές λειτουργικές και οικονομικές απώλειες και εγείρουν κρίσιμα νομικά και ηθικά ζητήματα. Συνολικά, η ΤΝ έχει μετασχηματίσει το προσαρμοστικό malware σε μια εξελιγμένη απειλή που επηρεάζει όχι μόνο την κυβερνοασφάλεια αλλά και τον οικονομικό και κοινωνικό τομέα.

4.2.5 Deepfakes και Παραπληροφόρηση με χρήση της Τεχνητής Νοημοσύνης

Η τεχνολογία των deepfakes, βασισμένη σε αλγορίθμους βαθιάς μάθησης όπως τα GANs, επιτρέπει την παραγωγή εξαιρετικά ρεαλιστικών ψευδών βίντεο και ηχητικών αρχείων, καθιστώντας δύσκολη τη διάκριση από το αυθεντικό υλικό (Chesney & Citron, 2019). Αν και μπορεί να αξιοποιηθεί δημιουργικά, η επιθετική χρήση της έχει καταγραφεί ως εργαλείο κοινωνικής μηχανικής, παραπληροφόρησης και πολιτικής αποσταθεροποίησης (Kietzmann et al., 2020; Paris & Donovan, 2019; Westerlund, 2019).

Μελέτες δείχνουν ότι τα deepfakes διαβρώνουν την εμπιστοσύνη στα μέσα ενημέρωσης, ενισχύουν γεωπολιτικές εντάσεις και μπορούν να επηρεάσουν εκλογικά αποτελέσματα (Vaccari & Chadwick, 2020; Shoaib et al., 2023). Στην Ελλάδα, έχουν καταγραφεί περιστατικά παραποιημένων βίντεο με δημοσιογράφους, sextortion και αύξηση 37% στη χρήση τους σε fake news και πολιτικές εκστρατείες (ΕΕΤΤ, 2023; ΕΡΤ, 2024).

Οι επιπτώσεις είναι οικονομικές, ηθικές και πολιτικές: βλάβη στο κύρος επιχειρήσεων, απώλεια εμπιστοσύνης των πολιτών στην ψηφιακή τεχνολογία, χειραγώγηση ψηφοφόρων και πρόκληση διεθνών εντάσεων (Citron & Chesney, 2019; Paris & Donovan, 2019).

4.2.6 Εξαγωγή δεδομένων μέσω AI-driven social engineering με χρήση της Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη έχει ενισχύσει την κοινωνική μηχανική, επιτρέποντας την αυτοματοποιημένη ανάλυση δημόσιων προφίλ και επικοινωνιών για την εξαγωγή προσωπικών δεδομένων. Μέσω NLP και προγνωστικής ανάλυσης, δημιουργούνται πειστικά σενάρια εξαπάτησης (Borges et al., 2022; Tsiatsos & Karyda, 2023). Γενετικά μοντέλα όπως

LLMs και GANs παράγουν εξατομικευμένα μηνύματα, email και οπτικοακουστικό υλικό, οδηγώντας σε spear-phishing, smishing και vishing με υψηλή πειστικότητα (Kietzmann et al., 2020; Shoaib et al., 2023).

Η ΤΝ μπορεί να συνδυάζει δεδομένα από κοινωνικά δίκτυα και δημόσιες βάσεις, δημιουργώντας ολοκληρωμένα προφίλ για μαζικές αλλά φαινομενικά «χειροποίητες» επιθέσεις (Fernández Gambín et al., 2024). Οι επιπτώσεις περιλαμβάνουν αυξημένο κίνδυνο παραβίασης δεδομένων και οικονομικών απωλειών, ενίσχυση παραπληροφόρησης και διάβρωση της εμπιστοσύνης στα μέσα ενημέρωσης, καθώς και απειλές για την κυβερνοασφάλεια και την πολιτική σταθερότητα.

4.2.7 Ανίχνευση και εκμετάλλευση ανθρώπινων αδυναμιών

Η τεχνητή νοημοσύνη ενισχύει τις επιθέσεις κοινωνικής μηχανικής όχι μόνο με ψευδή περιεχόμενα αλλά και με την εκμετάλλευση ψυχολογικών και συναισθηματικών αδυναμιών. Μέσω NLP και πολυτροπικών μοντέλων, οι επιτιθέμενοι μπορούν να ανιχνεύσουν σημάδια κόπωσης ή άγχους και να επιλέξουν το κατάλληλο χρονικό σημείο για την επίθεση (Borges et al., 2022; Tsiatsos & Karyda, 2023). Γενετικά μοντέλα όπως LLMs και GANs παράγουν εξατομικευμένα μηνύματα και πολυμεσικό υλικό, οδηγώντας σε εξελιγμένες μορφές spear-phishing, smishing και vishing που είναι δύσκολο να ανιχνευθούν (Kietzmann et al., 2020; Shoaib et al., 2023).

Η ΤΝ μπορεί να συνδυάζει δεδομένα από κοινωνικά δίκτυα και δημόσιες βάσεις, δημιουργώντας ολιστικά προφίλ στόχων για εξατομικευμένες επιθέσεις μεγάλης κλίμακας, οι οποίες μοιάζουν «χειροποίητες» αλλά παράγονται μαζικά (Fernández Gambín et al., 2024). Οι επιπτώσεις είναι πολυδιάστατες: σε οργανωσιακό επίπεδο αυξάνεται ο κίνδυνος παραβίασης δεδομένων και οικονομικών απωλειών, ενώ σε κοινωνικό επίπεδο ενισχύεται η παραπληροφόρηση και η διάβρωση της εμπιστοσύνης στα μέσα ενημέρωσης. Η αυτοματοποίηση και η κλιμάκωση αυτών των επιθέσεων καθιστούν την ανίχνευση δυσκολότερη, απειλώντας την κυβερνοασφάλεια και την πολιτική σταθερότητα.

4.2.8 Adversarial attacks σε συστήματα ΤΝ

Οι adversarial attacks αποτελούν κρίσιμη απειλή για τα συστήματα τεχνητής νοημοσύνης, καθώς βασίζονται στη δημιουργία παραπλανητικών δεδομένων εισόδου που οδηγούν τα μοντέλα σε λανθασμένες αποφάσεις. Η θεμελιώδης εργασία των Goodfellow, Shlens και Szegedy (2014) έδειξε ότι ακόμη και μικρές, αόρατες διαταραχές μπορούν να προκαλέσουν

σοβαρά σφάλματα ταξινόμησης. Οι επιθέσεις διακρίνονται σε white-box, όπου ο αντίπαλος έχει πλήρη γνώση του μοντέλου, και σε black-box, όπου αξιοποιείται η μεταφερσιμότητα παραδειγμάτων από surrogate μοντέλα (Qiu et al., 2019). Μπορούν να είναι στοχευμένες ή μη-στοχευμένες, με διαφορετικά ποσοστά επιτυχίας (Wang et al., 2023).

