



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

**Ηθική Χρήση Τεχνητής Νοημοσύνης κατά την Ανάλυση
και Ανάπτυξη Πληροφοριακών Συστημάτων**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

ΑΛΕΞΑΝΔΡΗ ΧΑΡΑΛΑΜΠΟΥ
ΚΟΝΤΟΓΙΑΝΝΗ ΝΙΚΟΛΑΟΥ

Επιβλέπουσα: ΔΙΑΜΑΝΤΟΠΟΥΛΟΥ ΒΑΣΙΛΙΚΗ

Μέλη εξεταστικής επιτροπής: ΑΛΕΞΟΠΟΥΛΟΣ ΧΑΡΑΛΑΜΠΟΣ, ΜΑΛΙΑΤΣΟΣ
ΚΩΝΣΤΑΝΤΙΝΟΣ

Σάμος, [Ιούλιος 2026]

© [2026]

των

ΑΛΕΞΑΝΔΡΗ ΧΑΡΑΛΑΜΠΟΥ

ΚΟΝΤΟΓΙΑΝΝΗ ΝΙΚΟΛΑΟΥ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Πίνακας περιεχομένων

Περιεχόμενα

1	Εισαγωγή	1
1.1	Εισαγωγή στην Τεχνητή Νοημοσύνη	1
1.2	Η Σημασία του Θέματος	3
1.3	Σκοπός και Αντικείμενο της Εργασίας	4
1.4	Επισκόπηση Δομής της Διπλωματικής Εργασίας	4
1.5	Μεθοδολογία της Βιβλιογραφικής Ανασκόπησης	5
1.5.1	Επισκόπηση	5
1.5.2	Στρατηγική Αναζήτηση Βιβλιογραφίας.....	6
1.5.3	Κριτήρια Επιλογής.....	6
1.5.4	Κριτήρια Εξαίρεσης.....	6
1.5.5	Αναλυτική Προσέγγιση	6
1.5.6	Περιορισμοί	7
2	Ιστορική Επισκόπηση της Τεχνητής Νοημοσύνης και των Ηθικών Ζητημάτων της.....	8
2.1	Εισαγωγή.....	8
2.2	Ηθικές Προκλήσεις, Λογοδοσία και Ρυθμίσεις στην Τεχνητή Νοημοσύνη	9
2.3	Η Σύγχρονη Τεχνητή Νοημοσύνη: Εξελίξεις, Προκλήσεις και Ηθικά Διλήμματα	9
2.4	Η Άνοδος των Αυτόνομων Συστημάτων	11
2.4.1	Ανάπτυξη Αυτόνομων Τεχνολογιών	11
2.4.2	Προβλήματα Ασφάλειας και Αξιοπιστίας	11
2.4.3	Μεροληψία και Διακρίσεις σε Αυτόνομα Συστήματα.....	11
2.4.4	Επιπτώσεις Απορρήτου.....	12
2.4.5	Παγκόσμια Ανταπόκριση και Ηθικές Κατευθυντήριες Γραμμές	12
2.5	Η Τεχνητή Νοημοσύνη στον Δημόσιο Λόγο	15
2.5.1	Η Αυξανόμενη Ευαισθητοποίηση του Κοινού για την Τεχνητή Νοημοσύνη	15
2.5.2	Συζητήσεις για την Προστασία της Ιδιωτικότητας και την Επιτήρηση.....	16
2.5.3	Τεχνητή Νοημοσύνη και Ανησυχίες για την Απασχόληση	16
2.5.4	Δημόσια Εμπλοκή στην Πολιτική και τη Ρύθμιση της Τεχνητής Νοημοσύνης.....	16
2.5.5	Τεχνητή Νοημοσύνη και Ηθική Επίγνωση.....	16
2.6	Το Σύγχρονο και Μελλοντικό Πλαίσιο της Ηθικής της Τεχνητής Νοημοσύνης.....	17
2.6.1	Τρέχον Ηθικό τοπίο στην Τεχνητή Νοημοσύνη.....	17
2.6.2	Προβλέποντας το Μέλλον της Ηθικής της Τεχνητής Νοημοσύνης	18
2.6.3	Ενσωμάτωση της Ηθικής στην Σχεδίαση και Ανάπτυξη Συστημάτων Τεχνητής Νοημοσύνης	19
3	Βασικά Θέματα στην Ηθική της Τεχνητής Νοημοσύνης.....	21
3.1	Αλγοριθμική Προκατάληψη και Δικαιοσύνη στην Τεχνητή Νοημοσύνη	21
3.1.1	Πηγές Μεροληψίας στην Τεχνητή Νοημοσύνη.....	21
3.1.2	Αντίκτυπος στην Κοινωνία.....	21

3.1.3	Προκλήσεις στον Μετριάσμό της Προκατάληψης.....	22
3.1.4	Προσεγγίσεις για τη Διασφάλιση της Δικαιοσύνης.....	22
3.1.5	Ηθικά και Δεοντολογικά Πλαίσια για την Αντιμετώπιση της Αλγοριθμικής Μεροληψίας	23
3.2	Προστασία Δεδομένων και Διακυβέρνηση Δεδομένων στην Τεχνητή Νοημοσύνη	26
3.2.1	Κατανόηση των Ανησυχιών περί Προστασίας Δεδομένων στην Τεχνητή Νοημοσύνη.....	26
3.2.2	Η Πρόκληση της Διακυβέρνησης Δεδομένων στην Τεχνητή Νοημοσύνη.....	26
3.2.3	Εξισορρόπηση της Χρήσης Δεδομένων με την Προστασία της Ιδιωτικότητας	27
3.2.4	Νομικά Πλαίσια και Κανονιστική Συμμόρφωση	27
3.2.5	Τεχνολογικές Λύσεις για τη Προστασία της Ιδιωτικότητας.....	29
3.2.6	Μοντέλα Ηθικής Διακυβέρνησης Δεδομένων	29
3.2.7	Προστασία της Ιδιωτικότητας και Υπεύθυνη Χρήση της Τεχνητής Νοημοσύνης.....	30
3.3	Αυτονομία και Υπευθυνότητα στα Συστήματα Τεχνητής Νοημοσύνης.....	30
3.3.1	Η Αυξανόμενη Αυτονομία των Συστημάτων Τεχνητής Νοημοσύνης.....	30
3.3.2	Προκλήσεις στην Απόδοση Λογοδοσίας.....	31
3.3.3	Ανάπτυξη Πλαισίων Λογοδοσίας για Τεχνητή Νοημοσύνη.....	31
3.3.4	Νομικές και Ηθικές Επιπτώσεις της Αυτονομίας της Τεχνητής Νοημοσύνης	32
3.3.5	Ο Ρόλος της Διαφάνειας στη Λογοδοσία της Τεχνητής Νοημοσύνης.....	32
3.3.6	Μελλοντικές Κατευθύνσεις στην Αυτονομία και την Υπευθυνότητα.....	32
3.4	Διαφάνεια και Επεξήγηση στην Τεχνητή Νοημοσύνη	33
3.4.1	Μελλοντικές Κατευθύνσεις στην Αυτονομία και την Υπευθυνότητα.....	33
3.4.2	Ξεπερνώντας τις Προκλήσεις της Διαφάνειας της Τεχνητής Νοημοσύνης.....	33
3.4.3	Υπογράμμιση της Σημασίας της Εξηγήσιμης Τεχνητής Νοημοσύνης	34
3.4.4	Η Αλληλεπίδραση της Διαφάνειας και της Ηθικής Τεχνητής Νοημοσύνης	34
3.4.5	Μελλοντικές Κατευθύνσεις για Διαφανή και Εξηγήσιμη Τεχνητή Νοημοσύνη	35
3.5	Αλληλεπίδραση Ανθρώπου – Τεχνητής Νοημοσύνης και Κοινωνικές Επιπτώσεις	35
3.5.1	Δυναμική της Αλληλεπίδρασης Ανθρώπου – Τεχνητής Νοημοσύνης	35
3.5.2	Κοινωνικός Αντίκτυπος της Τεχνητής Νοημοσύνης.....	36
3.5.3	Ηθικές Θεωρήσεις στην Αλληλεπίδραση Ανθρώπου – Τεχνητής Νοημοσύνης	36
3.5.4	Η Τεχνητή Νοημοσύνη και η Μεταβαλλόμενη Φύση της Εργασίας	36
3.5.5	Πολιτιστικές και Κοινωνικές Επιπτώσεις της Τεχνητής Νοημοσύνης.....	36
3.5.6	Μελλοντικές Προοπτικές για τη Συνύπαρξη Ανθρώπου – Τεχνητής Νοημοσύνης	37
3.6	Παγκόσμια Ηθική και Πολιτική για την Τεχνητή Νοημοσύνη.....	37
3.6.1	Αναγκαιότητα Παγκόσμιας Ηθικής στην Τεχνητή Νοημοσύνη.....	37
3.6.2	Προκλήσεις στη Διαμόρφωση Παγκόσμιων Πολιτικών Τεχνητής Νοημοσύνης.....	38
3.6.3	Σημασία της Διεθνούς Συνεργασίας.....	38
3.6.4	Ανάπτυξη Πολιτικών για την Τεχνητή Νοημοσύνη χωρίς Αποκλεισμούς και Πολιτισμικής Ευαισθησίας	38
3.6.5	Ο Ρόλος των Διεθνών Ρυθμιστικών Φορέων.....	39
3.6.6	Μελλοντικές Κατευθύνσεις για την Παγκόσμια Ηθική και Πολιτική της Τεχνητής Νοημοσύνης	39

4	Περιπτώσιολογικές Μελέτες στην Ηθική της Τεχνητής Νοημοσύνης.....	40
4.1	Εισαγωγή στις Μελέτες Περίπτωσης Ηθικών Διλημμάτων.....	40
4.2	Τεχνητή Νοημοσύνη στην Επιβολή του Νόμου	40
4.3	Αυτόνομες Οπλικές Τεχνολογίες	42
4.4	Τεχνητή Νοημοσύνη στην Ιατρική Διάγνωση	45
4.5	Διαχείριση Προσωπικών Δεδομένων και Ιδιωτικότητα.....	46
4.6	Συγκριτική Συζήτηση και Συμπεράσματα Κεφαλαίου	47
5	Αναδυόμενες Τάσεις και Μελλοντικές Κατευθύνσεις στην Ηθική της Τεχνητής Νοημοσύνης.....	49
5.1	Εισαγωγή στο μέλλον της ηθικής της Τεχνητής Νοημοσύνης	49
5.2	Νέες Εξελίξεις στην Τεχνητή Νοημοσύνη και οι Ηθικές τους Προκλήσεις.....	49
5.3	Ρυθμιστικά Πλαίσια και Πολιτικές	52
5.3.1	Οι Πρωτοβουλίες στην Ευρωπαϊκή Ένωση.....	52
5.3.2	Οι Κατευθυντήριες Γραμμές για Αξιοπίστη και Υπεύθυνη Τεχνητή Νοημοσύνη	53
5.3.3	Ο Ρόλος Διεθνών Πλαισίων.....	53
5.4	Ηθικά πρότυπα και τεχνολογίες υπεύθυνης Σχεδίασης και Ανάπτυξης Συστημάτων Τεχνητής Νοημοσύνης	55
5.4.1	Ηθικές Κατευθυντήριες Αρχές	55
5.4.2	Τεχνολογικά Εργαλεία για Υπεύθυνη Σχεδίαση και Ανάπτυξη Συστημάτων Τεχνητής Νοημοσύνης	55
5.4.3	Αυτορρύθμιση και Συνεργατικές Προσπάθειες.....	56
6	Μελέτη Περίπτωσης και Συγκριτική Ανάλυση Κύκλων Ζωής Λογισμικού, Λειτουργικών Συστημάτων και Συστημάτων Τεχνητής Νοημοσύνης.....	57
6.1	Σκοπός, Ερευνητική Στόχευση και Ορισμός της Μελέτης Περίπτωσης	57
6.2	Περίπτωση Α: Κύκλος Ζωής, Ρόλοι, Διαδικασία, Παραδοτέα.....	60
6.2.1	Έναρξη Έργου και Αποσαφήνιση Αναγκών: Ποιο Πρόβλημα Επιλύεται.....	60
6.2.2	Ρόλοι: Τι κάνει ο αναλυτής και τι κάνει ο developer στην πράξη.....	61
6.2.3	Διαδικασία Ανάπτυξης: Από Απαιτήσεις σε Παραγωγή.....	61
6.2.4	Παραδοτέα: Τι «Παράγεται» σε Κάθε Στάδιο.....	62
6.2.5	Ένα μικρό παράδειγμα μέσα στο case: “Αίτημα με έγκριση”	62
6.3	Περίπτωση Β: Κύκλος ζωής συστήματος Τεχνητής Νοημοσύνης ρόλοι, διαδικασίες και παραδοτέα.....	63
6.3.1	Τι Αλλάζει όταν Εισάγεται Τεχνητή Νοημοσύνη στο Ίδιο Πληροφοριακό Σύστημα.....	64
6.3.2	Ο Ρόλος του Αναλυτή στη Σχεδίαση Συστημάτων Τεχνητής Νοημοσύνης.....	64
6.3.3	Ο Ρόλος του developer/ML Engineer στην Υλοποίηση Συστημάτων ΤΝ	64
6.3.4	Διαδικασία Σχεδίασης και Ανάπτυξης Συστημάτων ΤΝ: Από τον Ορισμό του Προβλήματος έως την Τελική Αξιολόγηση του Μοντέλου.....	65
6.3.5	Παραδοτέα κατά την Υλοποίηση Συστημάτων Τεχνητής Νοημοσύνης	65
6.3.6	Τι Αναδεικνύει η Περίπτωση Β	66
6.4	Συγκριτική Ανάλυση των Δυο Μελετών Περίπτωσης(Ανάπτυξη Λογισμικού έναντι Ανάπτυξης Λογισμικού με Χρήση Τεχνητής Νοημοσύνης).....	68
6.5	Κύκλος ζωής Λειτουργικού Συστήματος.....	70

6.5.1	Σχεδίαση και Αρχική Ανάπτυξη	70
6.5.2	Κυκλοφορία και Παραγωγική Λειτουργία	71
6.5.3	Ενημερώσεις, Διορθώσεις και Διαχείριση Ευπαθειών	71
6.5.4	Μακροχρόνια Συντήρηση και Σταθερότητα	71
6.5.5	Τέλος Κύκλου Ζωής και Απόσυρση (End-Of-Life)	72
6.5.6	Τι μας Δείχνει ο Κύκλος Ζωής του Λειτουργικού Συστήματος	72
6.6	Κύκλος Ζωής Μοντέλου Συστημάτων Τεχνητής Νοημοσύνης (Model Lifecycle)	73
6.6.1	Διατύπωση Προβλήματος και Σκοπού του Μοντέλου	73
6.6.2	Συλλογή, Επιλογή και Τεκμηρίωση Δεδομένων	74
6.6.3	Εκπαίδευση και Αξιολόγηση του Μοντέλου	74
6.6.4	Ενσωμάτωση και Παραγωγική Λειτουργία	74
6.6.5	Συνεχής Παρακολούθηση, Υποβάθμιση και Επανεκπαίδευση	75
6.6.6	Απόσυρση, Αντικατάσταση και Ηθικό Τέλος Κύκλου Ζωής	75
6.6.7	Τι Αποκαλύπτει ο Κύκλος Ζωής της Τεχνητής Νοημοσύνης	75
6.7	Ηθικές και Λειτουργικές Απαιτήσεις Συστημάτων Τεχνητής Νοημοσύνης	76
6.7.1	Διαφάνεια (transparency): Από την Τεχνική Ορατότητα στη Θεσμική Κατανόηση	76
6.7.2	Μεροληψία (Bias): Δομικό Πρόβλημα και όχι Τεχνική Αστοχία	77
6.7.3	Ηθική εκ Σχεδιασμού (Ethics-By-Design): Από τις Αρχές στις Απαιτήσεις	78
6.7.4	Λειτουργικές Απαιτήσεις για Υπεύθυνη Επεξεργασία και Λήψη Αποφάσεων	78
6.7.5	Συνολική Αποτίμηση	79
6.8	Συμπεράσματα Κεφαλαίου	79
7	Σύνθεση και Συμπεράσματα	81
7.1	Ανακεφαλαίωση των Βασικών Σημείων	81
7.2	Συμπεράσματα για τη Σημασία της Ηθικής στην Ανάπτυξη της Τεχνητής Νοημοσύνης	83
7.2.1	Αποφυγή Βλαβών και Άδικων Αποτελεσμάτων	83
7.2.2	Ενίσχυση και διατήρηση της εμπιστοσύνης του κοινού	84
7.2.3	Προώθηση του δημόσιου συμφέροντος και του κοινού καλού	84
7.3	Τελικές Παρατηρήσεις και Προτάσεις για Υπεύθυνη Αξιοποίηση της Τεχνητής Νοημοσύνης ..	85
	Βιβλιογραφία	88