Η έρευνα δείχνει ότι οι επιθέσεις δεν περιορίζονται στην όραση υπολογιστών αλλά επεκτείνονται σε φωνή, κείμενο και δίκτυα, αποδεικνύοντας την πολυτροπική φύση του φαινομένου (Abomakhelb et al., 2023). Η μεταφερσιμότητα των adversarial παραδειγμάτων επιτρέπει την εκμετάλλευση πολλαπλών μοντέλων με κοινά μοτίβα ευαλωτότητας, αυξάνοντας τον κίνδυνο μαζικών επιθέσεων (Ahmed et al., 2024).

Οι επιπτώσεις είναι σοβαρές: υπονομεύεται η αξιοπιστία και η ασφάλεια των συστημάτων TN, μειώνεται η εμπιστοσύνη στη χρήση τους σε κρίσιμες εφαρμογές και αυξάνεται ο κίνδυνος λειτουργικών και κοινωνικών συνεπειών σε πεδία όπως η υγεία, η αυτόνομη οδήγηση και οι τηλεπικοινωνίες (Wang et al., 2023).

4.2.9 Επιθέσεις σε αυτόνομα συστήματα και IoT

Η ενσωμάτωση τεχνητής νοημοσύνης σε αυτόνομα συστήματα και δίκτυα IoT έχει διευρύνει την επιφάνεια απειλής, καθιστώντας τα ιδιαίτερα ευάλωτα λόγω ετερογένειας και περιορισμένης υπολογιστικής ισχύος (Radanliev et al., 2024). Οι επιθέσεις μπορούν να επηρεάσουν κρίσιμες λειτουργίες όπως δρομολόγηση και ανίχνευση ανωμαλιών, οδηγώντας σε λειτουργική αστάθεια (Khazane et al., 2024). Σε αυτόνομα οχήματα και drones, η παραποίηση σημάτων GPS και η χειραγώγηση μοντέλων αντίληψης μπορεί να προκαλέσει εσφαλμένες αποφάσεις πλοήγησης και συγκρούσεις (IEEE, 2025; Xing et al., 2025).

Παράλληλα, η συλλογή και συσχέτιση δεδομένων από IoT συσκευές ενισχύει τον κίνδυνο παραβίασης ιδιωτικότητας και μαζικής επιτήρησης (Radanliev et al., 2024). Οι επιθέσεις αυτές δεν αποτελούν μόνο τεχνικές προκλήσεις αλλά πολυδιάστατες απειλές που επηρεάζουν τη φυσική ασφάλεια, τη λειτουργική διαθεσιμότητα και την κοινωνική εμπιστοσύνη. Συνολικά, η βιβλιογραφία αναδεικνύει την ανάγκη για ολιστική προσέγγιση στην ασφάλεια των αυτόνομων συστημάτων και του IoT (Xing et al., 2025).

4.2.10 Παραποίηση νομικών και οικονομικών εγγράφων

Η χρήση γενετικών μοντέλων όπως GANs και LLMs έχει δημιουργήσει νέες μορφές επιθέσεων που ξεπερνούν την απλή πλαστογραφία, παράγοντας έγγραφα με υψηλή γλωσσική και μορφολογική πιστότητα που μοιάζουν αυθεντικά και δύσκολα ανιχνεύσιμα (Kumar et al.,

2023). Η ΤΝ μπορεί να δημιουργήσει ψευδή συμβόλαια, οικονομικές καταστάσεις και φορολογικά έγγραφα, διευκολύνοντας επιθέσεις απάτης στο πλαίσιο της AI-enabled forgery (Kietzmann et al., 2023).

Σε οικονομικό επίπεδο, παραποιημένα τιμολόγια και τραπεζικά έγγραφα μπορούν να οδηγήσουν σε απώλειες μεγάλης κλίμακας (Zhang et al., 2024), ενώ σε νομικό επίπεδο η δημιουργία ψευδών συμβολαίων και δικαστικών εγγράφων υπονομεύει την εμπιστοσύνη στους θεσμούς και τη διαφάνεια του συστήματος (Floridi, 2023). Συνολικά, η βιβλιογραφία συγκλίνει ότι οι επιθέσεις αυτές αποτελούν νέα κατηγορία γενετικών απειλών, με άμεσες επιπτώσεις στην ασφάλεια, την οικονομία και τη δικαιοσύνη, ενισχύοντας την απάτη και την κοινωνικο-οικονομική χειραγώγηση (Kietzmann et al., 2023).

4.3 Επιπτώσεις από Μορφές Επιθέσεων με Χρήση ΤΝ

Η συστηματική μελέτη δείχνει ότι κάθε μορφή επίθεσης με χρήση τεχνητής νοημοσύνης έχει διαφορετικά χαρακτηριστικά και επιπτώσεις σε τεχνικό, οικονομικό, νομικό και κοινωνικό επίπεδο. Οι αυτοματοποιημένες επιθέσεις phishing οδηγούν σε απώλεια δεδομένων και οικονομικών πόρων, υπονομεύοντας την εμπιστοσύνη στις ψηφιακές επικοινωνίες (Bose & Leung, 2023). Η παραβίαση CAPTCHA διευκολύνει μη εξουσιοδοτημένη πρόσβαση και ενισχύει αυτοματοποιημένες επιθέσεις (Yan et al., 2022). Οι επιθέσεις zero-day επιταχύνονται μέσω ΤΝ, αυξάνοντας τον κίνδυνο για κρίσιμες υποδομές (Shen et al., 2023).

Το προσαρμοστικό κακόβουλο λογισμικό αποκτά ανθεκτικότητα και παραμένει αόρατο για μεγαλύτερο διάστημα (Demetrio et al., 2021). Τα deepfakes υπονομεύουν την αξιοπιστία της πληροφόρησης και ενισχύουν την παραπληροφόρηση σε πολιτικό και κοινωνικό επίπεδο (Chesney & Citron, 2019). Η AI-driven social engineering δημιουργεί εξατομικευμένα σενάρια εξαπάτησης, οδηγώντας σε παραβίαση ιδιωτικότητας (Hadnagy, 2021). Η ανίχνευση ανθρώπινων αδυναμιών μέσω ΤΝ ενισχύει την χειραγώγηση και υπονομεύει την αυτονομία των χρηστών (West, 2020).

Οι adversarial attacks αποδεικνύουν τη δομική ευαλωτότητα των μοντέλων ΤΝ, προκαλώντας λανθασμένες αποφάσεις σε κρίσιμες εφαρμογές (Goodfellow et al., 2014; Qiu et al., 2019). Οι επιθέσεις σε αυτόνομα συστήματα και IoT μπορούν να οδηγήσουν σε απώλεια ελέγχου, φυσικές συνέπειες και παραβίαση ιδιωτικότητας (Radanliev et al., 2024; Xing et al., 2025). Τέλος, η παραποίηση νομικών και οικονομικών εγγράφων μέσω γενετικών μοντέλων υπονομεύει θεσμούς και δημιουργεί κινδύνους για την κοινωνικο-οικονομική σταθερότητα (Kumar et al., 2023; Floridi, 2023).