Λίστα Σχημάτων

Διάγραμμα 1: «Διάγραμμα Νομικών Πλαισίων και Κανονιστικής Συμμόρφωσης στην Τεχνητή Νοημοσύνη».....	29
Διάγραμμα 2. «Διάγραμμα Σύγκρισης Διαδικασιών Ανάπτυξης Λογισμικού και Ανάπτυξης Συστημάτων Τεχνητής Νοημοσύνης».....	67

Λίστα Πινάκων

Πίνακας 1: Κύριες πτυχές των αυτονόμων συστημάτων.....	12
Πίνακας 2: Αλγοριθμική μεροληψία και δικαιοσύνη στην τεχνητή νοημοσύνη.....	24
Πίνακας 3. Η χρήση της Τεχνητής Νοημοσύνης στην επιβολή του νόμου.....	43
Πίνακας 4: Ζητήματα Ηθικής, Νομικής και Διεθνούς Διάστασης των Αυτόνομων Οπλικών Τεχνολογιών.....	45
Πίνακας 5: Νέες Εξελίξεις στην Τεχνητή Νοημοσύνη και Κύριες Ηθικές Προκλήσεις.....	53
Πίνακας 6: Ρυθμιστικά Πλαίσια, Πολιτικές και Διεθνείς Πρωτοβουλίες για την Υπεύθυνη Ανάπτυξη της Τεχνητής Νοημοσύνης.....	55

Ακρωνύμια

ACM	Association for Computing Machinery
AGI	Artificial General Intelligence
GDPR	General Data Protection Regulation
IEEE	Institute of Electrical and Electronics Engineers
IoT	Διαδίκτυο των Πραγμάτων
ISO	International Organization for Standardization
ML engineer	Machine Learning Engineer
OECD	Organization for Economic Co-operation and Development
OHE	Οργανισμός Ηνωμένων Εθνών
TN	Τεχνητή Νοημοσύνη
UNESCO	United Nations Educational, Scientific and Cultural Organization
UX	User Experience
XAI	Explainable AI
MKO	Μη Κυβερνητική Οργάνωση
ΟΟΣΑ	Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει τη σχέση μεταξύ της τεχνητής νοημοσύνης (TN) και της ηθικής, εστιάζοντας στις προκλήσεις και τις επιπτώσεις των εξελίξεων της TN στην κοινωνία. Το έγγραφο ξεκινά με μια ιστορική επισκόπηση της ανάπτυξης της TN, από τα θεωρητικά θεμέλια που έθεσαν επιστήμονες του 20ού αιώνα, όπως ο Alan Turing, μέσω των πρώτων επιτυχιών των δεκαετιών του 1960 και 1970, μέχρι τη στασιμότητα κατά τη διάρκεια του λεγόμενου «χειμώνα της TN». Ιδιαίτερη έμφαση δίνεται στις επαναστάσεις που προκάλεσε η μηχανική μάθηση και η μετάβαση στα σύγχρονα αυτόνομα συστήματα.

Στη συνέχεια, αναλύονται οι ηθικοί προβληματισμοί που προέκυψαν μέσω της ανάπτυξης αυτών των τεχνολογιών, συμπεριλαμβανομένης της αλγοριθμικής προκατάληψης, της διαφάνειας στη λήψη αποφάσεων και της διαχείρισης δεδομένων. Συζητείται, επίσης, ο ρόλος της τεχνητής νοημοσύνης στην επεξεργασία τεράστιων ποσοτήτων προσωπικών δεδομένων και οι κίνδυνοι για την ιδιωτικότητα και τα δικαιώματα των πολιτών. Παράλληλα, εξετάζεται συγκριτικά η ανάπτυξη κλασικού λογισμικού και λογισμικού με τη χρήση της TN, αναδεικνύοντας διαφορές στους ρόλους, τις διαδικασίες και κυρίως στην αξιοπιστία, η οποία στην TN εξαρτάται από δεδομένα και συνεχή αξιολόγηση, ενισχύοντας την ανάγκη για διαχείριση κινδύνων. Το σημαντικό μέρος της διπλωματικής εργασίας αφιερώνεται σε μια συζήτηση για τα αυτόνομα συστήματα, όπως τα αυτοκινούμενα αυτοκίνητα, και τα ζητήματα λογοδοσίας που προκύπτουν όταν οι μηχανές λαμβάνουν αποφάσεις χωρίς ανθρώπινη παρέμβαση.

Η εργασία ενσωματώνει μια κριτική επισκόπηση της βιβλιογραφίας σχετικά με τα ηθικά ζητήματα στην TN και παρέχει εφαρμογές σε πραγματικές εφαρμογές που φέρνουν αυτά τα διλήμματα στο προσκήνιο. Τέλος, προτείνονται κατευθυντήριες γραμμές για την υπεύθυνη ανάπτυξη της τεχνητής νοημοσύνης με στόχο την προώθηση της καινοτομίας με σεβασμό στις κοινωνικές αξίες, τα ανθρώπινα δικαιώματα και τη δικαιοσύνη.

Η συνολική δομή του παρόντος εγγράφου επιτρέπει στον αναγνώστη να κατανοήσει τόσο τις τεχνικές όσο και τις ηθικές πτυχές της TN και παρέχει ένα πλαίσιο για την κριτική ανάλυση και διαχείριση των ηθικών προκλήσεων που παρουσιάζει αυτή η ταχέως εξελισσόμενη τεχνολογία.

Abstract

This paper examines the relationship between artificial intelligence (AI) and ethics, focusing on the challenges and implications arising from the development of AI in society. The paper begins with a historical overview of the evolution of AI, from the theoretical foundations laid by 20th century scientists such as Alan Turing, through the early successes of the 1960s and 1970s, to the stagnation during the so-called “AI winter.” Particular emphasis is placed on the revolution brought about by machine learning and the move to modern autonomous systems.

It then analyzes the ethical concerns that have emerged from these technological advances, including algorithmic bias, transparency in decision-making, and data management. In addition, the role of AI in processing vast amounts of personal data and the risks to privacy and citizens' rights will be examined. At the same time, the comparative development of classical software and AI systems is examined, highlighting differences in roles, processes, and especially in reliability, which in AI depends on data and continuous evaluation, thereby reinforcing the need for risk management. The bulk of the paper is devoted to a discussion of autonomous systems such as self-driving cars and the accountability issues that arise when machines make decisions without human intervention.

The study provides a critical literature review of the ethical issues surrounding AI and offers examples that highlight these dilemmas. Finally, we propose guidelines for responsible AI development, aiming to promote innovation while respecting social values, human rights, and equity.

The overall structure of this paper allows the reader to understand both the technical and ethical aspects of AI and provides a framework for critically analyzing and managing the ethical challenges posed by this rapidly evolving technology.

1

Εισαγωγή

1.1 Εισαγωγή στην Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη (ΤΝ) έχει αναδειχθεί ως μια μεταμορφωτική δύναμη στον σύγχρονο κόσμο, επηρεάζοντας βαθιά διάφορες πτυχές της κοινωνίας και της καθημερινής ζωής. Αυτή η τεχνολογική επανάσταση φέρνει μαζί της μια σειρά από ηθικές προκλήσεις και προβληματισμούς που είναι κρίσιμο να αντιμετωπιστούν. Ο πυρήνας αυτής της διπλωματικής εργασίας είναι να διερευνήσει τη διασταύρωση της τεχνητής νοημοσύνης και της ηθικής, εμβαθύνοντας στο πώς οι εξελίξεις της τεχνητής νοημοσύνης αμφισβητούν τους υπάρχοντες ηθικούς κανόνες και απαιτούν μια επαναξιολόγηση των ηθικών αρχών στο πλαίσιο της ψηφιακής εξέλιξης.

Η σφαίρα της τεχνητής νοημοσύνης περιλαμβάνει ένα ευρύ φάσμα τεχνολογιών, από απλούς αλγόριθμους που εκτελούν συγκεκριμένες εργασίες έως πολύπλοκα συστήματα ικανά να μάθουν και να προσαρμοστούν. Καθώς τα συστήματα ΤΝ μιμούνται όλο και περισσότερο τις ανθρώπινες γνωστικές λειτουργίες, το όριο μεταξύ των δυνατοτήτων της μηχανής και της ανθρώπινης νοημοσύνης γίνεται θολή, εγείροντας σημαντικά ηθικά ερωτήματα (Russell & Norvig, 2010). Αυτή η διπλωματική εργασία εξετάζει αυτά τα ερωτήματα, εστιάζοντας στις ηθικές συνέπειες της αυτονομίας της τεχνητής νοημοσύνης, των διαδικασιών λήψης αποφάσεων και της ενσωμάτωσής της στον ιστό της κοινωνίας.

Η σημασία της ηθικής στην τεχνητή νοημοσύνη υπογραμμίζεται από τον πιθανό αντίκτυπο αυτών των τεχνολογιών στις ανθρώπινες ζωές. Τα συστήματα τεχνητής νοημοσύνης γίνονται αναπόσπαστα σε τομείς όπως η υγειονομική περίθαλψη, όπου βοηθούν στη διάγνωση ασθενειών και οι μεταφορές, όπου τροφοδοτούν αυτόνομα οχήματα. Αυτές οι εφαρμογές αποφέρουν τεράστια οφέλη, αλλά θέτουν επίσης κινδύνους και ηθικά διλήμματα, όπως ανησυχίες σχετικά με το απόρρητο, την ασφάλεια και τη λογοδοσία (Crawford & Calo, 2016· Eubanks, 2018). Για

παράδειγμα, ο ρόλος της τεχνητής νοημοσύνης στην επεξεργασία και την επιτήρηση δεδομένων εγείρει ερωτήματα σχετικά με τα δικαιώματα απορρήτου και την ηθική χρήση των προσωπικών πληροφοριών (O'Neil, 2017).

Επιπλέον, η ικανότητα του TN να μαθαίνει και να προσαρμόζεται μέσω αλγορίθμων έχει φέρει στο φως ζητήματα μεροληψίας και δικαιοσύνης. Οι αλγόριθμοι, εκπαιδευμένοι σε ιστορικά δεδομένα, μπορούν ακούσια να διαιωνίσουν και να ενισχύσουν τις κοινωνικές προκαταλήψεις, οδηγώντας σε μεροληπτικά αποτελέσματα σε τομείς όπως η πρόσληψη θέσεων εργασίας, οι εγκρίσεις δανείων και η ποινική καταδίκη (Caliskan, Bryson & Narayanan, 2017; Sullins, 2012). Αυτό υπογραμμίζει την ηθική επιταγή να διασφαλιστεί ότι τα συστήματα τεχνητής νοημοσύνης σχεδιάζονται και εφαρμόζονται με έντονη επίγνωση των πιθανών προκαταλήψεων και των επιπτώσεών τους σε διαφορετικά τμήματα της κοινωνίας.

Επιπλέον, η ταχεία πρόοδος της τεχνητής νοημοσύνης αμφισβητεί την παραδοσιακή ανθρωποκεντρική προοπτική της ηθικής. Η ανάπτυξη αυτόνομων συστημάτων ικανών να λαμβάνουν αποφάσεις χωρίς ανθρώπινη παρέμβαση εγείρει ερωτήματα σχετικά με την υπευθυνότητα και τη λογοδοσία. Αυτό οδηγεί σε ένα «κενό ευθύνης» όπου καθίσταται δύσκολο να αποδοθεί ευθύνη για τις ενέργειες των συστημάτων τεχνητής νοημοσύνης, ειδικά όταν ενεργούν με απρόβλεπτους τρόπους (Matthias, 2004; Johnson & Verdicchio, 2019).

Συνοπτικά, η εμφάνιση της τεχνητής νοημοσύνης θέτει βαθιές ηθικές προκλήσεις που απαιτούν προσεκτική εξέταση και προληπτική διαχείριση. Αυτή η διπλωματική εργασία στοχεύει να συμβάλει στη συζήτηση για την ηθική της τεχνολογίας αναλύοντας κριτικά πώς η τεχνητή νοημοσύνη προκαλεί και αναδιαμορφώνει τους ηθικούς κανόνες και προτείνοντας ένα πλαίσιο για την πλοήγηση σε αυτές τις προκλήσεις. Η ίδια η τεχνητή νοημοσύνη δεν διαθέτει ηθική κρίση, οι ηθικές αξίες προέρχονται από τους ανθρώπους που τη σχεδιάζουν, την αναπτύσσουν και τη χρησιμοποιούν. Μέσα από μια ολοκληρωμένη ανασκόπηση της υπάρχουσας βιβλιογραφίας, συμπεριλαμβανομένων φιλοσοφικών συζητήσεων, εγγράφων πολιτικής και πρακτικών περιπτώσεων μελετών, Αυτή η διπλωματική εργασία επιδιώκει να παρέχει μια διαφοροποιημένη κατανόηση του ηθικού τοπίου της τεχνητής νοημοσύνης και να προσφέρει πληροφορίες για το πώς μπορούμε να διασφαλίσουμε την ηθική ανάπτυξη και ανάπτυξη τεχνολογιών τεχνητής νοημοσύνης για τη βελτίωση της κοινωνίας.

1.2 Η Σημασία του Θέματος

Οι ηθικές διαστάσεις της Τεχνητής Νοημοσύνης (TN) δεν είναι απλώς ακαδημαϊκά ερωτήματα αλλά πιεστικές ανησυχίες που έχουν σημαντικές επιπτώσεις στην κοινωνία. Καθώς τα συστήματα τεχνητής νοημοσύνης γίνονται πιο διαδεδομένα και εξελιγμένα, επηρεάζουν όλο και περισσότερο διάφορες πτυχές της ζωής μας, από καθημερινές εργασίες έως κρίσιμες αποφάσεις στην υγειονομική περίθαλψη, την επιβολή του νόμου και τις χρηματοοικονομικές διαδικασίες. Η σημασία της μελέτης της ηθικής της τεχνητής νοημοσύνης έγκειται στη δυνατότητά της να διαμορφώνει κοινωνικούς κανόνες, να επηρεάζει τη χάραξη πολιτικής και να επαναπροσδιορίζει την ανθρώπινη αλληλεπίδραση με την τεχνολογία.

Η διάχυτη επιρροή της τεχνητής νοημοσύνης στη σύγχρονη κοινωνία δεν μπορεί να υποτιμηθεί. Οι εφαρμογές της τεχνητής νοημοσύνης κυμαίνονται από τον απλό αυτοματισμό έως τις πολύπλοκες διαδικασίες λήψης αποφάσεων που κάποτε ήταν αποκλειστικός τομέας των ανθρώπων. Αυτή η αλλαγή εγείρει κρίσιμα ηθικά ερωτήματα, ιδίως όσον αφορά τη λογοδοσία, τη διαφάνεια και τη δικαιοσύνη. Τα συστήματα TN, καθοδηγούμενα από αλγόριθμους, λειτουργούν συχνά ως «μαύρα κουτιά», καθιστώντας δύσκολη την κατανόηση των διαδικασιών λήψης των αποφάσεων τους (O'Neil, 2017). Αυτή η έλλειψη διαφάνειας προκαλεί σημαντική ανησυχία, ειδικά όταν αυτά τα συστήματα λαμβάνουν αποφάσεις που επηρεάζουν τις ζωές και τα μέσα διαβίωσης των ανθρώπων.

Η ταχεία πρόοδος της τεχνητής νοημοσύνης φέρνει επίσης στο επίκεντρο τις ηθικές εκτιμήσεις που αφορούν τη μετατόπιση της ανθρώπινης εργασίας. Καθώς τα συστήματα τεχνητής νοημοσύνης γίνονται ικανά να εκτελούν εργασίες που παραδοσιακά εκτελούνται από ανθρώπους, υπάρχει αυξανόμενη ανησυχία για την απώλεια θέσεων εργασίας και τη μεταβαλλόμενη φύση της εργασίας (Crawford & Calo, 2016). Αυτή η μετάβαση παρουσιάζει ηθικές προκλήσεις σχετικά με την κοινωνική ευθύνη για τη διαχείριση αυτών των αλλαγών με τρόπο που ελαχιστοποιεί τη βλάβη και μεγιστοποιεί το όφελος.

Επιπλέον, η ανάπτυξη της τεχνητής νοημοσύνης μας προκαλεί να επανεξετάσουμε τα παραδοσιακά ηθικά πλαίσια. Οι τεχνολογίες τεχνητής νοημοσύνης, ιδιαίτερα εκείνες που παρουσιάζουν μορφές αυτονομίας και αυτομάθησης, πιέζουν τα όρια των υφιστάμενων ηθικών θεωριών και αρχών (Wallach & Allen, 2008; Coeckelbergh, 2020). Αυτό απαιτεί επαναξιολόγηση και πιθανή προσαρμογή αυτών των πλαισίων για την κατάλληλη αντιμετώπιση των μοναδικών προκλήσεων που θέτει η τεχνητή νοημοσύνη.

Συνοπτικά, η μελέτη της ηθικής της τεχνητής νοημοσύνης είναι ζωτικής σημασίας λόγω του εκτεταμένου αντίκτυπου της τεχνολογίας στην κοινωνία. Εξετάζοντας τις ηθικές επιπτώσεις της

τεχνητής νοημοσύνης, μπορούμε να καθοδηγήσουμε την ανάπτυξή της σε μια κατεύθυνση που ευθυγραμμίζεται με τις ανθρώπινες αξίες και τους κοινωνικούς στόχους. Αυτή η διπλωματική εργασία στοχεύει να συμβάλει σε αυτόν τον κρίσιμο τομέα μελέτης, προσφέροντας μια ολοκληρωμένη ανάλυση των ηθικών προκλήσεων που θέτει η τεχνητή νοημοσύνη και προτείνοντας πλαίσια για την αντιμετώπιση αυτών των προκλήσεων με υπεύθυνο και δίκαιο τρόπο.

1.3 Σκοπός και Αντικείμενο της Εργασίας

Αυτή η διπλωματική εργασία υποστηρίζει ότι η ταχεία πρόοδος και ενσωμάτωση της Τεχνητής Νοημοσύνης (TN) σε διάφορους κοινωνικούς τομείς απαιτεί μια κριτική εξέταση των ηθικών συνεπειών της. Στόχος της είναι να διερευνήσει τις πολύπλευρες ηθικές προκλήσεις που θέτει η τεχνητή νοημοσύνη, συμπεριλαμβανομένων ζητημάτων αυτονομίας, μεροληψίας, διαφάνειας και λογοδοσίας. Διεξάγοντας μια ανασκόπηση της υπάρχουσας βιβλιογραφίας, αυτή η έρευνα επιδιώκει να συνθέσει διαφορετικές προοπτικές σχετικά με την ηθική της τεχνητής νοημοσύνης και να προτείνει ένα πλαίσιο για υπεύθυνη και ισορροπημένη ανάπτυξη τεχνητής νοημοσύνης. Ο στόχος είναι να συμβάλει σε έναν ισορροπημένο λόγο που καθοδηγεί την τεχνολογική καινοτομία και διασφαλίζει την ευθυγράμμιση της τεχνητής νοημοσύνης με τις ανθρώπινες αξίες και την κοινωνική ευημερία.

1.4 Επισκόπηση Δομής της Διπλωματικής Εργασίας

Αυτή η διπλωματική εργασία είναι δομημένη για να παρέχει μια ολοκληρωμένη διερεύνηση των ηθικών επιπτώσεων της Τεχνητής Νοημοσύνης (TN), και να αποσαφηνίσει συστηματικά τις πολυπλοκότητες αυτού του ταχέως εξελισσόμενου τομέα. Μετά από αυτήν την εισαγωγή, η πτυχιακή εκτυλίσσεται ως εξής:

Κεφάλαιο 1: Μεθοδολογία της Βιβλιογραφικής Ανασκόπησης - Αυτό το κεφάλαιο περιγράφει τις μεθόδους έρευνας που χρησιμοποιήθηκαν στη διπλωματική εργασία, συμπεριλαμβανομένων των κριτηρίων για την επιλογή της βιβλιογραφίας και της προσέγγισης που ακολουθήθηκε για την ανάλυση και τη σύνθεση του ποικίλου φάσματος των πηγών. Θέτει τα θεμέλια για μια αυστηρή και συστηματική εξερεύνηση του θέματος.

Κεφάλαιο 2: Ιστορική επισκόπηση της τεχνητής νοημοσύνης και των ηθικών ζητημάτων της. Εδώ, παρακολουθούμε την εξέλιξη της τεχνητής νοημοσύνης από την έναρξή της έως την τρέχουσα κατάστασή της, επισημαίνοντας βασικά ηθικά ζητήματα που έχουν προκύψει με την πάροδο του

χρόνου. Αυτό το ιστορικό πλαίσιο παρέχει ένα υπόβαθρο για την κατανόηση των τρεχουσών ηθικών συζητήσεων στην τεχνητή νοημοσύνη.

Κεφάλαιο 3: Βασικά θέματα στα ηθικά ζητήματα της τεχνητής νοημοσύνης. Αυτό το κεφάλαιο εμβαθύνει στα βασικά ηθικά θέματα που εγείρονται με τη χρήση της τεχνητής νοημοσύνης, όπως η αλγοριθμική μεροληψία, η αυτονομία, η διαφάνεια και ο αντίκτυπος της τεχνητής νοημοσύνης στο απόρρητο και την απασχόληση. Κάθε θέμα εξετάζεται κριτικά, αντλώντας γνώσεις από σημαντικές εργασίες στους τομείς της υγειονομικής περίθαλψης, των χρηματοοικονομικών και της εκπαίδευσης.

Κεφάλαιο 4: Μελέτες περιπτώσεων στη δεοντολογία της τεχνητής νοημοσύνης.

Κεφάλαιο 5: Αναδυόμενες τάσεις και μελλοντικές κατευθύνσεις στην ηθική της τεχνητής νοημοσύνης.

Κεφάλαιο 6: Μελέτη Περίπτωσης και Συγκριτική Ανάλυση Κύκλων Ανάπτυξης — Λογισμικό, Λειτουργικά Συστήματα και Τεχνητή Νοημοσύνη.

Κεφάλαιο 7: Σύνθεση και Συμπεράσματα.

Εξετάζοντας συστηματικά την ηθική της τεχνητής νοημοσύνης μέσω αυτής της δομημένης προσέγγισης, η πτυχιακή στοχεύει να παρέχει μια διαφοροποιημένη κατανόηση των ηθικών προκλήσεων που θέτει η τεχνητή νοημοσύνη και να συνεισφέρει πολύτιμες γνώσεις στον συνεχιζόμενο λόγο σε αυτόν τον κρίσιμο τομέα μελέτης.

1.5 Μεθοδολογία της Βιβλιογραφικής Ανασκόπησης

1.5.1 Επισκόπηση

Η μεθοδολογία αυτής της βιβλιογραφικής ανασκόπησης είναι κεντρική στη διπλωματική εργασία, καθώς περιγράφει τη συστηματική προσέγγιση που ακολουθείται για τη συγκέντρωση, ανάλυση και σύνθεση του τεράστιου όγκου της βιβλιογραφίας σχετικά με την ηθική της Τεχνητής Νοημοσύνης (TN). Αυτό το κεφάλαιο περιγράφει λεπτομερώς τις διαδικασίες που εμπλέκονται στην επιλογή της σχετικής βιβλιογραφίας, τα κριτήρια για ένταξη και αποκλεισμό και τις αναλυτικές μεθόδους που χρησιμοποιούνται για να εξασφαλιστεί μια ολοκληρωμένη και ισορροπημένη ανασκόπηση.

1.5.2 Στρατηγική Αναζήτηση Βιβλιογραφίας

Η αναζήτηση της βιβλιογραφίας ξεκίνησε με τον εντοπισμό βασικών βάσεων δεδομένων και ψηφιακών βιβλιοθηκών, συμπεριλαμβανομένων των IEEE Xplore, JSTOR και Google Scholar. Οι λέξεις-κλειδιά και οι φράσεις που χρησιμοποιήθηκαν στην αναζήτηση περιλάμβαναν “Ηθική Τεχνητής Νοημοσύνης”, “Τεχνητή Νοημοσύνη και Κοινωνία”, “Αλγοριθμική προκατάληψη”, “Αυτονομία Τεχνητής Νοημοσύνης” και “Απόρρητο δεδομένων στη Τεχνητή Νοημοσύνη”. Η αναζήτηση δεν περιορίστηκε σε ακαδημαϊκά περιοδικά, αλλά περιλάμβανε επίσης πρακτικά συνεδρίων, βιβλία, κυβερνητικές εκθέσεις και αξιόπιστους διαδικτυακούς πόρους για να καταγράψει το εύρος της συζήτησης σχετικά με την ηθική της τεχνητής νοημοσύνης.

1.5.3 Κριτήρια Επιλογής

Η επιλογή της βιβλιογραφίας βασίστηκε σε διάφορα κριτήρια:

- **Συνάφεια:** Το υλικό που απαιτείται για την άμεση αντιμετώπιση της τεχνητής νοημοσύνης και τις ηθικές της επιπτώσεις.
- **Αξιοπιστία:** Προτιμήθηκαν άρθρα με κριτές, βιβλία που εκδόθηκαν από έγκριτους εκδότες και αναφορές από αναγνωρισμένα ιδρύματα.
- **Πρόσφατες μελέτες :** Δεδομένης της ταχέως εξελισσόμενης φύσης της τεχνητής νοημοσύνης, προτιμήθηκαν δημοσιεύσεις από το 2010.
- **Ποικιλομορφία απόψεων:** Για να εξασφαλιστεί μια ισορροπημένη άποψη, συμπεριλήφθηκε βιβλιογραφία που περιλάμβανε διάφορες απόψεις, συμπεριλαμβανομένων τεχνολογικών, φιλοσοφικών, κοινωνιολογικών και νομικών προοπτικών.

1.5.4 Κριτήρια Εξαίρεσης

Οι εργασίες εξαιρούνταν εάν:

- Δεν αφορούσε άμεσα την ηθική της τεχνητής νοημοσύνης.
- Δεν προέρχονταν από αξιόπιστες πηγές ή κανάλια με αξιολογήσεις από ομότιμους.
- Ήταν ξεπερασμένα, προσφέροντας απαρχαιωμένες προοπτικές για την τεχνολογία και την ηθική της τεχνητής νοημοσύνης.

1.5.5 Αναλυτική Προσέγγιση

Η ανάλυση της επιλεγμένης βιβλιογραφίας περιλάμβανε διάφορα στάδια:

Θεματική Ανάλυση: Η βιβλιογραφία κατηγοριοποιήθηκε σε θέματα (π.χ. αυτονομία, μεροληψία, διαφάνεια). Αυτή η θεματική προσέγγιση διευκόλυνε μια δομημένη εξερεύνηση των ηθικών διαστάσεων της τεχνητής νοημοσύνης.

Κριτική αξιολόγηση: Κάθε εργασία αξιολογήθηκε κριτικά για τα επιχειρήματά της, τα στοιχεία και τα συμπεράσματά της. Αξιολογήθηκαν τα δυνατά σημεία, οι αδυναμίες και η συμβολή κάθε μελέτης στον τομέα της ηθικής της τεχνητής νοημοσύνης.

Σύνθεση: Βασικά ευρήματα και επιχειρήματα σε όλη τη βιβλιογραφία συντέθηκαν για να παρέχουν μια συνεκτική επισκόπηση της τρέχουσας κατάστασης της ηθικής της τεχνητής νοημοσύνης. Αυτή η σύνθεση είχε ως στόχο να εντοπίσει κοινά σημεία, διαφορές και αναδυόμενες τάσεις στον λόγο.

1.5.6 Περιορισμοί

Αυτή η ανασκόπηση αναγνωρίζει ορισμένους περιορισμούς:

- Η τεράστια και διεπιστημονική φύση της ηθικής της τεχνητής νοημοσύνης σημαίνει ότι δεν θα μπορούσε να καλυφθεί εξαντλητικά όλη η σχετική βιβλιογραφία.
- Η ταχεία ανάπτυξη της τεχνητής νοημοσύνης μπορεί να οδηγήσει στην εμφάνιση νέων ηθικών ζητημάτων που δεν καλύπτονται στην υπάρχουσα βιβλιογραφία.
- Η μεθοδολογία που περιγράφεται σε αυτό το κεφάλαιο παρείχε μια δομημένη και αυστηρή προσέγγιση για τη διερεύνηση των ηθικών επιπτώσεων της τεχνητής νοημοσύνης. Επιλέγοντας προσεκτικά και αναλύοντας κριτικά ένα ευρύ φάσμα βιβλιογραφίας, αυτή η ανασκόπηση στοχεύει να προσφέρει μια ολοκληρωμένη και ισορροπημένη κατανόηση των ηθικών προκλήσεων που θέτει η τεχνητή νοημοσύνη, θέτοντας το έδαφος για τη λεπτομερή εξερεύνηση στα επόμενα κεφάλαια.

2

Ιστορική Επισκόπηση της Τεχνητής Νοημοσύνης και των Ηθικών Ζητημάτων της

2.1 Εισαγωγή

Αυτό το κεφάλαιο παρέχει μια ιστορική επισκόπηση της Τεχνητής Νοημοσύνης χαρτογραφώντας την εξέλιξή της από τη θεωρητική έννοια στην τρέχουσα κατάστασή της, και την αντίστοιχη εμφάνιση ηθικών θεωρήσεων. Αυτή η ιστορική προοπτική είναι κρίσιμη για την κατανόηση του πλαισίου στο οποίο βρίσκονται οι τρέχουσες ηθικές συζητήσεις γύρω από την τεχνητή νοημοσύνη. Η γέννηση της τεχνητής νοημοσύνης (TN) ξεκίνησε στα μέσα του εικοστού αιώνα, όταν οι θεωρητικές βάσεις τέθηκαν από τον Άλαν Τούρινγκ, ο οποίος εισήγαγε το τεστ Τούρινγκ για την αξιολόγηση της νοημοσύνης των μηχανών το 1950. Η διάσκεψη του 1956 στο Ντάρτμουθ, στην οποία συμμετείχαν σημαντικές προσωπικότητες όπως ο Τζον Μακάρθι και ο Μάρβιν Μίνσκι, σηματοδότησε την έναρξη της ανάπτυξης του πεδίου της τεχνητής νοημοσύνης. Οι πρώτες επιτυχίες της τεχνητής νοημοσύνης στις δεκαετίες του 1960 και του 1970, που χαρακτηρίζονταν από τη συμβολική τεχνητή νοημοσύνη και τα λογικά συστήματα, δημιούργησαν αισιοδοξία για την ανάπτυξη νοημοσύνης ανθρώπινου επιπέδου, αλλά οι τεχνικοί περιορισμοί οδήγησαν σε μια περίοδο στασιμότητας γνωστή ως «χειμώνας της τεχνητής νοημοσύνης». Η αρχική έρευνα επικεντρώθηκε στη συμβολική τεχνητή νοημοσύνη, αλλά δυσκολεύτηκε να επιλύσει πιο σύνθετα προβλήματα του πραγματικού κόσμου. Σημαντική πρόοδος σημειώθηκε με τη μετάβαση στη μηχανική μάθηση και τη βαθιά μάθηση, η οποία επέτρεψε την ανάπτυξη πιο ισχυρών συστημάτων TN, όπως η αναγνώριση προτύπων και τα αυτόνομα συστήματα.