Συνολικά, οι μορφές επιθέσεων με ΤΝ συνιστούν πολυδιάστατες απειλές που επηρεάζουν την κυβερνοασφάλεια, την οικονομία, την κοινωνική εμπιστοσύνη και τη θεσμική διαφάνεια, αναδεικνύοντας την ανάγκη για ολιστική στρατηγική αντιμετώπισης.

4.4 Επίλογος

Η συνοπτική παρουσίαση των ευρημάτων ανέδειξε ότι η τεχνητή νοημοσύνη λειτουργεί ως διττός μηχανισμός, αφενός ενισχύει την καινοτομία και την αυτοματοποίηση, αφετέρου προσφέρει στους επιτιθέμενους νέα εργαλεία για την ανάπτυξη εξελιγμένων μορφών κυβερνοεπιθέσεων. Επιπλέον, τα ευρήματα αποδεικνύουν πως η τεχνητή νοημοσύνη επηρεάζει όχι μόνο την κυβερνοασφάλεια αλλά και την κοινωνική εμπιστοσύνη, την οικονομική σταθερότητα και τη θεσμική διαφάνεια (Brundage et al., 2018; Chesney & Citron, 2019; Stylianou et al., 2025).

Η πολυδιάστατη φύση των επιπτώσεων καταδεικνύει ότι οι απειλές δεν περιορίζονται σε τεχνικό επίπεδο, αλλά επεκτείνονται σε ψυχολογικές, κοινωνικές και πολιτικές διαστάσεις, υπονομεύοντας την ασφάλεια και την αξιοπιστία του ψηφιακού οικοσυστήματος. Συνεπώς, η κατανόηση των ευρημάτων αποτελεί κρίσιμο βήμα για την ανάπτυξη ολοκληρωμένων στρατηγικών αντιμετώπισης, οι οποίες θα εξεταστούν στα επόμενα κεφάλαια.

Κεφάλαιο 5ο: Συζήτηση - συμπεράσματα

5.1 Εισαγωγή

Η μελέτη που προηγήθηκε ανέδειξε την πολυπλοκότητα και τη διττή φύση της τεχνητής νοημοσύνης στον χώρο των κυβερνοεπιθέσεων. Η ανάλυση ανέδειξε ότι η τεχνητή νοημοσύνη δεν αποτελεί απλώς τεχνολογική καινοτομία, αλλά έναν στρατηγικό πολλαπλασιαστή ισχύος για τους κυβερνοεγκληματίες, μετασχηματίζοντας τις μορφές και την ένταση των επιθέσεων. Τα δεδομένα αποκαλύπτουν πώς οι διαφορετικές τεχνικές — από τις αυτοματοποιημένες επιθέσεις phishing και την παραβίαση CAPTCHA έως τα zero-day exploits, το adaptive malware και τα deepfakes — επηρεάζουν όχι μόνο την κυβερνοασφάλεια αλλά και την κοινωνική εμπιστοσύνη, την οικονομική σταθερότητα και τη θεσμική διαφάνεια. Παράλληλα, αναδεικνύεται η ανάγκη για ολιστική στρατηγική αντιμετώπισης, η οποία θα συνδυάζει τεχνικές λύσεις, θεσμικά μέτρα και κοινωνική ευαισθητοποίηση.

5.2 Συμπεράσματα

Σκοπός της παρούσας ήταν η καταγραφή, ομαδοποίηση και συστηματοποίηση των τακτικών των κυβερνοεπιθέσεων που αξιοποιούν τεχνητή νοημοσύνη, ανάλογα με το είδος της επίθεσης.

Τα δεδομένα έδειξαν ότι η τεχνητή νοημοσύνη (ΤΝ) λειτουργεί ως καταλύτης μετασχηματισμού του κυβερνοεγκλήματος. Οι αυτοματοποιημένες επιθέσεις phishing, η παραβίαση CAPTCHA, οι zero day επιθέσεις και το προσαρμοστικό κακόβουλο λογισμικό αποδεικνύουν την τεχνική ευελιξία και την αυξημένη ανθεκτικότητα των απειλών (Bose & Leung, 2023; Demetrio et al., 2021; Shen et al., 2023; Yan et al., 2022).

Παράλληλα, τα deepfakes και η παραποίηση νομικών και οικονομικών εγγράφων υπονομεύουν την αξιοπιστία της πληροφόρησης και των θεσμών, ενισχύοντας την παραπληροφόρηση και τη θεσμική διάβρωση (Chesney & Citron, 2019; Floridi, 2023; Kumar et al., 2023; Zhang et al., 2024).

Η κοινωνική μηχανική που υποστηρίζεται από την τεχνητή νοημοσύνη, είτε μέσω AI driven social engineering είτε μέσω ανίχνευσης ψυχολογικών αδυναμιών, ενισχύει την εξατομίκευση των επιθέσεων και οδηγεί σε παραβίαση ιδιωτικότητας και χειραγώγηση χρηστών (Borges et al., 2022; Hadnagy, 2021; West, 2020). Οι adversarial attacks αποδεικνύουν τη δομική ευαλωτότητα των μοντέλων ΤΝ, καθώς ακόμη και μικρές διαταραχές μπορούν να οδηγήσουν

σε λανθασμένες αποφάσεις σε κρίσιμες εφαρμογές (Goodfellow et al., 2014; Qiu et al., 2019; Wang et al., 2023). Τέλος, οι επιθέσεις σε αυτόνομα συστήματα και IoT αναδεικνύουν τον κίνδυνο φυσικών συνεπειών, παραβίασης ιδιωτικότητας και λειτουργικής αστάθειας (Radanliev et al., 2024; Xing et al., 2025).

Σε όλες τις διαστάσεις —τεχνική, οικονομική, νομική, κοινωνική και δεοντολογική— η ΤΝ μετατρέπει τις ανθρώπινες και τεχνολογικές αδυναμίες σε σημεία εισόδου για επιθέσεις υψηλής κρισιμότητας. Η βιβλιογραφία συγκλίνει στην ανάγκη για ολιστική στρατηγική αντιμετώπισης, που θα συνδυάζει τεχνικές λύσεις, θεσμικά μέτρα και κοινωνική ευαισθητοποίηση, ώστε να περιοριστεί η ισχύς της ΤΝ ως εργαλείου απειλής και να αξιοποιηθεί υπεύθυνα για το συλλογικό καλό.

5.3 Απαντήσεις στα ερευνητικά ερωτήματα

5.3.1 Ποιες είναι οι κύριες τακτικές και τεχνικές κυβερνοεπιθέσεων που αξιοποιούν τεχνητή νοημοσύνη;

Σε ότι αφορά το πρώτο ερευνητικό ερώτημα, η ΤΝ αξιοποιείται για την αυτοματοποίηση και εξατομίκευση επιθέσεων, μέσω τεχνικών όπως η επεξεργασία φυσικής γλώσσας (NLP), η βαθιά μάθηση (deep learning), τα Generative Adversarial Networks (GANs) και τα Large Language Models (LLMs).