2.2 Ηθικές Προκλήσεις, Λογοδοσία και Ρυθμίσεις στην Τεχνητή Νοημοσύνη

Η ανάπτυξη της τεχνητής νοημοσύνης έφερε στο προσκήνιο σημαντικά ηθικά ζητήματα που αναπτύχθηκαν παράλληλα με τις τεχνολογικές εξελίξεις. Αρχικά, οι ηθικές συζητήσεις επικεντρώθηκαν κυρίως σε θεωρητικά ζητήματα, όπως η δυνατότητα της τεχνητής νοημοσύνης να ξεπεράσει την ανθρώπινη νοημοσύνη. Ωστόσο, καθώς η τεχνολογία άρχισε να χρησιμοποιείται ευρύτερα, οι ηθικές ανησυχίες μετατοπίστηκαν σε πιο πρακτικά και άμεσα ζητήματα.

Καθώς η χρήση της ΤΝ αυξήθηκε σε τομείς όπως η επεξεργασία δεδομένων και η λήψη αποφάσεων, προέκυψαν ανησυχίες σχετικά με την προστασία της ιδιωτικότητας, την ασφάλεια των δεδομένων και την πιθανότητα αλγοριθμικής μεροληψίας. Η χρήση συστημάτων ΤΝ σε αυτοκινούμενα αυτοκίνητα και μη επανδρωμένα αεροσκάφη έχει εγείρει ζητήματα ευθύνης, ιδίως σε περίπτωση ατυχημάτων, δημιουργώντας το λεγόμενο «κενό ευθύνης».

Η αλγοριθμική μεροληψία αποτελεί επίσης σημαντική ανησυχία, καθώς τα συστήματα ΤΝ τείνουν να αναπαράγουν τις προκαταλήψεις που υπάρχουν στα δεδομένα εκπαίδευσης, γεγονός που μπορεί να επηρεάσει αρνητικά τη λήψη αποφάσεων σε σημαντικούς τομείς όπως η δικαιοσύνη, η απασχόληση και η υγειονομική περίθαλψη. Ταυτόχρονα, οι αυξανόμενες δυνατότητες επιτήρησης της ΤΝ θέτουν σε κίνδυνο τα δικαιώματα της ιδιωτικής ζωής των ανθρώπων.

Τέλος, υπάρχει αυξανόμενη ανάγκη για την ανάπτυξη πολιτικών και ρυθμιστικών πλαισίων για την αντιμετώπιση ηθικών ζητημάτων. Οι προσπάθειες αυτές αποσκοπούν στη διασφάλιση ότι η ανάπτυξη της ΤΝ συνάδει με τις κοινωνικές αξίες, την ανθρώπινη αξιοπρέπεια και τα δικαιώματα.

2.3 Η Σύγχρονη Τεχνητή Νοημοσύνη: Εξελίξεις, Προκλήσεις και Ηθικά Διλήμματα

Η σύγχρονη εποχή της τεχνητής νοημοσύνης (ΤΝ) είναι μια εποχή ταχείας ανάπτυξης και σημαντικών αλλαγών, που καθοδηγείται από τη συνεχώς αυξανόμενη υπολογιστική ισχύ, τη διαθεσιμότητα τεράστιων ποσοτήτων δεδομένων και την πρόοδο των αλγοριθμικών τεχνικών. Χαρακτηριστικό αυτής της εποχής είναι η ανάγκη αντιμετώπισης των ηθικών προκλήσεων που θέτει η χρήση της ΤΝ σε διάφορους κλάδους, καθώς και η ενσωμάτωσή της στην καθημερινή ζωή. Η ραγδαία ανάπτυξη της υπολογιστικής ισχύος και της επιστήμης των δεδομένων οδήγησε σε ανανεωμένο ενδιαφέρον για την ΤΝ από τα τέλη του 20ού αιώνα. Η επεξεργασία τεράστιων ποσοτήτων δεδομένων επέτρεψε την ανάπτυξη εξελιγμένων αλγορίθμων που επιτρέπουν στην ΤΝ

να εκτελεί λειτουργίες που ήταν προηγουμένως αδύνατες, όπως η αναγνώριση προτύπων και η λήψη σύνθετων αποφάσεων.

Η τεχνητή νοημοσύνη έχει ήδη διεισδύσει στην καθημερινή μας ζωή μέσω εφαρμογών όπως οι υπηρεσίες συνεχούς ροής περιεχομένου (streaming), οι φωνητικοί βοηθοί και τα έξυπνα συστήματα στα έξυπνα κινητά τηλέφωνα (smartphones). Ωστόσο, αυτή η εκτεταμένη παρουσία εγείρει σοβαρές ανησυχίες για την προστασία της ιδιωτικής ζωής, καθώς πολλά από αυτά τα συστήματα συλλέγουν και επεξεργάζονται προσωπικά δεδομένα χωρίς τη ρητή συγκατάθεση του χρήστη.

Παράλληλα, η ανάπτυξη αυτόνομων συστημάτων, όπως τα αυτοκινούμενα αυτοκίνητα και τα ρομποτικά συστήματα, εγείρει νέα ηθικά ζητήματα, ιδίως όσον αφορά την ασφάλεια και την ευθύνη σε περίπτωση βλαβών ή ατυχημάτων, όπως υποστηρίζει ο Matthias (2004): όταν οι μηχανές λαμβάνουν αποφάσεις χωρίς ανθρώπινη παρέμβαση, αυτό το κενό ευθύνης γίνεται σοβαρό πρόβλημα.

Επιπλέον, η αλγοριθμική μεροληψία έχει καταστεί ένα από τα σημαντικότερα ηθικά ζητήματα. Τα συστήματα τεχνητής νοημοσύνης που εκπαιδεύονται σε ιστορικά δεδομένα συχνά αναπαράγουν τις ανθρώπινες προκαταλήψεις, οδηγώντας σε άδικα αποτελέσματα σε τομείς όπως η απασχόληση, η πίστωση και η επιβολή του νόμου. Οι Caliskan, Bryson και Narayanan (2017) υπογραμμίζουν την ανάγκη προσεκτικού σχεδιασμού τέτοιων συστημάτων.

Ταυτόχρονα, η χρήση της τεχνητής νοημοσύνης σε τεχνολογίες επιτήρησης, όπως η αναγνώριση προσώπου, εγείρει σοβαρά ερωτήματα σχετικά με την ιδιωτικότητα των δεδομένων και τις πολιτικές ελευθερίες. Οι εξελιγμένες τεχνολογίες επιτήρησης είναι χρήσιμες για την ασφάλεια και την επιβολή του νόμου, αλλά ενέχουν σημαντικούς κινδύνους για την προστασία της ιδιωτικής ζωής.

Αναπτύσσονται παγκόσμιες κατευθυντήριες γραμμές και πολιτικές για την αντιμετώπιση των ηθικών προκλήσεων της ΤΝ- οι Jobin, Iyenka και Wayena (2019) αναλύουν την παγκόσμια κατάσταση της ηθικής της ΤΝ και υπογραμμίζουν την ανάγκη για ένα σύστημα που θα διασφαλίζει την υπεύθυνη καινοτομία και θα προστατεύει τις κοινωνικές αξίες.

Καθώς η τεχνητή νοημοσύνη συνεχίζει να επεκτείνεται και να επηρεάζει όλες τις πτυχές της κοινωνίας, τα ηθικά ζητήματα γύρω από την ιδιωτικότητα, την προκατάληψη, την αυτονομία και τη λογοδοσία γίνονται όλο και πιο πολύπλοκα. Η ανάπτυξη μιας υπεύθυνης προσέγγισης για τη διαχείριση και την αξιοποίηση της τεχνητής νοημοσύνης είναι ζωτικής σημασίας και για να διασφαλιστεί ότι τα οφέλη της τεχνολογίας είναι συμβατά με τις κοινωνικές αξίες και τα ανθρώπινα δικαιώματα.

2.4 Η Άνοδος των Αυτόνομων Συστημάτων

Η άνοδος των αυτόνομων συστημάτων στη σύγχρονη εποχή της Τεχνητής Νοημοσύνης (TN) αντιπροσωπεύει ένα σημαντικό άλμα στις δυνατότητες της τεχνολογίας και τις επιπτώσεις της στην κοινωνία. Καθώς τα συστήματα τεχνητής νοημοσύνης έχουν εξελιχθεί για να λαμβάνουν αποφάσεις και να λειτουργούν με υψηλό βαθμό ανεξαρτησίας, έχουν επιφέρει ηθικές, νομικές και κοινωνικές προκλήσεις.

2.4.1 Ανάπτυξη Αυτόνομων Τεχνολογιών

Τα αυτόνομα συστήματα, όπως τα αυτοοδηγούμενα οχήματα και τα drones, είναι σχεδιασμένα να λειτουργούν χωρίς την ανάγκη συνεχούς ανθρώπινης παρέμβασης, αξιοποιώντας προηγμένους αλγόριθμους και τεχνικές μηχανικής μάθησης για να λαμβάνουν αποφάσεις αυτόνομα, με βάση τα δεδομένα που συλλέγουν σε πραγματικό χρόνο. Αυτά τα συστήματα είναι σχεδιασμένα να λειτουργούν χωρίς συνεχή ανθρώπινη καθοδήγηση, λαμβάνοντας αποφάσεις με βάση τους αλγόριθμους προγραμματισμού και μηχανικής μάθησης. Η πρόοδος στη μηχανική μάθηση, ιδιαίτερα σε τομείς όπως η βαθιά μάθηση και η ενισχυτική μάθηση, ήταν ζωτικής σημασίας σε αυτήν την εξέλιξη.

2.4.2 Προβλήματα Ασφάλειας και Αξιοπιστίας

Η ασφάλεια και η αξιοπιστία των αυτόνομων συστημάτων είναι σημαντικά ζητήματα, αλλά εξίσου σημαντικές είναι και οι ηθικές ανησυχίες. Στην περίπτωση των αυτοοδηγούμενων αυτοκινήτων, για παράδειγμα, προκύπτουν ερωτήματα σχετικά με τη λήψη αποφάσεων σε κρίσιμες καταστάσεις, όπως τα σενάρια αποφυγής ατυχημάτων. Αυτές οι ανησυχίες δεν αφορούν μόνο την τεχνική αξιοπιστία των συστημάτων αλλά και τις ηθικές αρχές που πρέπει να καθοδηγούν τις διαδικασίες λήψης αποφάσεων. Ο προγραμματισμός αυτών των συστημάτων περιλαμβάνει περίπλοκες ηθικές αποφάσεις, όπως η προτεραιότητα ορισμένων τύπων ασφάλειας έναντι άλλων, κάτι που έχει σημαντικές επιπτώσεις.

2.4.3 Μεροληψία και Διακρίσεις σε Αυτόνομα Συστήματα

Μια άλλη σημαντική ανησυχία είναι η πιθανότητα μεροληψίας και διάκρισης σε αυτόνομα συστήματα. Δεδομένου ότι αυτά τα συστήματα μαθαίνουν από δεδομένα, είναι επιρρεπή σε εγγενείς προκαταλήψεις που υπάρχουν στα σύνολα δεδομένων εκπαίδευσής τους. Οι Caliskan, Bryson και Narayanan (2017) δείχνουν πώς οι προκαταλήψεις μπορούν να ενσωματωθούν σε συστήματα τεχνητής νοημοσύνης, οδηγώντας σε μεροληπτικά αποτελέσματα. Αυτό είναι ιδιαίτερα ανησυχητικό στο πλαίσιο αυτόνομων συστημάτων που χρησιμοποιούνται σε κρίσιμους τομείς όπως

η επιβολή του νόμου, οι αποφάσεις δανεισμού από χρηματοπιστωτικά ιδρύματα και οι προσλήψεις εργαζομένων, όπου η μεροληπτική λήψη αποφάσεων μπορεί να έχει σημαντικές αρνητικές επιπτώσεις σε άτομα και ομάδες.

2.4.4 Επιπτώσεις Απορρήτου

Τα αυτόνομα συστήματα βασίζονται συχνά σε εκτεταμένες δυνατότητες συλλογής και επεξεργασίας δεδομένων, εγείροντας ανησυχίες σχετικά με το απόρρητο και την επιτήρηση. Η ικανότητα αυτών των συστημάτων να συλλέγουν και να αναλύουν προσωπικές πληροφορίες εγκυμονεί κινδύνους για το απόρρητο και την αυτονομία του ατόμου.

2.4.5 Παγκόσμια Ανταπόκριση και Ηθικές Κατευθυντήριες Γραμμές

Η άνοδος των αυτόνομων συστημάτων έχει προκαλέσει μια παγκόσμια ανταπόκριση όσον αφορά τις ηθικές κατευθυντήριες γραμμές και τα ρυθμιστικά πλαίσια. Οι Jobin, Ienca και Vayena (2019) διερευνούν το τοπίο των κατευθυντήριων γραμμών ηθικής τεχνητής νοημοσύνης, σημειώνοντας την αυξανόμενη έμφαση σε αρχές όπως η διαφάνεια, η λογοδοσία και η δικαιοσύνη στο σχεδιασμό και την ανάπτυξη αυτόνομων συστημάτων. Αυτές οι κατευθυντήριες γραμμές είναι ζωτικής σημασίας για τη διασφάλιση ότι τα αυτόνομα συστήματα αναπτύσσονται και χρησιμοποιούνται με τρόπο που ευθυγραμμίζονται με τις κοινωνικές αξίες και τους ηθικούς κανόνες.

Η άνοδος των αυτόνομων συστημάτων σηματοδοτεί μια κομβική στιγμή στην εξέλιξη της τεχνητής νοημοσύνης, που χαρακτηρίζεται από σημαντικές ηθικές, νομικές και κοινωνικές προκλήσεις. Η αυτονομία αυτών των συστημάτων απαιτεί μια επαναξιολόγηση των παραδοσιακών εννοιών της ευθύνης, της προκατάληψης, της ασφάλειας και της ιδιωτικής ζωής. Καθώς αυτές οι τεχνολογίες συνεχίζουν να προχωρούν και να ενσωματώνονται σε διάφορες πτυχές της ζωής, είναι επιτακτική ανάγκη να αντιμετωπιστούν αυτές οι προκλήσεις μέσω ισχυρών πλαισίων και πολιτικών ηθικής. Αυτό διασφαλίζει ότι η ανάπτυξη αυτόνομων συστημάτων συμβάλλει θετικά στην κοινωνία, ελαχιστοποιώντας τις πιθανές βλάβες, σεβόμενοι τις ανθρώπινες αξίες.