Οι τεχνικές αυτές επιτρέπουν:

- Την δημιουργία πειστικών phishing μηνυμάτων και deepfakes.
- Την παραβίαση CAPTCHA με OCR και αναγνώριση εικόνας.
- Την ανάλυση μεγάλων δεδομένων για zero day exploits.
- Την παραγωγή προσαρμοστικού malware που μεταβάλλει τη συμπεριφορά του σε πραγματικό χρόνο.
- Την ανίχνευση ψυχολογικών αδυναμιών και δημιουργία εξατομικευμένων σεναρίων κοινωνικής μηχανικής.

5.3.2 Ποιοι τύποι επιθέσεων παρατηρούνται και ποια είναι τα βασικά χαρακτηριστικά τους;

Σε ότι αφορά το δεύτερο ερευνητικό ερώτημα, τα δεδομένα έδειξαν ότι οι βασικοί τύποι επιθέσεων είναι:

- Phishing με ΤΝ: εξατομικευμένα και πολυγλωσσικά μηνύματα με υψηλή πειστικότητα.
- Παραβίαση CAPTCHA: χρήση νευρωνικών δικτύων για παράκαμψη μηχανισμών επαλήθευσης.
- Zero day exploits: ταχεία αναγνώριση και εκμετάλλευση άγνωστων ευπαθειών.
- Adaptive malware: δυναμική αλλαγή υπογραφής και συμπεριφοράς για αποφυγή ανίχνευσης.
- Deepfakes: παραγωγή ψευδών οπτικοακουστικών δεδομένων για παραπληροφόρηση.
- AI driven social engineering: αυτοματοποιημένη ανάλυση προφίλ και δημιουργία εξατομικευμένων επιθέσεων.
- Adversarial attacks: στρατηγικές διαταραχές στα δεδομένα εισόδου που παραπλανούν μοντέλα ΤΝ.
- Επιθέσεις σε IoT και αυτόνομα συστήματα: παραποίηση αισθητήρων και δικτυακών καναλιών με φυσικές συνέπειες.
- Παραποίηση νομικών και οικονομικών εγγράφων: παραγωγή πλαστών εγγράφων με υψηλή πιστότητα.

5.3.3 Ποιες είναι οι επιπτώσεις αυτών των επιθέσεων και σε ποιους τομείς εκδηλώνονται;

Σε ότι αφορά το τρίτο ερευνητικό ερώτημα, τα δεδομένα έδειξαν ότι οι επιπτώσεις είναι πολυδιάστατες και διακρίνονται σε:

- Τεχνικές: αυξημένη πολυπλοκότητα, δυσκολία ανίχνευσης, λειτουργική αστάθεια.
- Οικονομικές: απώλειες πόρων, ζημιές σε επιχειρήσεις, χειραγώγηση αγορών.
- Νομικές/θεσμικές: υπονόμευση εμπιστοσύνης σε θεσμούς, παραβίαση δικαιοσύνης και διαφάνειας.
- Κοινωνικές/ψυχολογικές: παραπληροφόρηση, διάβρωση εμπιστοσύνης στα μέσα ενημέρωσης, άγχος και μείωση αυτονομίας χρηστών.
- Δεοντολογικές: μαζική επιτήρηση, κοινωνικο οικονομική χειραγώγηση, θεσμική διάβρωση.

5.3.4 Ποια είναι τα ιδιαίτερα χαρακτηριστικά που πρέπει να έχουν τα μέτρα προστασίας και άμυνας έναντι των επιθέσεων που βασίζονται στην τεχνητή νοημοσύνη;

Τέλος, σε ότι αφορά το τέταρτο ερευνητικό ερώτημα, τα δεδομένα έδειξαν ότι τα ιδιαίτερα χαρακτηριστικά που πρέπει να έχουν τα μέτρα προστασίας και άμυνας έναντι των επιθέσεων που βασίζονται στην τεχνητή νοημοσύνη είναι:

- Ολιστικά: να καλύπτουν τεχνικές, κοινωνικές και θεσμικές διαστάσεις.
- Προσαρμοστικά: να εξελίσσονται δυναμικά απέναντι σε adaptive malware και adversarial attacks.
- Προληπτικά: να εντοπίζουν zero day ευπάθειες και deepfake περιεχόμενο πριν την εξάπλωση.
- Διαφανή και αξιόπιστα: να ενισχύουν την εμπιστοσύνη των χρηστών και των θεσμών.
- Συνεργατικά: να συνδυάζουν τεχνολογικές λύσεις, θεσμικά μέτρα και κοινωνική ευαισθητοποίηση.

5.4 Συζήτηση

Η ανάλυση των μορφών επιθέσεων με χρήση τεχνητής νοημοσύνης καταδεικνύει ότι η ΤΝ έχει μετατραπεί σε στρατηγικό πολλαπλασιαστή ισχύος για τους κυβερνοεγκληματίες, ενισχύοντας την αποτελεσματικότητα, την κλίμακα και την πολυπλοκότητα των επιθέσεων. Οι αυτοματοποιημένες πρακτικές phishing, η παραβίαση CAPTCHA, οι zero day exploits και το adaptive malware δείχνουν ότι οι τεχνικές άμυνας που βασίζονται σε στατικά μοτίβα δεν επαρκούν πλέον (Bose & Leung, 2023; Demetrio et al., 2021; Shen et al., 2023). Παράλληλα, οι deepfakes και η παραποίηση εγγράφων υπογραμμίζουν ότι οι επιθέσεις δεν περιορίζονται στο τεχνικό πεδίο, αλλά επεκτείνονται σε κοινωνικό, πολιτικό και θεσμικό επίπεδο, υπονομεύοντας την εμπιστοσύνη και τη διαφάνεια (Chesney & Citron, 2019; Floridi, 2023; Kumar et al., 2023).

Η κοινωνική μηχανική που υποστηρίζεται από ΤΝ, είτε μέσω αυτοματοποιημένης ανάλυσης προφίλ είτε μέσω εντοπισμού ψυχολογικών αδυναμιών, αναδεικνύει την ανάγκη για μέτρα προστασίας που δεν περιορίζονται σε τεχνικές λύσεις αλλά ενσωματώνουν και κοινωνικές/ψυχολογικές παραμέτρους (Borges et al., 2022; West, 2020). Οι adversarial attacks αποδεικνύουν τη δομική ευαλωτότητα των μοντέλων ΤΝ, θέτοντας σε αμφισβήτηση την αξιοπιστία τους σε κρίσιμες εφαρμογές όπως η υγεία και η αυτόνομη οδήγηση (Goodfellow et al., 2014; Qiu et al., 2019). Αντίστοιχα, οι επιθέσεις σε IoT και αυτόνομα

συστήματα δείχνουν ότι η κυβερνοασφάλεια συνδέεται πλέον άμεσα με τη φυσική ασφάλεια και την κοινωνική εμπιστοσύνη (Radanliev et al., 2024; Xing et al., 2025).