Πίνακας 1. Κύριες πτυχές των αυτόνομων συστημάτων

Θεματική Ενότητα	Κύρια Σημεία	Βασικές Προκλήσεις	Ενδεικτικές Πηγές
Ανάπτυξη Αυτόνομων Τεχνολογιών	Αυτοοδηγούμενα οχήματα, drones, ρομποτικές εφαρμογές	Ανάγκη αξιόπιστης και συνεχούς ενημέρωσης δεδομένων.	Caliskan, Bryson & Narayanan (2017) Rahwan κ.ά. (2019)

	λειτουργούν με υψηλό βαθμό ανεξαρτησίας. Αξιοποίηση αλγορίθμων μηχανικής μάθησης (ιδίως βαθιάς και ενισχυτικής) για λήψη αποφάσεων σε πραγματικό χρόνο.	Εξασφάλιση τεχνικής ευστάθειας και απόκρισης σε απρόβλεπτες καταστάσεις.	
Προβλήματα Ασφάλειας και Αξιοπιστίας	Ερωτήματα σχετικά με τη λήψη αποφάσεων σε κρίσιμες καταστάσεις (π.χ. αποφυγή ατυχημάτων). Η ανάγκη ενσωμάτωσης ηθικών αρχών στον προγραμματισμό (π.χ. ποιες ανθρώπινες ζωές προστατεύονται κατά προτεραιότητα).	Τεχνολογικές βλάβες ή αστοχίες που ενδέχεται να έχουν σοβαρές συνέπειες. Διχογνωμίες σχετικά με το ποιος φέρει την ευθύνη σε περίπτωση λάθους ή ατυχήματος (κατασκευαστής, προγραμματιστής, χρήστης).	Rahwan κ.ά. (2019)
Μεροληψία και Διακρίσεις σε Αυτόνομα Συστήματα	Οι αλγόριθμοι ενδέχεται να «κληρονομούν» προκαταλήψεις από τα δεδομένα εκπαίδευσης.	Ανάγκη για διαφάνεια στον τρόπο εκπαίδευσης και λειτουργίας των μοντέλων.	Caliskan, Bryson & Narayanan (2017)

	Μεροληπτικά αποτελέσματα σε κρίσιμους τομείς (επιβολή νόμου, αποφάσεις δανεισμού, προσλήψεις).	Κοινωνικές και νομικές επιπτώσεις όταν επηρεάζονται άτομα ή ευάλωτες ομάδες.	
Επιπτώσεις Απορρήτου	Η συλλογή και επεξεργασία μεγάλων όγκων δεδομένων από αυτόνομα συστήματα μπορεί να απειλήσει την ιδιωτικότητα. Η δυνατότητα «αδιάκοπης» επιτήρησης γεννά ανησυχίες για ενδεχόμενη κατάχρηση.	Θέσπιση σαφούς νομικού πλαισίου για τη χρήση και επεξεργασία προσωπικών δεδομένων. Διασφάλιση της αυτονομίας των πολιτών και προστασία από ανεπιθύμητη καταγραφή δραστηριοτήτων.	Rahwan κ.ά. (2019)
Παγκόσμια Ανταπόκριση και Ηθικές Κατευθυντήριες Γραμμές	Οι διεθνείς οργανισμοί και οι κυβερνήσεις διαμορφώνουν κατευθυντήριες γραμμές για την ηθική ανάπτυξη της ΤΝ. Έμφαση σε αξίες όπως διαφάνεια,	Ετερογένεια των νομικών και πολιτισμικών πλαισίων σε διεθνές επίπεδο. Δυσκολία επιβολής κοινών προτύπων σε πολυεθνικές εταιρείες τεχνολογίας.	Jobin, Ienca & Vayena (2019)

	λογοδοσία και δικαιοσύνη.		
--	---------------------------	--	--

2.5 Η Τεχνητή Νοημοσύνη στον Δημόσιο Λόγο

Η ενσωμάτωση της Τεχνητής Νοημοσύνης (TN) σε διάφορες πτυχές της καθημερινής ζωής όχι μόνο έχει φέρει επανάσταση στις τεχνολογικές δυνατότητες αλλά και έχει επηρεάσει σημαντικά τον δημόσιο διάλογο. Η συζήτηση γύρω από την τεχνητή νοημοσύνη έχει ξεπεράσει τους ακαδημαϊκούς και τεχνολογικούς κύκλους, και έγινε θέμα ευρέως διαδεδομένου δημόσιου ενδιαφέροντος κατά την τελευταία δεκαετία, κυρίως από την δεκαετία του 2010 και έπειτα . Η συζήτηση έχει απασχολήσει κυβερνήσεις, διεθνείς οργανισμούς (όπως τα Ηνωμένα Έθνη, η UNESCO, ο ΟΟΣΑ), ακαδημαϊκά ιδρύματα, τεχνολογικές εταιρείες (Google, Microsoft, OpenAI κλπ.), καθώς και ευρύτερες κοινωνικές και επαγγελματικές κοινότητες. Κομβικό ρόλο έχουν παίξει επίσης οι ειδικοί σε ζητήματα ηθικής και νομικής της TN. Τα αποτελέσματα της δημόσιας συζήτησης περιλαμβάνουν τη διαμόρφωση πολιτικών και κανονισμών, όπως ο Γενικός Κανονισμός για την Προστασία Δεδομένων (GDPR) στην Ευρωπαϊκή Ένωση, η χάραξη ηθικών κατευθυντήριων γραμμών για την ανάπτυξη και χρήση TN, καθώς και η αυξημένη επίγνωση των κοινωνικών, νομικών και ηθικών ζητημάτων που προκύπτουν από τη χρήση της τεχνητής νοημοσύνης. Η δημόσια συζήτηση έχει συμβάλει επίσης στην προώθηση της διαφάνειας και της υπευθυνότητας στον τομέα αυτό. Αυτή η αλλαγή αντανακλά την αυξανόμενη αναγνώριση του αντίκτυπου της τεχνητής νοημοσύνης στην κοινωνία και τις ηθικές, νομικές και κοινωνικές επιπτώσεις που συνοδεύουν την ανάπτυξη και τη χρήση της.

2.5.1 Η Αυξανόμενη Ευαισθητοποίηση του Κοινού για την Τεχνητή Νοημοσύνη

Καθώς οι τεχνολογίες τεχνητής νοημοσύνης έχουν γίνει πιο διαδεδομένες, η ευαισθητοποίηση του κοινού τόσο για τα οφέλη όσο και για τους πιθανούς κινδύνους τους έχει αυξηθεί. Αυτή η ευαισθητοποίηση οφείλεται στην κάλυψη των μέσων ενημέρωσης, στη δημοφιλή κουλτούρα και σε υποθέσεις υψηλού προφίλ που αφορούν την τεχνητή νοημοσύνη, όπως παραβιάσεις δεδομένων, κακή χρήση προσωπικών πληροφοριών και αμφιλεγόμενες χρήσεις στην παρακολούθηση και την επιβολή του νόμου. Το βιβλίο της O'Neil (2017) "Weapons of Math Destruction" υπογραμμίζει πώς η τεχνητή νοημοσύνη και τα μεγάλα δεδομένα μπορούν να διαιωνίσουν την ανισότητα, ενισχύοντας την ανησυχία του κοινού σχετικά με τη δικαιοσύνη και τη διαφάνεια αυτών των συστημάτων.

2.5.2 Συζητήσεις για την Προστασία της Ιδιωτικότητας και την Επιτήρηση

Ένας από τους πιο εξέχοντες τομείς ανησυχίας του κοινού σχετικά με την τεχνητή νοημοσύνη είναι το απόρρητο και η επιτήρηση. Η ικανότητα των συστημάτων τεχνητής νοημοσύνης να επεξεργάζονται τεράστιες ποσότητες προσωπικών δεδομένων έχει ως αποτέλεσμα τη διάβρωση του απορρήτου και την πιθανότητα κακής χρήσης. Η ηθική διαχείριση των δεδομένων, όπως συζητήθηκε από τους Rahwan και συνεργατών (2019) και Rumbold (2023), έχει γίνει ένα κρίσιμο ζήτημα στον δημόσιο διάλογο, τονίζοντας την ανάγκη για αυστηρή προστασία της ιδιωτικής ζωής και τη διαφάνεια στη χρήση δεδομένων.

2.5.3 Τεχνητή Νοημοσύνη και Ανησυχίες για την Απασχόληση

Ένας άλλος σημαντικός τομέας του δημόσιου διαλόγου περιστρέφεται γύρω από τον αντίκτυπο της τεχνητής νοημοσύνης στην απασχόληση. Η αυτοματοποίηση των θέσεων εργασίας από την τεχνητή νοημοσύνη και τη ρομποτική έχει οδηγήσει σε φόβους για εκτεταμένη μετατόπιση θέσεων εργασίας και αλλαγές στη φύση της εργασίας. Αυτή η ανησυχία εκτείνεται πέρα από την απώλεια θέσεων εργασίας και περιλαμβάνει τις συνέπειες για την ποιότητα της εργασίας, την ανισότητα εισοδήματος και τις δεξιότητες που απαιτούνται σε ένα ολοένα και πιο αυτοματοποιημένο εργατικό δυναμικό. Ο δημόσιος λόγος για αυτό το θέμα αντανακλά μια ευρύτερη ανησυχία σχετικά με τον κοινωνικό αντίκτυπο της τεχνητής νοημοσύνης και την ανάγκη για πολιτικές που αντιμετωπίζουν αυτές τις προκλήσεις.

2.5.4 Δημόσια Εμπλοκή στην Πολιτική και τη Ρύθμιση της Τεχνητής Νοημοσύνης

Ο αυξανόμενος αντίκτυπος της τεχνητής νοημοσύνης σε διάφορες πτυχές της ζωής έχει οδηγήσει σε αυξανόμενη ζήτηση για συμμετοχή του κοινού στη διαμόρφωση της πολιτικής και των κανονισμών για την τεχνητή νοημοσύνη. Οι Jobin, Ienca και Vayena (2019) συζητούν την ανάπτυξη παγκόσμιων κατευθυντήριων γραμμών για την ηθική της τεχνητής νοημοσύνης, αντανακλώντας μια κίνηση προς πιο περιεκτικές και συμμετοχικές προσεγγίσεις για τη διακυβέρνηση της τεχνητής νοημοσύνης. Η συμμετοχή του κοινού θεωρείται κρίσιμη για τη διασφάλιση ότι η ανάπτυξη της τεχνητής νοημοσύνης ευθυγραμμίζεται με τις κοινωνικές αξίες και αντιμετωπίζει τις ανησυχίες του κοινού.

2.5.5 Τεχνητή Νοημοσύνη και Ηθική Επίγνωση

Ο δημόσιος διάλογος για την τεχνητή νοημοσύνη οδήγησε επίσης σε αυξημένη ευαισθητοποίηση σχετικά με τις ηθικές επιπτώσεις της τεχνολογίας. Οι συζητήσεις σχετικά με τις ηθικές ευθύνες των προγραμματιστών τεχνητής νοημοσύνης, τον ηθικό σχεδιασμό των συστημάτων τεχνητής

νοημοσύνης και τις μακροπρόθεσμες επιπτώσεις της τεχνητής νοημοσύνης στην κοινωνία είναι όλο και πιο συχνές. Αυτή η αλλαγή αντικατοπτρίζει μια ευρύτερη αναγνώριση της ανάγκης για ενσωμάτωση ηθικών παραγόντων στην ανάπτυξη τεχνολογιών τεχνητής νοημοσύνης.

Η παρουσία της Τεχνητής Νοημοσύνης στον δημόσιο λόγο αντανακλά την αυξανόμενη σημασία της στην κοινωνία και τις περίπλοκες προκλήσεις που παρουσιάζει. Από αυτό το σημείο προχωρούμε σε πιο συγκεκριμένα ζητήματα. Εκτός από την επισκόπηση του πεδίου, είναι χρήσιμο να επικεντρωθούμε στην ανάλυση των λύσεων που προτείνουν οι ερευνητές, με αναφορά σε βασικούς τομείς της έρευνας, όπως η ιδιωτικότητα, η προκατάληψη, η αυτονομία και η λογοδοσία. Με αυτόν τον τρόπο, θα μπορούσε να παρουσιαστεί μια πιο ολοκληρωμένη άποψη των προκλήσεων και των προτεινόμενων λύσεων στην ανάπτυξη και τη ρύθμιση της τεχνητής νοημοσύνης. Καθώς η τεχνητή νοημοσύνη συνεχίζει να εξελίσσεται, η διατήρηση ενός ισχυρού και περιεκτικού δημόσιου λόγου θα είναι ζωτικής σημασίας για τη διαμόρφωση ενός μέλλοντος όπου οι τεχνολογίες τεχνητής νοημοσύνης αναπτύσσονται και χρησιμοποιούνται με τρόπους που είναι ωφέλιμοι, ηθικοί και ευθυγραμμισμένοι με τις κοινωνικές αξίες και τα ανθρώπινα δικαιώματα.

2.6 Το Σύγχρονο και Μελλοντικό Πλαίσιο της Ηθικής της Τεχνητής Νοημοσύνης

Ο τομέας της ηθικής της τεχνητής νοημοσύνης, στη διασταύρωση τεχνολογίας και ηθικής, εξελίσσεται ταχέως ως απάντηση στις συνεχείς εξελίξεις στην Τεχνητή Νοημοσύνη (TN). Επί του παρόντος, η ηθική της τεχνητής νοημοσύνης αντιμετωπίζει μια σειρά από περίπλοκα ζητήματα, που κυμαίνονται από το απόρρητο των δεδομένων έως την αλγοριθμική μεροληψία, ενώ προβλέπει επίσης μελλοντικές προκλήσεις που μπορεί να προκύψουν από περαιτέρω εξελίξεις στην τεχνολογία τεχνητής νοημοσύνης.

2.6.1 Τρέχον Ηθικό τοπίο στην Τεχνητή Νοημοσύνη

Στο παρόν πλαίσιο, η ηθική της τεχνητής νοημοσύνης επικεντρώνεται σε μεγάλο βαθμό στην αντιμετώπιση άμεσων ανησυχιών όπως το απόρρητο δεδομένων, η επιτήρηση, η αλγοριθμική δικαιοσύνη και η λογοδοσία των αυτόνομων συστημάτων. Η εξάπλωση της τεχνητής νοημοσύνης σε διάφορους τομείς έχει εντείνει τις ανησυχίες σχετικά με την προστασία των προσωπικών δεδομένων, όπως φαίνεται από τους Rahwan και των συνεργατών (2019) και Rumbold (2023). Η ηθική διαχείριση αυτών των δεδομένων είναι ζωτικής σημασίας για τη διατήρηση του ιδιωτικού απορρήτου και της εμπιστοσύνης στα συστήματα τεχνητής νοημοσύνης.

Η αλγοριθμική μεροληψία και η δικαιοσύνη βρίσκονται επίσης στην πρώτη γραμμή των σημερινών δεοντολογικών συζητήσεων της τεχνητής νοημοσύνης. Μελέτες όπως αυτές των Caliskan, Bryson και Narayanan (2017) έχουν δείξει πως οι προκαταλήψεις στα δεδομένα εκπαίδευσης μπορούν να οδηγήσουν σε μεροληπτικά αποτελέσματα στα συστήματα τεχνητής νοημοσύνης. Αυτό έχει ωθήσει μια κίνηση προς την ανάπτυξη πιο διαφανών και δίκαιων τεχνολογιών τεχνητής νοημοσύνης, ειδικά σε τομείς με σημαντικές κοινωνικές επιπτώσεις, όπως η ποινική δικαιοσύνη, η υγειονομική περίθαλψη και η απασχόληση.

Η ευθύνη και η λογοδοσία των συστημάτων τεχνητής νοημοσύνης, ιδιαίτερα των αυτόνομων, είναι ένας άλλος κρίσιμος τομέας εστίασης. Το «κενό ευθύνης» που συζητήθηκε από τον Mathias (2004) θέτει μια μοναδική πρόκληση, καθώς αμφισβητεί τη δυνατότητα εφαρμογής των παραδοσιακών νομικών και ηθικών πλαισίων σε σενάρια όπου τα συστήματα τεχνητής νοημοσύνης λαμβάνουν ανεξάρτητες αποφάσεις. Αυτό οδήγησε σε εκκλήσεις για σαφείς κανονισμούς και κατευθυντήριες γραμμές για τη διασφάλιση της λογοδοσίας στην ανάπτυξη και ανάπτυξη της τεχνητής νοημοσύνης.

2.6.2 Προβλέποντας το Μέλλον της Ηθικής της Τεχνητής Νοημοσύνης

Κοιτάζοντας στο μέλλον, η ηθική της τεχνητής νοημοσύνης αναμένεται να αντιμετωπίσει πιο περίπλοκες και διαφοροποιημένες προκλήσεις, καθώς οι τεχνολογίες τεχνητής νοημοσύνης γίνονται όλο και πιο εξελιγμένες. Ένας τομέας μελλοντικής ανησυχίας είναι οι ηθικές επιπτώσεις των προηγμένων συστημάτων τεχνητής νοημοσύνης ικανών να επιδεικνύουν συμπεριφορές παρόμοιες με την ανθρώπινη γνώση ή τη συναισθηματική νοημοσύνη. Αυτό εγείρει βαθιά ερωτήματα σχετικά με τα δικαιώματα των οντοτήτων (π.χ. Ευφυή Συστήματα Τεχνητής Νοημοσύνης, Ψηφιακοί Πράκτορες, Συνειδητές ή Ευαίσθητες Οντότητες κ.α) τεχνητής νοημοσύνης, τη φύση της συνείδησης και τη δυνατότητα της τεχνητής νοημοσύνης να επηρεάσει την ανθρώπινη συμπεριφορά και την κοινωνία γενικότερα.