Συνολικά, τα ευρήματα αναδεικνύουν την ανάγκη για ολιστική στρατηγική αντιμετώπισης, η οποία θα συνδυάζει:

- Τεχνικές λύσεις (προσαρμοστικά συστήματα ανίχνευσης, AI based defense).
- Θεσμικά μέτρα (κανονιστικά πλαίσια, διεθνής συνεργασία).
- Κοινωνική ευαισθητοποίηση (εκπαίδευση χρηστών, ενίσχυση εμπιστοσύνης).

Η μελλοντική έρευνα καλείται να εστιάσει σε τρεις κατευθύνσεις:

1. Ανάπτυξη ανθεκτικών μοντέλων ΤΝ που μπορούν να αντιμετωπίσουν adversarial attacks και adaptive malware.
2. Διερεύνηση κοινωνικών και ψυχολογικών παραμέτρων ώστε να περιοριστεί η αποτελεσματικότητα της AI driven social engineering.
3. Διαμόρφωση διεθνών πολιτικών και δεοντολογικών πλαισίων για την υπεύθυνη χρήση της ΤΝ, με στόχο την προστασία θεσμών και κοινωνικής συνοχής.

5.5 Ερευνητικά κενά

Η έρευνα ανέδειξε συγκεκριμένα ερευνητικά κενά που σχετίζονται με:

- Περιορισμένη εμπειρική τεκμηρίωση: Οι περισσότερες μελέτες εστιάζουν σε θεωρητικά μοντέλα ή εργαστηριακά πειράματα (π.χ. adversarial attacks, deepfakes), αλλά λείπουν εκτεταμένα εμπειρικά δεδομένα από πραγματικά περιστατικά σε οργανισμούς και κοινωνίες.
- Διασταύρωση τεχνικών και κοινωνικών επιπτώσεων: Υπάρχει πλούσια βιβλιογραφία για τις τεχνικές διαστάσεις (zero day, adaptive malware), αλλά περιορισμένη ανάλυση για το πώς οι τεχνικές αυτές συνδέονται με κοινωνικές συνέπειες όπως παραπληροφόρηση, ψυχολογική χειραγώγηση και πολιτική αποσταθεροποίηση.
- Ανθεκτικότητα και άμυνα με ΤΝ: Ενώ η ΤΝ μελετάται ως εργαλείο επίθεσης, υπάρχει έλλειψη συστηματικής έρευνας για το πώς μπορεί να αξιοποιηθεί αποτελεσματικά ως εργαλείο άμυνας (π.χ. AI based detection, resilient architectures).
- Δεοντολογικές και νομικές προεκτάσεις: Οι μελέτες αναγνωρίζουν την παραποίηση εγγράφων και την υπονόμευση θεσμών, αλλά δεν υπάρχει επαρκής διερεύνηση για το

πώς πρέπει να διαμορφωθούν διεθνή κανονιστικά πλαίσια και δεοντολογικές αρχές για την υπεύθυνη χρήση της ΤΝ.

- Διαθεματική προσέγγιση: Λείπει η σύνδεση μεταξύ διαφορετικών πεδίων (π.χ. κυβερνοασφάλεια, ψυχολογία, κοινωνιολογία, νομική επιστήμη) ώστε να υπάρξει μια ολοκληρωμένη κατανόηση των επιθέσεων με ΤΝ.

Συνολικά, η υπάρχουσα βιβλιογραφία δείχνει ότι η ΤΝ έχει μετασχηματίσει τις κυβερνοεπιθέσεις σε πολυδιάστατες απειλές, αλλά λείπει μια ολιστική, διεπιστημονική και εμπειρικά τεκμηριωμένη προσέγγιση που να συνδέει τεχνικές, κοινωνικές, οικονομικές και θεσμικές επιπτώσεις με συγκεκριμένες στρατηγικές άμυνας.

5.6 Προτάσεις για μελλοντική έρευνα

Η παρούσα ανασκόπηση ανέδειξε ότι, παρά την πλούσια βιβλιογραφία γύρω από τις επιθέσεις με χρήση τεχνητής νοημοσύνης, εξακολουθούν να υπάρχουν σημαντικά ερευνητικά κενά που απαιτούν περαιτέρω διερεύνηση. Πρώτον, χρειάζεται εμπειρική τεκμηρίωση σε πραγματικά περιβάλλοντα, με μελέτες περίπτωσης και ποσοτικά δεδομένα από οργανισμούς και κοινωνίες, ώστε να καταγραφεί η πραγματική έκταση και αποτελεσματικότητα των επιθέσεων με ΤΝ. Δεύτερον, η μελλοντική έρευνα οφείλει να υιοθετήσει μια διεπιστημονική προσέγγιση που θα συνδέει τεχνικές, κοινωνικές, ψυχολογικές και νομικές διαστάσεις, προκειμένου να αναπτυχθούν πιο ολιστικά μέτρα άμυνας.

Επιπλέον, απαιτείται συστηματική διερεύνηση για την ανάπτυξη ανθεκτικών συστημάτων ΤΝ, τα οποία θα μπορούν να λειτουργήσουν ως εργαλεία άμυνας, αξιοποιώντας resilient architectures, adversarial training και AI-based detection. Παράλληλα, η έρευνα πρέπει να εστιάσει στη διαμόρφωση διεθνών πολιτικών και δεοντολογικών πλαισίων για την υπεύθυνη χρήση της ΤΝ, ιδίως σε πεδία όπως η παραποίηση εγγράφων, η παραπληροφόρηση και η μαζική επιτήρηση. Τέλος, απαιτείται περαιτέρω μελέτη για το πώς οι τεχνικές επιθέσεις μεταφράζονται σε κοινωνικές και πολιτικές επιπτώσεις, όπως η διάβρωση της εμπιστοσύνης, η χειραγώγηση της κοινής γνώμης και η αποσταθεροποίηση θεσμών.

Με βάση τα ευρήματα, προτείνονται οι ακόλουθες κατευθύνσεις:

1. Εμπειρική τεκμηρίωση σε πραγματικά περιβάλλοντα: Απαιτούνται μελέτες περίπτωσης και ποσοτικά δεδομένα από οργανισμούς και κοινωνίες, ώστε να καταγραφεί η πραγματική έκταση και αποτελεσματικότητα των επιθέσεων με ΤΝ.

2. Διεπιστημονική προσέγγιση: Η μελλοντική έρευνα πρέπει να συνδέσει τεχνικές, κοινωνικές, ψυχολογικές και νομικές διαστάσεις, ώστε να υπάρξει ολοκληρωμένη κατανόηση των επιπτώσεων και να αναπτυχθούν πιο ολιστικά μέτρα άμυνας.
3. Ανάπτυξη ανθεκτικών συστημάτων ΤΝ: Χρειάζεται συστηματική διερεύνηση για το πώς η ίδια η ΤΝ μπορεί να χρησιμοποιηθεί ως εργαλείο άμυνας, π.χ. με resilient architectures, adversarial training και AI based detection.
4. Δεοντολογικά και κανονιστικά πλαίσια: Η έρευνα πρέπει να εστιάσει στη διαμόρφωση διεθνών πολιτικών και δεοντολογικών αρχών για την υπεύθυνη χρήση της ΤΝ, ειδικά σε πεδία όπως η παραποίηση εγγράφων, η παραπληροφόρηση και η μαζική επιτήρηση.
5. Σύνδεση με κοινωνικές συνέπειες: Απαιτείται περαιτέρω μελέτη για το πώς οι τεχνικές επιθέσεις μεταφράζονται σε κοινωνικές και πολιτικές επιπτώσεις, όπως η διάβρωση της εμπιστοσύνης, η χειραγώγηση της κοινής γνώμης και η αποσταθεροποίηση θεσμών.

5.7 Επίλογος

Η παρούσα μελέτη ανέδειξε ότι η τεχνητή νοημοσύνη έχει μετασχηματίσει ριζικά το πεδίο των κυβερνοεπιθέσεων, προσδίδοντάς τους νέα χαρακτηριστικά πολυπλοκότητας, εξατομίκευσης και ανθεκτικότητας. Από τις αυτοματοποιημένες επιθέσεις phishing και την παραβίαση CAPTCHA έως τα zero-day exploits, το προσαρμοστικό κακόβουλο λογισμικό και τα deepfakes, η ΤΝ αποδείχθηκε ότι λειτουργεί ως καταλύτης για την ανάπτυξη εξελιγμένων απειλών που επηρεάζουν όχι μόνο την κυβερνοασφάλεια αλλά και την κοινωνική εμπιστοσύνη, την οικονομική σταθερότητα και τη θεσμική διαφάνεια (Brundage et al., 2018; Chesney & Citron, 2019; Floridi, 2023).

Η ανάλυση των ερευνητικών ερωτημάτων κατέδειξε ότι οι επιθέσεις με χρήση ΤΝ έχουν πολυδιάστατες επιπτώσεις — τεχνικές, οικονομικές, κοινωνικές, ψυχολογικές και δεοντολογικές — γεγονός που καθιστά αναγκαία την ανάπτυξη ολιστικών στρατηγικών αντιμετώπισης. Η προστασία απέναντι σε αυτές τις απειλές δεν μπορεί να περιοριστεί σε τεχνικές λύσεις· απαιτεί συνδυασμό θεσμικών μέτρων, διεθνούς συνεργασίας και κοινωνικής ευαισθητοποίησης.

Συνολικά, η εργασία αυτή συμβάλλει στην κατανόηση του τρόπου με τον οποίο η ΤΝ αξιοποιείται σε κυβερνοεπιθέσεις, αναδεικνύοντας τόσο τις προκλήσεις όσο και τις ευκαιρίες

που δημιουργεί. Η μελλοντική έρευνα καλείται να εμβαθύνει σε εμπειρικά δεδομένα, να υιοθετήσει διεπιστημονικές προσεγγίσεις και να διαμορφώσει δεοντολογικά πλαίσια που θα εξασφαλίζουν την υπεύθυνη χρήση της ΤΝ. Μόνο έτσι μπορεί να περιοριστεί η ισχύς της ως εργαλείου απειλής και να αξιοποιηθεί δημιουργικά για το συλλογικό καλό.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Abomakhelb, A., Abd Jalil, K., Buja, A. G., Alhammadi, A., & Alenezi, A. M. (2023). A comprehensive review of adversarial attacks and defense strategies in deep neural networks. Technologies, 13(5), 202. https://www.mdpi.com/2227-7080/13/5/202, προσπέλαση: 01/10/2025

Ahmed, S. Q., Vokkaliga Ganesh, B., Sampath Kumar, S., Mishra, P., Anand, R., & Akurathi, B. (2024). A comprehensive review of adversarial attacks on machine learning. arXiv preprint. https://arxiv.org/abs/2412.11384, προσπέλαση: 01/10/2025

Alansary, S. A., Ayyad, S. M., Talaat, F. M., & Saafan, M. M. (2025). Emerging AI threats in cybercrime: A review of zero-day attacks via machine, deep, and federated learning. Knowledge and Information Systems. https://link.springer.com/article/10.1007/s10115-025-02556-6, προσπέλαση: 04/10/2025

Alansary, M., Zhang, Y., & Ullah, I. (2025). Adaptive malware and AI-driven cyberattacks. Springer.

Brown, H., Hutton, B., Clifford, T., & Graham, I. D. (2021). Using the PRISMA statement and its extensions to write protocols and reports. Systematic Reviews, 10, Article 156. https://doi.org/10.1186/s13643-021-01671-z.

Brown, C., Smith, A., & Zhao, L. (2021). AI-enhanced cyber threats: A systematic review. Journal of Cybersecurity Research, 14(3), 211–229. https://doi.org/10.1016/j.jcsr.2021.03.005.

Borges, M., Silva, A., & Almeida, J. (2022). AI-driven social engineering: Profiling and deception in cybersecurity. Journal of Information Security, 11(3), 145–160.

Bose, I., & Leung, A. C. M. (2023). AI-powered phishing attacks: Emerging threats and countermeasures. Computers & Security, 125, 103050.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502.

Chesney, R., & Citron, D. K. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Affairs, 98(1), 147–155.

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1819.

Citron, D. K. (2020). Sexual privacy in the digital age. Yale Law Journal, 130(7), 1870–1934.

CLTC – Center for Long-Term Cybersecurity. (2025). Beyond phishing: Exploring the rise of AI-enabled cybercrime. https://cltc.berkeley.edu/2025/01/16/beyond-phishing-exploring-the-rise-of-ai-enabled-cybercrime/, προσπέλαση: 10/10/2025

Copeland, B. J. (2025). History of artificial intelligence. Encyclopaedia Britannica. https://www.britannica.com/science/history-of-artificial-intelligence , προσπέλαση: 10/10/2025

Δαμαλίτης, Α-Σ., (2024). Παραπληροφόρηση στα ψηφιακά μέσα κοινωνικής δικτύωσης & η στρατηγική και νομοθετική αντιμετώπιση της από την Ευρωπαϊκή Ένωση. Μεταπτυχιακή Διπλωματική Εργασία. Πανεπιστήμιο Πειραιώς. https://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/16817/Damalitis-2210.pdf?sequence=3&isAllowed=y, προσπέλαση: 12/10/2025

Davies, R. (2023). Psychological tactics of AI-driven phishing: Emotional manipulation and digital trust erosion. Proceedings of the IACIS 2023 Conference. https://iacis.org/iis/2023/2_iis_2023_71-83.pdf, προσπέλαση: 18/10/2025

Demetrio, L., Biggio, B., Fumera, G., & Roli, F. (2021). Adversarial malware: Impact and defense. IEEE Security & Privacy, 19(2), 32–41.

Dinh, N., & Ogiela, L. (2022). Human-artificial intelligence approaches for secure analysis in CAPTCHA codes. EURASIP Journal on Information Security, 2022(8). https://doi.org/10.1186/s13635-022-00134-9.