Μια άλλη μελλοντική πρόκληση είναι η ηθική διακυβέρνηση της τεχνητής νοημοσύνης σε παγκόσμια κλίμακα. Καθώς οι τεχνολογίες τεχνητής νοημοσύνης ξεπερνούν τα εθνικά σύνορα, υπάρχει μια αυξανόμενη ανάγκη για διεθνή συνεργασία και συναίνεση σχετικά με τα ηθικά πρότυπα και τους κανονισμούς.

Η πιθανότητα υπερευφυνών συστημάτων Τεχνητής Νοημοσύνης φαίνεται επίσης σε μελλοντικές συζητήσεις. Οι ανησυχίες που διατυπώνονται από συγγραφείς όπως ο Bostrom (2016) σχετικά με τους υπαρξιακούς κινδύνους που θέτει η τεχνητή νοημοσύνη που ξεπερνά την ανθρώπινη νοημοσύνη απαιτούν προσεκτική εξέταση και προληπτικά μέτρα. Αυτό περιλαμβάνει τη

διασφάλιση ισχυρών μέτρων ασφαλείας, δεοντολογικών κατευθυντήριων γραμμών και μηχανισμών ελέγχου για την πρόληψη πιθανής κακής χρήσης ή ακούσιων συνεπειών της υπερέξυπνης τεχνητής νοημοσύνης.

2.6.3 Ενσωμάτωση της Ηθικής στην Σχεδίαση και Ανάπτυξη Συστημάτων Τεχνητής Νοημοσύνης

Το παρόν και το μέλλον της ηθικής της τεχνητής νοημοσύνης τονίζουν επίσης την ανάγκη ενσωμάτωσης ηθικών κριτηρίων στη διαδικασία ανάπτυξης της τεχνητής νοημοσύνης να στοχεύουν σε θεμελιώδεις αξίες, όπως η δικαιοσύνη, η διαφάνεια, η ιδιωτικότητα, η λογοδοσία, και η ισότητα. Αυτό απαιτεί διεπιστημονική συνεργασία, φέρνοντας σε επαφή τεχνολόγους, ηθικολόγους, υπεύθυνους χάραξης πολιτικής και άλλους ενδιαφερόμενους για να διασφαλιστεί ότι οι τεχνολογίες τεχνητής νοημοσύνης ευθυγραμμίζονται με τις ανθρώπινες αξίες και τους κοινωνικούς στόχους.

Η σύγχρονη προσέγγιση στην σχεδίαση και ανάπτυξη συστημάτων Τεχνητής Νοημοσύνης δεν περιορίζεται πλέον αποκλειστικά στην τεχνική αποτελεσματικότητα των συστημάτων. Η ηθική ενσωματώνεται πλέον ως βασική απαίτηση ήδη από τα πρώτα στάδια σχεδιασμού και ανάπτυξης. Σύμφωνα με τη φιλοσοφία του AI Act, η ΤΝ πρέπει να αναπτύσσεται με ανθρωποκεντρικό προσανατολισμό, διασφαλίζοντας ότι λειτουργεί ως εργαλείο προς όφελος του ανθρώπου και όχι ως μηχανισμός που υποκαθιστά πλήρως την ανθρώπινη κρίση.

Η προσέγγιση αυτή βασίζεται στην ανάγκη προστασίας της ανθρώπινης ευημερίας, της ασφάλειας και των θεμελιωδών δικαιωμάτων. Για τον λόγο αυτό, η ηθική δεν θεωρείται πλέον εξωτερικό στοιχείο της ανάπτυξης ενός συστήματος, αλλά συστατικό μέρος του κύκλου ζωής του. Η ανάπτυξη αξιόπιστων συστημάτων ΤΝ προϋποθέτει τη συνεχή αξιολόγηση πιθανών κινδύνων, τη διασφάλιση ανθρώπινης εποπτείας και τη λήψη μέτρων που ενισχύουν τη διαφάνεια και τη λογοδοσία.

Η ενσωμάτωση ηθικών αρχών από τη φάση του σχεδιασμού συμβάλλει στη δημιουργία συστημάτων που γίνονται περισσότερο αποδεκτά από την κοινωνία, ενισχύουν την εμπιστοσύνη των χρηστών και μειώνουν τον κίνδυνο αρνητικών κοινωνικών επιπτώσεων.

Καθώς η τεχνητή νοημοσύνη συνεχίζει να προοδεύει, οι ηθικοί προβληματισμοί πρέπει να προσαρμοστούν για την αντιμετώπιση τόσο των τρεχουσών προκλήσεων όσο και για την πρόβλεψη των μελλοντικών εξελίξεων. Αυτό απαιτεί συνεχή διάλογο, έρευνα και συνεργασία πέρα από τους κλάδους και τα σύνορα. Δίνοντας προτεραιότητα σε ηθικούς παράγοντες στην ανάπτυξη της τεχνητής νοημοσύνης, η κοινωνία μπορεί να εκμεταλλευτεί τα οφέλη της τεχνητής νοημοσύνης

μετριάζοντας τους κινδύνους της, διασφαλίζοντας ένα μέλλον όπου η τεχνητή νοημοσύνη συμβάλλει θετικά στην ανθρώπινη ευημερία και πρόοδο.

3

Βασικά Θέματα στην Ηθική της Τεχνητής Νοημοσύνης

Η συζήτηση σχετικά με την τεχνητή νοημοσύνη (ΤΝ) εγείρει συνεχώς μια σειρά σημαντικών ερωτημάτων που αντικατοπτρίζουν το ποικίλο και πολύπλοκο ηθικό τοπίο αυτού του ταχέως εξελισσόμενου τομέα. Αυτό το κεφάλαιο εμβαθύνει σε αυτά τα θέματα, παρέχοντας μια ολοκληρωμένη επισκόπηση των πρωταρχικών ηθικών ανησυχιών που σχετίζονται με την τεχνητή νοημοσύνη, όπως προσδιορίζονται στην τρέχουσα βιβλιογραφία.

3.1 Αλγοριθμική Προκατάληψη και Δικαιοσύνη στην Τεχνητή Νοημοσύνη

3.1.1 Πηγές Μεροληψίας στην Τεχνητή Νοημοσύνη

Οι πηγές της αλγοριθμικής μεροληψίας είναι πολύπλευρες και συχνά αλληλένδετες. Μπορούν να προέρχονται από προκατειλημμένα δεδομένα εκπαίδευσης, εσφαλμένο σχεδιασμό αλγορίθμων ή παρερμηνεία αποτελεσμάτων της τεχνητής νοημοσύνης. Για παράδειγμα, εάν ένα σύστημα τεχνητής νοημοσύνης εκπαιδεύεται σε ιστορικά δεδομένα προσλήψεων που αντικατοπτρίζουν μια ανισορροπία μεταξύ των φύλων σε ένα συγκεκριμένο επάγγελμα, το σύστημα μπορεί άθελά του να μάθει να ευνοεί το ένα φύλο έναντι του άλλου. Αυτός ο τύπος μεροληψίας στα συστήματα τεχνητής νοημοσύνης όχι μόνο διαιωνίζει τις υπάρχουσες κοινωνικές ανισότητες, αλλά τις ενισχύει με την πάροδο του χρόνου, εάν δεν ελεγχθεί.

3.1.2 Αντίκτυπος στην Κοινωνία

Ο αντίκτυπος της αλγοριθμικής μεροληψίας στην κοινωνία μπορεί να είναι βαθύς και εκτεταμένος. Σε τομείς όπως η επιβολή του νόμου, η προκατειλημμένη τεχνητή νοημοσύνη μπορεί να οδηγήσει σε άδικο προφίλ και αυξημένη παρακολούθηση ορισμένων ομάδων. Κατά την πρόσληψη, μπορεί να διαιωνίσει τις φυλετικές ανισότητες. Αυτό το είδος συστημικής προκατάληψης διαβρώνει την εμπιστοσύνη στις τεχνολογίες τεχνητής νοημοσύνης και μπορεί να έχει σοβαρές συνέπειες για άτομα που στοχοποιούνται άδικα ή υφίστανται διακρίσεις. Η μελέτη των Caliskan, Bryson και

Narayanan (2017) υπογραμμίζει πώς αυτές οι προκαταλήψεις μπορούν να εκδηλωθούν σε φαινομενικά ουδέτερες τεχνολογίες, αποκαλύπτοντας τις υποκείμενες προκαταλήψεις που κωδικοποιούνται μέσα τους.

3.1.3 Προκλήσεις στον Μετριάσμό της Προκατάληψης

Ο μετριάσμος της αλγοριθμικής μεροληψίας παρουσιάζει πολλές προκλήσεις. Μια σημαντική πρόκληση είναι ο εντοπισμός και η μέτρηση της μεροληψίας στα συστήματα τεχνητής νοημοσύνης. Λόγω της περίπλοκης και συχνά αδιαφανούς φύσης των αλγορίθμων τεχνητής νοημοσύνης, καθιστώντας δύσκολη την ανίχνευση της μεροληψίας ή εκδηλώνεται μόνο σε ορισμένα πλαίσια. Επιπλέον, υπάρχει η πρόκληση του ορισμού της δικαιοσύνης με τρόπο που να είναι παγκοσμίως αποδεκτός και να εφαρμόζεται σε διαφορετικά σενάρια. Οι διαφορετικοί πολιτισμικοί και κοινωνικοί κανόνες σημαίνουν ότι αυτό που θεωρείται δίκαιο σε ένα πλαίσιο μπορεί να μην θεωρείται δίκαιο σε ένα άλλο. Αυτό είναι ιδιαίτερα σημαντικό για τα συστήματα τεχνητής νοημοσύνης, καθώς τα δεδομένα που χρησιμοποιούνται για την εκπαίδευσή τους και οι ηθικές επιλογές που γίνονται κατά τον σχεδιασμό τους αντικατοπτρίζουν συχνά τις αξίες και τις παραδοχές της κοινωνίας από την οποία προέρχονται. Ως αποτέλεσμα, αυτό που θεωρείται δίκαιο σε ένα πλαίσιο μπορεί να θεωρηθεί άδικο σε ένα άλλο, δημιουργώντας σημαντικές προκλήσεις για την εφαρμογή της τεχνητής νοημοσύνης σε διαφορετικά πολιτισμικά και κοινωνικά περιβάλλοντα.

3.1.4 Προσεγγίσεις για τη Διασφάλιση της Δικαιοσύνης

Η αντιμετώπιση της αλγοριθμικής μεροληψίας και η διασφάλιση της δικαιοσύνης στην τεχνητή νοημοσύνη περιλαμβάνει πολλαπλές προσεγγίσεις. Μια προσέγγιση είναι η διαφοροποίηση των δεδομένων που χρησιμοποιούνται για την εκπαίδευση συστημάτων τεχνητής νοημοσύνης ώστε να διασφαλίζεται ότι είναι αντιπροσωπευτικά διαφορετικών ομάδων και προοπτικών. Αυτό απαιτεί προσεκτική επιμέλεια των συνόλων δεδομένων και κατανόηση του ιστορικού και κοινωνικού πλαισίου μέσα στο οποίο δημιουργήθηκαν τα δεδομένα. Μία άλλη προσέγγιση για τον μετριάσμό της μεροληψίας είναι η ανάπτυξη αλγορίθμων με τεχνικές που ενσωματώνουν μηχανική μάθηση με επίγνωση της δικαιοσύνης. Αυτοί οι αλγόριθμοι είναι σχεδιασμένοι να εντοπίζουν και να διορθώνουν προκαταλήψεις κατά τη διαδικασία μάθησης. Για παράδειγμα, η ενσωμάτωση τεχνικών όπως η "απομεροληπτική αναδιανομή" (debiasing) και η "αντιστάθμιση ευαισθησίας" (sensitivity adjustment) επιτρέπει στους αλγορίθμους να ανιχνεύουν και να εξισορροπούν τα δεδομένα που περιέχουν μεροληψία, εξασφαλίζοντας δίκαια αποτελέσματα για όλες τις ομάδες.

3.1.5 Ηθικά και Δεοντολογικά Πλαίσια για την Αντιμετώπιση της Αλγοριθμικής Μεροληψίας

Η ανάπτυξη ηθικών και δεοντολογικών πλαισίων αποτελεί κεντρικό ζήτημα στη διαμόρφωση πολιτικών που στοχεύουν στην καταπολέμηση της αλγοριθμικής μεροληψίας στην τεχνητή νοημοσύνη. Τα συστήματα τεχνητής νοημοσύνης επηρεάζονται άμεσα από τα δεδομένα που χρησιμοποιούνται για την εκπαίδευσή τους, και αυτά τα δεδομένα ενδέχεται να φέρουν τις προκαταλήψεις των ανθρώπων ή των κοινωνικών συνθηκών από τις οποίες προέρχονται. Ως αποτέλεσμα, οι μεροληπτικές αποφάσεις μπορούν να ενισχυθούν ή να αναπαραχθούν από τους αλγόριθμους, δημιουργώντας προβλήματα όπως διακρίσεις ή αδικίες.

Ηθικά και δεοντολογικά πλαίσια, όπως οι κατευθυντήριες γραμμές που αναπτύσσονται από οργανισμούς όπως η Ευρωπαϊκή Ένωση, το IEEE και άλλοι φορείς, προσφέρουν καίρια εργαλεία για την πρόληψη και τον μετριασμό της αλγοριθμικής μεροληψίας. Αυτά τα πλαίσια υπογραμμίζουν τη σημασία της ανάπτυξης αλγορίθμων που προάγουν τη διαφάνεια, τη δικαιοσύνη και τη λογοδοσία. Επιπλέον, καθοδηγούν τους προγραμματιστές να λαμβάνουν υπόψη τους τις κοινωνικές και ηθικές διαστάσεις κατά τον σχεδιασμό και την ανάπτυξη τεχνολογιών.

Μία βασική συνιστώσα αυτών των πλαισίων είναι η δημιουργία κατευθυντήριων γραμμών που προάγουν τη χρήση αντιπροσωπευτικών δεδομένων. Αυτό σημαίνει ότι τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των αλγορίθμων πρέπει να αντικατοπτρίζουν με ακρίβεια τη δημογραφική και πολιτισμική ποικιλομορφία των κοινωνιών, μειώνοντας έτσι τον κίνδυνο μεροληψίας. Η καταγραφή του ιστορικού και κοινωνικού πλαισίου των δεδομένων είναι απαραίτητη για την κατανόηση των προκαταλήψεων που μπορεί να εμπεριέχονται σε αυτά.

Πέρα από τη διαχείριση των δεδομένων, τα ηθικά πλαίσια ενθαρρύνουν τη χρήση τεχνικών όπως η "απομεροληπτική αναδιανομή" (debiasing) και η "αντιστάθμιση ευαισθησίας" (sensitivity adjustment). Αυτές οι τεχνικές επιτρέπουν στους αλγόριθμους να ανιχνεύουν και να διορθώνουν τις προκαταλήψεις κατά τη διάρκεια της μάθησής τους, προάγοντας πιο δίκαια αποτελέσματα. Με αυτόν τον τρόπο, οι αλγόριθμοι μπορούν να προσαρμόζονται σε διάφορα πολιτισμικά πλαίσια, εξασφαλίζοντας ότι οι αποφάσεις τους είναι δίκαιες, ανεξάρτητα από το κοινωνικό ή πολιτισμικό περιβάλλον.

Η εργασία των Jobin, Ienca και Vayena (2019), σχετικά με την ανάπτυξη κατευθυντήριων γραμμών ηθικής της τεχνητής νοημοσύνης, τονίζει την ανάγκη για παγκοσμίως αποδεκτά ηθικά πρότυπα που μπορούν να εφαρμοστούν σε διαφορετικά σενάρια και πολιτισμικά πλαίσια. Οι οδηγίες αυτές μπορούν να βοηθήσουν τους προγραμματιστές και τις εταιρείες τεχνητής νοημοσύνης να

πλοηγηθούν στις ηθικές πολυπλοκότητες που σχετίζονται με τον σχεδιασμό και την εφαρμογή των αλγορίθμων τους, προάγοντας τη δικαιοσύνη και μειώνοντας τις ανισότητες.