Doshi, R., Apthorpe, N., & Feamster, N. (2018). Machine learning DDoS detection for consumer Internet of Things devices. 2018 IEEE Security and Privacy Workshops (SPW), 29–35. https://doi.org/10.1109/SPW.2018.00013.

EETT, (2023). Έκθεση Πεπραγμένων 2023. https://www.eett.gr/wp-content/uploads/2024/12/%CE%88%CE%BA%CE%B8%CE%B5%CF%83%CE%B7-%CE%A0%CE%B5%CF%80%CF%81%CE%B1%CE%B3%CE%BC%CE%AD%CE%BD

%CF%89%CE%BD-%CE%95%CE%95%CE%A4%CE%A4-2023.pdf , προσπέλαση: 20/10/2025

Εθνικό Κέντρο Τεκμηρίωσης. (2023). Η εξέλιξη των ελληνικών επιστημονικών δημοσιεύσεων στον τομέα της Τεχνητής Νοημοσύνης. https://www.nationalcoalition.gov.gr/skills-intelligence/h-exelixi-ton-ellinikon-epistimonikon/ , προσπέλαση: 24/10/2025

EPT. (2025). Deep Fakes: Πώς μπορούμε να προστατευτούμε και ποιο είναι τελικά το πραγματικό πρόσωπο της τεχνητής νοημοσύνης; https://www.ertnews.gr/eidiseis/mono-sto-ertgr/deep-fakes-pos-mporoume-na-prostateytoume-kai-poio-einai-telika-to-pragmatiko-prosopo-tis-texnitis-noimosynis/ , προσπέλαση: 24/10/2025

EY Ελλάδος. (2023). Η AI αλλάζει το σκηνικό των κυβερνοεπιθέσεων. Forbes Greece. https://www.forbes.com , προσπέλαση: 27/10/2025

Euronews Next. (2024). AI-generated phishing attacks are getting smarter. Retrieved from https://www.euronews.com/next, προσπέλαση: 30/10/2025

Fernández Gambín, Á., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. Artificial Intelligence Review, 57(64). https://doi.org/10.1007/s10462-023-10679-x.

FireEye. (2020, December 13). Highly evasive attacker leverages SolarWinds supply chain to compromise multiple global victims with SUNBURST backdoor. https://www.fireeye.com/blog/threat-research/2020/12/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor.html, προσπέλαση: 30/10/2025

Floridi, L. (2023). AI and the future of trust in legal systems. AI & Society, 38(4), 1123–1135. https://doi.org/10.1007/s00146-023-01567-9.

Fritsch, L., Jaber, A., & Yazidi, A. (2023). An overview of artificial intelligence used in malware. In Nordic Artificial Intelligence Research and Development Conference (pp. 41–51). Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint. https://arxiv.org/abs/1412.6572.

Hadnagy, C. (2021). Social engineering: The science of human hacking. Wiley.

Hameed, I. A., Kumaran, V. P., Meena, U., & Swetha, T. (2025). Leveraging AI for enhanced cybersecurity: A comprehensive review. Discover Applied Sciences, 7, Article 584. https://doi.org/10.1007/s42452-025-06773-0.

HP Wolf Security. (2025, April 26). Καμπανάκι για ραγδαία αύξηση ψεύτικων τεστ CAPTCHA από κυβερνοεγκληματίες. In.gr. https://www.in.gr/2025/04/26/in-science/technology/kampanaki-gia-ragdaia-ayksisi-pseytikon-test-captcha-apo-kyvernoegklimaties/ , προσπέλαση: 31/10/2025

IEEE. (2025). Security challenges and solutions for autonomous vehicles and drones in the age of AI. IEEE Xplore. https://ieeexplore.ieee.org/document/10907826, προσπέλαση: 31/10/2025

ISACA. (2024). State of Cybersecurity 2024: Global update on workforce, resources and AI threats. https://www.isaca.org/resources/news-and-trends/newsletters/atisaca/2024/state-of-cybersecurity-report, προσπέλαση: 31/10/2025

Kaspersky. (2024). Rising concerns: AI-driven cyberattacks and the global defense gap. https://www.kaspersky.com/about/press-releases/rising-concerns-lingering-gaps-most-organizations-fear-ai-driven-cyberattacks-but-lack-key-defenses, προσπέλαση: 02/11/2025

Khazane, H., Ridouani, M., Salahdine, F., & Kaabouch, N. (2024). A holistic review of machine learning adversarial attacks in IoT networks. Future Internet, 16(1), 32. https://doi.org/10.3390/fi16010032.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? Business Horizons, 63(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006.

Kietzmann, J., Paschen, J., & Treen, E. (2023). Artificial intelligence in fraud and forgery: Risks and governance. Journal of Business Ethics, 182(3), 635–648. https://doi.org/10.1007/s10551-023-05321-7.

Kumar, R., Singh, A., & Patel, S. (2023). Generative adversarial networks for document forgery: A survey. Computers & Security, 125, 103048. https://doi.org/10.1016/j.cose.2023.103048.

Kumaran, V. P., Swetha, T., Meena, U., & Hameed, I. A. (2025). AI-driven threat intelligence for real-time cybersecurity. Open Access Research Journal of Science & Technology, 4(2), 135. https://oarjst.com/sites/default/files/OARJST-2024-0135.pdf , προσπέλαση: 02/11/2025

Li, Z., Zou, D., Xu, S., Ou, X., Jin, H., Wang, S., & Deng, Z. (2018). VulDeePecker: A deep learning-based system for vulnerability detection. NDSS Symposium 2018.

Manky, D., & Baram, G. (2025). Beyond phishing: Exploring the rise of AI-enabled cybercrime. UC Berkeley.

Mastercard Signal Insights. (2025). Cybercrime and AI: The $15.6 Trillion Threat. https://www.mastercard.com/us/en/news-and-trends/stories/2025/consumer-cybersecurity-survey.html, προσπέλαση: 03/11/2025

McCarthy, J. (2007). What is artificial intelligence? Stanford University. http://www-formal.stanford.edu/jmc/whatisai.html , προσέλαση: 03/11/2025

Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. Knowledge and Information Systems, 67, 6969–7055. https://doi.org/10.1007/s10115-025-02429-y.

OpenLearn. (2025). An introduction to artificial intelligence. Open University. https://www.open.edu/openlearn

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71.

Paris, B., & Donovan, J. (2019). Deepfakes and the infocalypse. Brookings Institution Report.

Pantserev, K. A. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In Cyber Defence in the Age of AI (pp. 37–55). Springer. https://link.springer.com/chapter/10.1007/978-3-030-35746-7_3.

PRISMA Executive. (2020). PRISMA statement. Retrieved from https://www.prisma-statement.org.

Plesner, A., & ETH Zurich. (2024, October 19). New research reveals CAPTCHAs vulnerable to AI exploits. Quanta Intelligence. https://quantaintelligence.ai/2024/10/19/technology/new-research-reveals-captchas-vulnerable-to-ai-exploits , προσπέλαση: 05/11/2025

Psychology Today. (2024). The psychological effects of AI clones and deepfakes. https://www.psychologytoday.com/us/blog/urban-survival/202401/the-psychological-effects-of-ai-clones-and-deepfakes , προσπέλαση: 05/11/2025

Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of artificial intelligence adversarial attack and defense technologies. Applied Sciences, 9(5), 909. https://doi.org/10.3390/app9050909.

Radanliev, P., De Roure, D., Maple, C., Nurse, J. R. C., Nicolescu, R., & Ani, U. (2024). AI security and cyber risk in IoT systems. Frontiers in Big Data, 7. https://doi.org/10.3389/fdata.2024.1402745.

Radanliev, P., De Roure, D., Maple, C., Nurse, J. R. C., Nicolescu, R., & Ani, U. (2024). AI security and cyber risk in IoT systems. arXiv preprint. https://arxiv.org/pdf/2410.09194 , προσπέλαση: 07/11/2025

Rai, D. H. (2024). Artificial Intelligence Through Time: A Comprehensive Historical Review. ResearchGate. https://www.researchgate.net/publication/385939923 , προσπέλαση: 08/11/2025

Ramaswamy, M. (2024). AI-powered advanced threat protection: A novel framework for next-generation malware defense. International Journal of Futuristic Research and Management, 5(22481).

Rodriguez-Vance, T. (2025). Geopolitical implications of artificial intelligence in cybersecurity: A comprehensive analysis. ResearchGate. https://www.researchgate.net/publication/382129529 , προσπέλαση: 10/11/2025

Roumani, Y. (2021). Patching zero-day vulnerabilities: An empirical analysis. Journal of Cybersecurity, 7(1), tyab023. https://doi.org/10.1093/cybsec/tyab023.

Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

Sadeghian, A., Zhang, J., & Amin, S. (2019). SemFuzz: Semantics-based automatic generation of vulnerability triggers. Proceedings of the 2019 IEEE Symposium on Security and Privacy.

Sarker, I. H., Kayes, A. S. M., & Watters, P. (2020). Cybersecurity data science: An overview from machine learning perspective. Journal of Big Data, 7(1), 1–29. https://doi.org/10.1186/s40537-020-00318-5

Salem, A. H., Azzam, S. M., Emam, O. E., & Abohany, A. A. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. Journal of Big Data, 11, Article 105. https://doi.org/10.1186/s40537-024-00957-y.

Samoili, S., Lopez Cobo, M., et al. (2020). AI Watch: Historical Evolution of Artificial Intelligence. European Commission. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC120469 , προσπέλαση: 11/11/2025

Sangfor Technologies. (2025). AI-powered cyber threats: From zero-day exploits to deepfakes and beyond. https://www.sangfor.com/blog/cybersecurity/ai-powered-cyber-threats-zero-day-deepfakes, προσπέλαση: 11/11/2025

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E., & Lockwood, C. (2021). How to properly use the PRISMA Statement. Systematic Reviews, 10(117). https://doi.org/10.1186/s13643-021-01671-z.

Shen, Y., Zhang, J., & Wang, L. (2023). AI-driven zero-day vulnerability discovery. Journal of Cybersecurity, 9(1), 45–62.

Shib Daily. (2024, March 3). Bots now mimic human behavior to bypass reCAPTCHA v3. Shib Daily Tech. https://www.shibdaily.tech/ai-bots-bypass-recaptcha-v3 , προσπέλαση: 11/11/2025

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI. arXiv preprint. https://arxiv.org/abs/2311.17394 , προσπέλαση: 13/11/2025

Shribe. (2024). PRISMA Literature Review (Flow Chart & Example). Retrieved from https://shribe.eu/prisma-literature-review/ , προσπέλαση: 13/11/2025

Singh, T. (2025). Artificial Intelligence-Driven Cyberattacks. In Cybersecurity, Psychology and People Hacking (pp. 167–188). Springer. https://doi.org.

Song, Y., Zhang, D., Wang, J., Wang, Y., Wang, Y., & Ding, P. (2025). Application of deep learning in malware detection: A review. Journal of Big Data, 12(99).

SQ Magazine. (2025). AI cyber attacks statistics 2025: Attacks, deepfakes, ransomware. https://sqmagazine.co.uk/ai-cyber-attacks-statistics/ , προσπέλαση: 13/11/2025

Swetha, T., Kumaran, V. P., Meena, U., & Hameed, I. A. (2025). Ethical challenges in AI-driven cybersecurity: A review. Discover Applied Sciences, 7, Article 585. https://link.springer.com/article/10.1007/s42452-025-06774-z , προσπέλαση: 15/11/2025

Tsiatsos, T., & Karyda, M. (2023). Artificial intelligence and social engineering: Emerging threats and countermeasures. Computers & Security, 124, 102947.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460.

Yamato, Y., Yamaguchi, S., & Yoshioka, K. (2019). DeepExploit: Automated penetration testing using deep reinforcement learning. Proceedings of the Black Hat USA 2019.

Ullah, I., Mahmoud, Q. H., & Alsaqour, R. (2022). A hybrid deep learning approach for malware detection using convolutional neural networks and long short-term memory. Computers & Security, 114, 102586. https://doi.org/10.1016/j.cose.2021.102586.

Ullah, I., Ahmad, S., & Khan, M. (2022). Artificial intelligence in malware evolution: Adaptive threats and countermeasures. International Journal of Information Security, 21(4), 567–582.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social media + Society, 6(1), 1-13. https://doi.org/10.1177/2056305120903408.

Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., Hossain, E., & Poor, H. V. (2023). Adversarial attacks and defenses in machine learning-powered networks: A contemporary survey. arXiv preprint. https://arxiv.org/abs/2303.06302 , προσπέλαση: 15/11/2025

West, D. M. (2020). The global race for AI: Human vulnerabilities and risks. Brookings Institution Press.

Westerlund, M. (2019). The emergence of deepfake technology: A review. Technology Innovation Management Review, 9(11), 39–52.

Wikipedia contributors. (2025). Preferred reporting items for systematic reviews and meta-analyses. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Preferred_reporting_items_for_systematic_reviews_and_meta-analyses , προσπέλαση: 13/11/2025

Xing, W., Li, M., Li, M., & Han, M. (2025). Towards robust and secure embodied AI: A survey on vulnerabilities and attacks. arXiv preprint. https://arxiv.org/pdf/2502.13175 , προσπέλαση: 17/11/2025

Yan, J., Li, H., & Zhang, Y. (2022). Breaking CAPTCHA with AI: Security implications. Information Security Journal, 31(4), 215–229.

Zhang, Y., Wang, S., & Liu, X. (2020). AI-driven vulnerability discovery: Challenges and opportunities. IEEE Access, 8, 219497–219509. https://doi.org/10.1109/ACCESS.2020.3042067.

Zhang, Y., Li, H., & Chen, X. (2024). AI-enabled financial fraud: Detecting and preventing document manipulation. Information Systems Frontiers, 26(2), 451–468. https://doi.org/10.1007/s10796-024-10321-9.