Συνολικά, η ανάπτυξη ηθικών και δεοντολογικών πλαισίων για την τεχνητή νοημοσύνη είναι ένα απαραίτητο βήμα για την καταπολέμηση της αλγοριθμικής μεροληψίας και τη διασφάλιση της δικαιοσύνης. Η εφαρμογή αυτών των πλαισίων είναι ζωτικής σημασίας για τη δημιουργία τεχνολογιών που σέβονται τις ανθρώπινες αξίες και τις κοινωνικές δικαιοδοσίες, συμβάλλοντας στη διαμόρφωση πιο δίκαιων και ηθικών αλγορίθμων.

Πίνακας 2: Αλγοριθμική Μεροληψία και Δικαιοσύνη στην Τεχνητή Νοημοσύνη

Θεματική Ενότητα	Κύρια Σημεία	Παράδειγμα
Πηγές Μεροληψίας στην Τεχνητή Νοημοσύνη	Προκατειλημμένα ή μη αντιπροσωπευτικά δεδομένα εκπαίδευσης. Εσφαλμένος σχεδιασμός ή αδιαφάνεια αλγορίθμων. Παρερμηνεία αποτελεσμάτων ή ανεπαρκής αξιολόγηση.	Εάν ένα σύστημα εκπαιδεύεται σε δεδομένα προσλήψεων με έμφυλη ανισορροπία, μπορεί να μάθει να ευνοεί το ένα φύλο.
Αντίκτυπος στην Κοινωνία	Ενίσχυση υπαρχουσών κοινωνικών και φυλετικών ανισοτήτων. Διάβρωση της εμπιστοσύνης των πολιτών προς τις τεχνολογίες ΤΝ. Σοβαρές επιπτώσεις σε άτομα που στοχοποιούνται άδικα (π.χ. σε επιβολή του νόμου).	Caliskan, Bryson & Narayanan (2017) αναδεικνύουν πώς «ουδέτερα» συστήματα μπορεί να ενσωματώνουν λανθάνουσες προκαταλήψεις.
Προκλήσεις στον Μετριάσμό της Προκατάληψης	Δυσκολία εντοπισμού και μέτρησης της μεροληψίας	Το τι θεωρείται δίκαιο σε έναν δυτικό πολιτισμό μπορεί να μην ισχύει σε έναν άλλο

	λόγω περίπλοκης ή «μαύρης κουτιού» φύσης αλγορίθμων. Διαφορετικές αντιλήψεις περί «δικαιοσύνης» ανάλογα με το πολιτισμικό ή κοινωνικό πλαίσιο. Ανάγκη σαφών κριτηρίων για την αξιολόγηση των δεδομένων και των αποτελεσμάτων.	πολιτισμό με διαφορετικές κοινωνικές νόρμες.
Προσεγγίσεις για τη Διασφάλιση της Δικαιοσύνης	Χρήση αντιπροσωπευτικών δεδομένων και έλεγχος του ιστορικού τους. Τεχνικές «απομερολυπτικής αναδιανομής» (debiasing) και «αντιστάθμισης ευαισθησίας» (sensitivity adjustment). Συνεχής παρακολούθηση και αναπροσαρμογή των μοντέλων για ελαχιστοποίηση της μεροληψίας.	Αλγόριθμοι που εντοπίζουν μεροληπτικές τάσεις στα δεδομένα και τροποποιούν το μοντέλο σε πραγματικό χρόνο για δίκαια αποτελέσματα.
Ηθικά και Δεοντολογικά Πλαίσια	Σημασία της διαφάνειας και της λογοδοσίας στον σχεδιασμό και την υλοποίηση της ΤΝ. Δημιουργία παγκόσμιων προτύπων (π.χ. από ΕΕ, IEEE) που διασφαλίζουν δικαιοσύνη	Jobin, Ienca & Vayena (2019) για παγκόσμια ηθικά πρότυπα στην ανάπτυξη της ΤΝ.

	σε διαφορετικά πολιτισμικά πλαίσια. Ανάγκη σύνδεσης τεχνικών πρακτικών με κοινωνικές και ηθικές αξίες.	
--	---	--

3.2 Προστασία Δεδομένων και Διακυβέρνηση Δεδομένων στην Τεχνητή Νοημοσύνη

3.2.1 Κατανόηση των Ανησυχιών περί Προστασίας Δεδομένων στην Τεχνητή Νοημοσύνη

Το απόρρητο στο πλαίσιο της Τεχνητής Νοημοσύνης (TN) είναι ένα πολύπλευρο ζήτημα, που περιστρέφεται κυρίως γύρω από τη συλλογή, αποθήκευση, επεξεργασία και χρήση προσωπικών δεδομένων. Τα συστήματα τεχνητής νοημοσύνης, ειδικά εκείνα που βασίζονται σε μεγάλα δεδομένα, έχουν τη δυνατότητα να αναλύουν τεράστιες ποσότητες πληροφοριών, που συχνά περιλαμβάνουν ευαίσθητα προσωπικά δεδομένα. Αυτό εγείρει σημαντικές ανησυχίες σχετικά με το απόρρητο του ατόμου και την προστασία των δεδομένων. Ο Rahwan και οι συνάδελφοι του (2019) τονίζουν τη σημασία της διαχείρισης αυτών των ανησυχιών, τονίζοντας την ηθική, νομική, τεχνολογική αλλά και οικονομική υποχρέωση προστασίας των προσωπικών δεδομένων στα συστήματα τεχνητής νοημοσύνης. Η ηθική διαχείριση των δεδομένων είναι ζωτικής σημασίας, όχι μόνο για το απόρρητο του ατόμου, αλλά και για τη ενίσχυση και διατήρηση της εμπιστοσύνης του κοινού στις τεχνολογίες τεχνητής νοημοσύνης.

3.2.2 Η Πρόκληση της Διακυβέρνησης Δεδομένων στην Τεχνητή Νοημοσύνη

Η διακυβέρνηση δεδομένων στην Τεχνητή Νοημοσύνη (TN) περιλαμβάνει πολιτικές, πρότυπα και πρακτικές που διασφαλίζουν ότι η διαχείριση των δεδομένων γίνεται με ηθικό και ασφαλή τρόπο. Η πρόκληση αυτή επιδεινώνεται από την πολυπλοκότητα και την αδιαφάνεια πολλών συστημάτων TN, γεγονός που καθιστά δύσκολη την κατανόηση του τρόπου χρήσης και επεξεργασίας των δεδομένων- όπως προτείνει ο Rumbold (2023), απαιτείται επειγόντως διαφάνεια που σέβεται τη συγκατάθεση των χρηστών και διασφαλίζει την ακεραιότητα των δεδομένων και πρακτικές διακυβέρνησης δεδομένων που σέβονται τη συγκατάθεση των χρηστών και διασφαλίζουν την ακεραιότητα των δεδομένων.

Για την αποτελεσματική αντιμετώπιση αυτών των προκλήσεων έχουν προταθεί λύσεις όπως η υιοθέτηση διαφανών μηχανισμών λήψης αποφάσεων που επιτρέπουν στους χρήστες να έχουν πρόσβαση στα δεδομένα τους και να κατανοούν τον τρόπο χρήσης τους. Επιπλέον, η υιοθέτηση αποκεντρωμένων μοντέλων διακυβέρνησης δεδομένων μπορεί να βελτιώσει την ασφάλεια και την ιδιωτικότητα και να επιτρέψει στους χρήστες να διαχειρίζονται καλύτερα τα προσωπικά τους δεδομένα - διεθνή πρότυπα ασφάλειας δεδομένων για να διασφαλιστεί ότι τα συστήματα ΤΝ συμμορφώνονται με τις παγκόσμιες κανονιστικές απαιτήσεις και προστατεύουν τα ατομικά δικαιώματα.

Η αποτελεσματική διακυβέρνηση των δεδομένων είναι το κλειδί για την αντιμετώπιση των προκλήσεων και την οικοδόμηση εμπιστοσύνης στο πλαίσιο της ΤΝ και μπορεί να επιτευχθεί μέσω σαφών ρυθμιστικών και τεχνικών λύσεων που δίνουν προτεραιότητα στην προστασία της ιδιωτικής ζωής και της ασφάλειας των χρηστών.

3.2.3 Εξισορρόπηση της Χρήσης Δεδομένων με την Προστασία της Ιδιωτικότητας

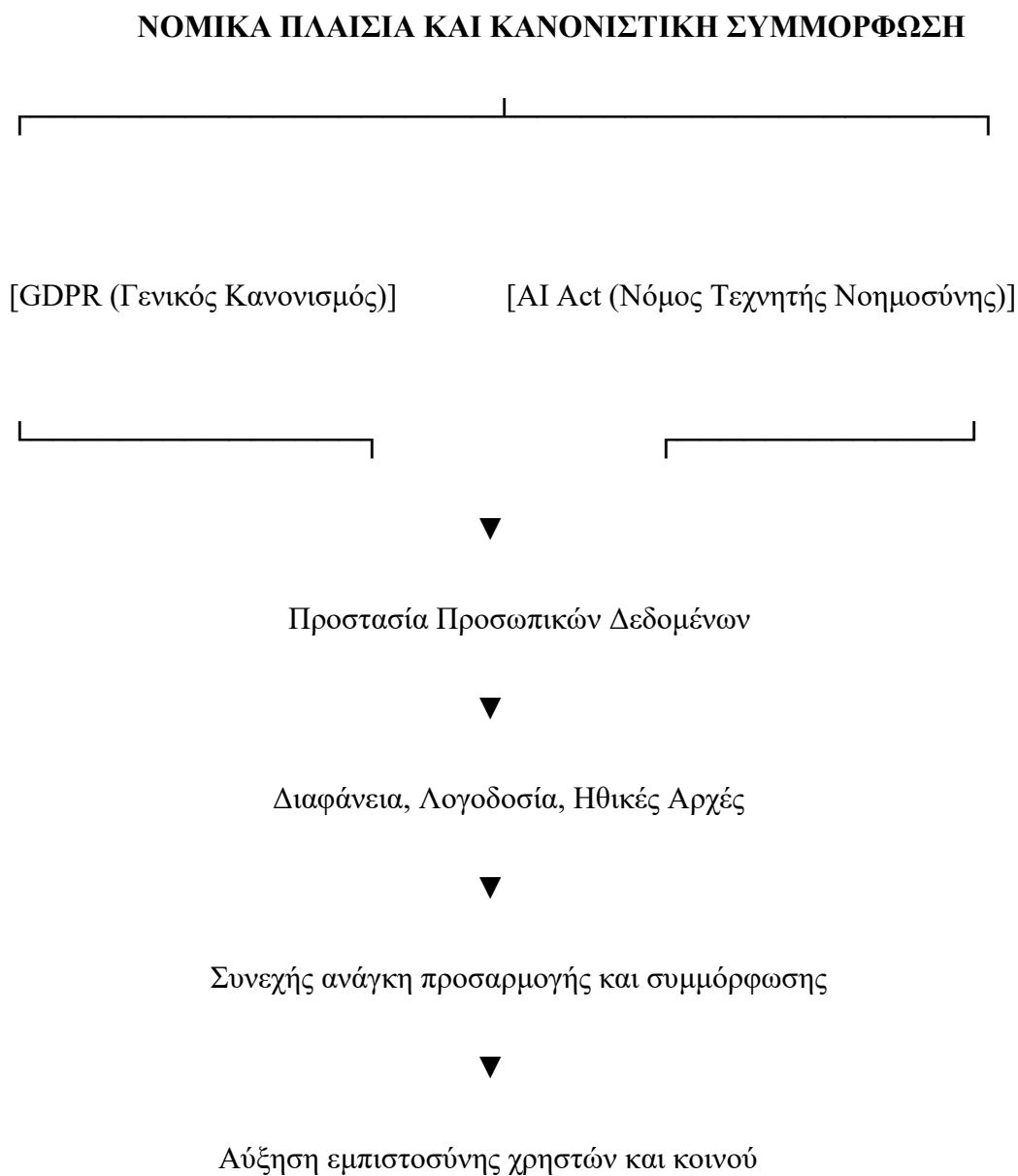
Μία από τις κεντρικές προκλήσεις στην τεχνητή νοημοσύνη είναι η εξισορρόπηση της χρησιμότητας των δεδομένων, η οποία μπορεί να οδηγήσει στην καινοτομία και να βελτιώσει τις υπηρεσίες, με την ανάγκη προστασίας του ιδιωτικού απορρήτου. Αυτή η ισορροπία είναι ιδιαίτερα ευαίσθητη σε τομείς όπως η υγειονομική περίθαλψη, τα οικονομικά και οι δημόσιες υπηρεσίες, όπου η τεχνητή νοημοσύνη έχει τη δυνατότητα να προσφέρει σημαντικά οφέλη αλλά και κινδύνους παραβίασης της ιδιωτικής ζωής. Η ανάπτυξη συστημάτων τεχνητής νοημοσύνης που μπορούν να αξιοποιήσουν αποτελεσματικά τα δεδομένα με σεβασμό του απορρήτου απαιτεί μια κατανόηση τόσο των τεχνολογικών δυνατοτήτων όσο και των ηθικών επιπτώσεων.

3.2.4 Νομικά Πλαίσια και Κανονιστική Συμμόρφωση

Τα νομικά πλαίσια διαδραματίζουν κρίσιμο ρόλο στη διαχείριση του απορρήτου στην τεχνητή νοημοσύνη. Κανονισμοί όπως ο Γενικός Κανονισμός για την Προστασία Δεδομένων (GDPR) στην Ευρωπαϊκή Ένωση έχουν θέσει προηγούμενα για την προστασία της ιδιωτικής ζωής και των δεδομένων, επιβάλλοντας αυστηρούς κανόνες για τον χειρισμό δεδομένων και παρέχοντας στα άτομα μεγαλύτερο έλεγχο των προσωπικών τους δεδομένων. Η συμμόρφωση με αυτούς τους κανονισμούς είναι απαραίτητη για τους προγραμματιστές και τους οργανισμούς τεχνητής νοημοσύνης, όχι μόνο για την αποφυγή νομικών επιπτώσεων αλλά και για την οικοδόμηση εμπιστοσύνης με τους χρήστες και το κοινό. Αυτά τα νομικά πλαίσια εξελίσσονται διαρκώς ως απάντηση στις προόδους της τεχνητής νοημοσύνης και της επιστήμης δεδομένων, δημιουργώντας μια συνεχή ανάγκη προσαρμογής και συμμόρφωσης των συστημάτων τεχνητής νοημοσύνης με τα

πιο πρόσφατα πρότυπα. Πρόσφατα, η Ευρωπαϊκή Ένωση προχώρησε σε περαιτέρω ρυθμίσεις με την πρόταση της AI Act, η οποία αποτελεί το πρώτο νομικό πλαίσιο που εστιάζει αποκλειστικά στη ρύθμιση της τεχνητής νοημοσύνης. Η AI Act στοχεύει στην κατάταξη και ρύθμιση των συστημάτων τεχνητής νοημοσύνης ανάλογα με τον βαθμό επικινδυνότητας που παρουσιάζουν για τα ανθρώπινα δικαιώματα, την ασφάλεια και τις θεμελιώδεις αξίες. Το πλαίσιο αυτό απαιτεί μεγαλύτερη διαφάνεια, λογοδοσία και συμμόρφωση με ηθικές αρχές κατά τη χρήση και ανάπτυξη συστημάτων τεχνητής νοημοσύνης.

Διάγραμμα 1. «Διάγραμμα Νομικών Πλαισίων και Κανονιστικής Συμμόρφωσης στην Τεχνητή Νοημοσύνη»



3.2.5 Τεχνολογικές Λύσεις για τη Προστασία της Ιδιωτικότητας

Οι τεχνολογικές εξελίξεις προσφέρουν πολλά υποσχόμενες λύσεις για τη διατήρηση της ιδιωτικής ζωής στην τεχνητή νοημοσύνη. Τεχνικές όπως το διαφορικό απόρρητο, η ομομορφική κρυπτογράφηση και η ομοσπονδιακή μάθηση αναπτύσσονται για να επιτρέψουν στα συστήματα τεχνητής νοημοσύνης να μαθαίνουν από δεδομένα χωρίς να διακυβεύεται το ατομικό απόρρητο. Το διαφορικό απόρρητο προσθέτει τυχαία στοιχεία στα δεδομένα, καθιστώντας δύσκολο τον προσδιορισμό ενός ατόμου, ενώ η ομομορφική κρυπτογράφηση επιτρέπει την επεξεργασία κρυπτογραφημένων δεδομένων χωρίς αποκρυπτογράφηση, προστατεύοντας έτσι την ιδιωτικότητα. Παράλληλα, η ομοσπονδιακή μάθηση επιτρέπει στα δεδομένα να παραμένουν τοπικά, εκπαιδεύοντας μοντέλα χωρίς τη μεταφορά των δεδομένων σε κεντρικούς διακομιστές. Αυτές οι τεχνολογίες επιτρέπουν την εξαγωγή χρήσιμων πληροφοριών από δεδομένα, διατηρώντας ταυτόχρονα τις ευαίσθητες λεπτομέρειες κρυφές. Η υιοθέτηση τέτοιων τεχνολογιών στην ανάπτυξη τεχνητής νοημοσύνης μπορεί να μειώσει σημαντικά τους κινδύνους προστασίας της ιδιωτικής ζωής και να ενισχύσει την ασφάλεια των δεδομένων.

3.2.6 Μοντέλα Ηθικής Διακυβέρνησης Δεδομένων

Η δημιουργία μοντέλων ηθικής διακυβέρνησης δεδομένων είναι ζωτικής σημασίας για την αντιμετώπιση των προβλημάτων απορρήτου στην τεχνητή νοημοσύνη. Αυτά τα μοντέλα θα πρέπει να περιλαμβάνουν αρχές όπως η διαφάνεια, η συναίνεση, η ελαχιστοποίηση δεδομένων και ο περιορισμός του σκοπού. Θα πρέπει επίσης να περιλαμβάνουν μηχανισμούς ώστε τα άτομα να κατανοούν, να ελέγχουν και να αμφισβητούν τη χρήση των δεδομένων τους. Η ανάπτυξη τέτοιων μοντέλων απαιτεί μια συλλογική προσέγγιση, με τη συμμετοχή ενδιαφερόμενων μερών από διαφορετικούς τομείς, συμπεριλαμβανομένης της τεχνολογίας, του δικαίου, της δεοντολογίας και της κοινωνίας των πολιτών. Ο τομέας της τεχνολογίας παίζει καίριο ρόλο, καθώς οι μηχανικοί και οι προγραμματιστές πρέπει να αναπτύσσουν συστήματα που να εφαρμόζουν τεχνικές προστασίας της ιδιωτικότητας και της ασφάλειας, όπως η κρυπτογράφηση και το διαφορικό απόρρητο. Από την άλλη πλευρά, ο τομέας του δικαίου είναι υπεύθυνος για τη ρύθμιση της χρήσης δεδομένων, με τη θέσπιση νομοθεσίας που διασφαλίζει τη συμμόρφωση με κανόνες προστασίας δεδομένων, όπως ο GDPR στην Ευρωπαϊκή Ένωση.

Η δεοντολογία είναι εξίσου σημαντική, καθώς αφορά τις ηθικές αρχές που πρέπει να διέπουν τη χρήση των δεδομένων, διασφαλίζοντας ότι οι αποφάσεις που λαμβάνονται από τα συστήματα τεχνητής νοημοσύνης είναι δίκαιες, διαφανείς και δεν επιφέρουν διακρίσεις. Τέλος, η κοινωνία των πολιτών παίζει κρίσιμο ρόλο, καθώς οι πολίτες και οι οργανώσεις που τους εκπροσωπούν πρέπει να συμμετέχουν στη διαμόρφωση των προτύπων και των πολιτικών για τη διακυβέρνηση

δεδομένων, διασφαλίζοντας ότι οι φωνές και οι ανησυχίες τους ακούγονται και λαμβάνονται υπόψη.

Η προστασία δεδομένων και η διακυβέρνηση δεδομένων στην τεχνητή νοημοσύνη είναι σύνθετοι και εξελισσόμενοι τομείς που απαιτούν συνεχή προσοχή και δράση. Η αντιμετώπιση των προβλημάτων απορρήτου στην τεχνητή νοημοσύνη περιλαμβάνει έναν συνδυασμό ηθικών πλαισίων, νομικής συμμόρφωσης, τεχνολογικών λύσεων και αποτελεσματικών μοντέλων διακυβέρνησης δεδομένων. Καθώς οι τεχνολογίες τεχνητής νοημοσύνης συνεχίζουν να προχωρούν και να ενσωματώνονται σε διάφορες πτυχές της ζωής, η σημασία της τήρησης των αρχών του απορρήτου και της ηθικής διακυβέρνησης δεδομένων γίνεται ολοένα και πιο κρίσιμη.

3.2.7 Προστασία της Ιδιωτικότητας και Υπεύθυνη Χρήση της Τεχνητής Νοημοσύνης

Η διασφάλιση ότι η τεχνητή νοημοσύνη σέβεται και προστατεύει το ατομικό απόρρητο είναι θεμελιώδους σημασίας για την οικοδόμηση μιας κοινωνίας όπου οι τεχνολογίες τεχνητής νοημοσύνης είναι αξιόπιστες και χρησιμοποιούνται με υπευθυνότητα. Η επίτευξη αυτού του στόχου απαιτεί μια σειρά λύσεων που περιλαμβάνουν την ενίσχυση της διαφάνειας στη χρήση των δεδομένων, με τους οργανισμούς να ενημερώνουν σαφώς τους χρήστες για τον τρόπο συλλογής και επεξεργασίας των προσωπικών τους δεδομένων. Επιπλέον, η υιοθέτηση αποκεντρωμένων μοντέλων διαχείρισης δεδομένων, όπως η ομοσπονδιακή μάθηση, ενισχύει τον έλεγχο των χρηστών στα δεδομένα τους, αποφεύγοντας τη συγκέντρωση ευαίσθητων πληροφοριών σε κεντρικούς διακομιστές. Τέλος, η εκπαίδευση και ευαισθητοποίηση των χρηστών σχετικά με τα δικαιώματά τους και τους κινδύνους που αφορούν την προστασία του απορρήτου τους αποτελούν σημαντικό βήμα για την ενίσχυση της υπευθυνότητας στη χρήση της τεχνητής νοημοσύνης.

3.3 Αυτονομία και Υπευθυνότητα στα Συστήματα Τεχνητής Νοημοσύνης

3.3.1 Η Αυξανόμενη Αυτονομία των Συστημάτων Τεχνητής Νοημοσύνης

Η αυτονομία των συστημάτων Τεχνητής Νοημοσύνης (TN) αναφέρεται στην ικανότητά τους να εκτελούν εργασίες, να λαμβάνουν αποφάσεις και ενδεχομένως να μαθαίνουν χωρίς άμεση ανθρώπινη παρέμβαση. Αυτή η αυτονομία, ενώ αποτελεί χαρακτηριστικό γνώρισμα της προηγμένης τεχνητής νοημοσύνης, εγείρει σημαντικά ηθικά ερωτήματα, ιδιαίτερα σχετικά με τη λογοδοσία. Καθώς τα συστήματα TN γίνονται πιο αυτόνομα, ο καθορισμός του ποιος είναι υπεύθυνος για τις αποφάσεις και τις ενέργειες αυτών των συστημάτων γίνεται όλο και πιο περίπλοκος.

3.3.2 Προκλήσεις στην Απόδοση Λογοδοσίας

Η κύρια πρόκληση στο πλαίσιο της αυτονομίας και της λογοδοσίας στην τεχνητή νοημοσύνη είναι να καθοριστεί ποιος είναι υπεύθυνος όταν ένα αυτόνομο σύστημα προκαλεί βλάβη ή λαμβάνει μια ανήθικη απόφαση. Τα παραδοσιακά νομικά και ηθικά πλαίσια είναι συχνά ανεπαρκώς εξοπλισμένα για να χειριστούν τα μοναδικά σενάρια που παρουσιάζονται από την αυτόνομη τεχνητή νοημοσύνη. Για παράδειγμα, εάν ένα αυτόνομο όχημα εμπλέκεται σε ένα ατύχημα, τίθεται το ερώτημα εάν την ευθύνη φέρει ο κατασκευαστής, ο προγραμματιστής λογισμικού, ο χρήστης ή το ίδιο το σύστημα τεχνητής νοημοσύνης. Αυτή η πολυπλοκότητα απαιτεί μια επανεξέταση των μηχανισμών λογοδοσίας στην εποχή της τεχνητής νοημοσύνης.

Η αντιμετώπιση του ζητήματος της λογοδοσίας στα συστήματα τεχνητής νοημοσύνης απαιτεί περισσότερο από τον απλό εντοπισμό υπευθύνων μετά την εμφάνιση ενός προβλήματος. Η αποτελεσματική διακυβέρνηση της τεχνητής νοημοσύνης βασίζεται στον σαφή καθορισμό ρόλων, ευθυνών και αρμοδιοτήτων σε όλα τα στάδια του κύκλου ζωής ενός συστήματος. Η ανάθεση συγκεκριμένων αρμοδιοτήτων για τη συμμόρφωση, την παρακολούθηση και την αξιολόγηση των συστημάτων ΤΝ συμβάλλει στη διασφάλιση της υπεύθυνης χρήσης τους και περιορίζει το φαινόμενο του κενού ευθύνης που συχνά εμφανίζεται σε περιβάλλοντα υψηλής αυτονομίας. Με τον τρόπο αυτό ενισχύεται η λογοδοσία των οργανισμών και δημιουργείται ένα πλαίσιο αποτελεσματικού ελέγχου των αποφάσεων που λαμβάνονται με τη βοήθεια της τεχνητής νοημοσύνης.

3.3.3 Ανάπτυξη Πλαισίων Λογοδοσίας για Τεχνητή Νοημοσύνη

Για την αντιμετώπιση αυτών των προκλήσεων, υπάρχει ανάγκη για ισχυρά πλαίσια λογοδοσίας ειδικά σχεδιασμένα για συστήματα τεχνητής νοημοσύνης. Αυτά τα πλαίσια θα πρέπει να περιλαμβάνουν σαφείς κατευθυντήριες γραμμές και πρότυπα για την ανάπτυξη και την ανάπτυξη της τεχνητής νοημοσύνης, διασφαλίζοντας ότι τα συστήματα τεχνητής νοημοσύνης είναι αξιόπιστα, ασφαλή και ηθικά ευθυγραμμισμένα. Οι Jobin, Ienca και Vayena (2019) τονίζουν τη σημασία της παγκόσμιας συνεργασίας για την ανάπτυξη τέτοιων πλαισίων, υπογραμμίζοντας την ανάγκη για διεθνή συναίνεση σχετικά με τα ηθικά πρότυπα και τους κανονισμούς για την τεχνητή νοημοσύνη. Η αυξανόμενη αυτονομία των συστημάτων ΤΝ δημιουργεί σημαντικές προκλήσεις σχετικά με την ευθύνη για τις αποφάσεις που λαμβάνονται από αυτά. Όσο περισσότερο ένα σύστημα λειτουργεί ανεξάρτητα από την άμεση ανθρώπινη παρέμβαση, τόσο πιο σημαντικό γίνεται να καθορίζονται μηχανισμοί εποπτείας και ελέγχου.

Το AI Act υιοθετεί την άποψη ότι η ανάπτυξη αυτόνομων συστημάτων πρέπει να συνοδεύεται από ουσιαστική ανθρώπινη εποπτεία. Η ανθρώπινη συμμετοχή παραμένει απαραίτητη ώστε να

διασφαλίζεται ότι οι αποφάσεις που παράγονται από τα συστήματα ΤΝ μπορούν να ελεγχθούν, να αξιολογηθούν και, εφόσον απαιτείται, να διορθωθούν.

Η προσέγγιση αυτή αποσκοπεί στην αποφυγή του λεγόμενου «κενού ευθύνης», δηλαδή της κατάστασης όπου δεν είναι σαφές ποιος φέρει την ευθύνη για τις επιπτώσεις μιας απόφασης που λήφθηκε με τη βοήθεια ΤΝ. Για τον λόγο αυτό, η ανάπτυξη υπεύθυνων συστημάτων ΤΝ απαιτεί σαφή κατανομή αρμοδιοτήτων μεταξύ σχεδιαστών, οργανισμών και χρηστών.

3.3.4 Νομικές και Ηθικές Επιπτώσεις της Αυτονομίας της Τεχνητής Νοημοσύνης

Η αυξανόμενη αυτονομία των συστημάτων τεχνητής νοημοσύνης εγείρει ηθικά ζητήματα, καθώς αμφισβητεί τις παραδοσιακές έννοιες της ηθικής ευθύνης και εξουσίας. Συγκεκριμένα, δημιουργείται το ερώτημα αν και πώς μπορούν τα αυτόνομα συστήματα να θεωρηθούν υπεύθυνα για τις πράξεις τους, κάτι που μέχρι τώρα αποδιδόταν αποκλειστικά στους ανθρώπους ή στους οργανισμούς που τα χρησιμοποιούν.

3.3.5 Ο Ρόλος της Διαφάνειας στη Λογοδοσία της Τεχνητής Νοημοσύνης

Η διαφάνεια είναι ένας βασικός παράγοντας για την αντιμετώπιση των προκλήσεων της αυτονομίας και της λογοδοσίας στην τεχνητή νοημοσύνη. Τα συστήματα τεχνητής νοημοσύνης που λειτουργούν με διαφάνεια επιτρέπουν την καλύτερη κατανόηση και τον έλεγχο των διαδικασιών λήψης αποφάσεων τους, καθιστώντας ευκολότερο τον εντοπισμό των ευθυνών. Αυτή η διαφάνεια δεν είναι μόνο τεχνική, που περιλαμβάνει την επεξήγηση των αλγορίθμων τεχνητής νοημοσύνης, αλλά και διαδικαστική, διασφαλίζοντας ότι οι διαδικασίες ανάπτυξης και λειτουργίας τεχνητής νοημοσύνης είναι ανοιχτές σε επίβλεψη και επανεξέταση.

3.3.6 Μελλοντικές Κατευθύνσεις στην Αυτονομία και την Υπευθυνότητα

Κοιτάζοντας το μέλλον, ο τομέας της ηθικής της τεχνητής νοημοσύνης πρέπει να συνεχίσει να εξελίσσεται και να προσαρμόζεται για να αντιμετωπίσει την αυξανόμενη αυτονομία των συστημάτων τεχνητής νοημοσύνης. Αυτό πιθανότατα θα περιλαμβάνει έναν συνδυασμό τεχνολογικών λύσεων, όπως η ανάπτυξη συστημάτων τεχνητής νοημοσύνης με ενσωματωμένους ηθικούς περιορισμούς και μηχανισμούς λογοδοσίας, καθώς και νομικές και ρυθμιστικές καινοτομίες. Επιπλέον, υπάρχει ανάγκη για συνεχή διάλογο και συνεργασία μεταξύ τεχνολόγων, υπευθύνων χάραξης πολιτικής και του κοινού για αντιμετώπιση των ηθικών περιπλοκών που θέτει η αυτόνομη τεχνητή νοημοσύνη.

Η αυτονομία των συστημάτων τεχνητής νοημοσύνης παρουσιάζει σημαντικές προκλήσεις όσον αφορά τη λογοδοσία, απαιτώντας μια πολυδιάστατη προσέγγιση που περιλαμβάνει τεχνικές, νομικές, δεοντολογικές και ρυθμιστικές στρατηγικές. Η διασφάλιση της λογοδοσίας στην ανάπτυξη αυτόνομης τεχνητής νοημοσύνης είναι ζωτικής σημασίας για τη ενίσχυση και διατήρηση της εμπιστοσύνης του κοινού και την αξιοποίηση του πλήρους δυναμικού των τεχνολογιών τεχνητής νοημοσύνης με υπεύθυνο και ηθικό τρόπο. Καθώς η τεχνητή νοημοσύνη συνεχίζει να προοδεύει, η καλλιέργεια μιας κουλτούρας λογοδοσίας θα είναι απαραίτητη για την πλοήγηση στο ηθικό τοπίο της αυτονομίας της τεχνητής νοημοσύνης.

3.4 Διαφάνεια και Επεξήγηση στην Τεχνητή Νοημοσύνη

3.4.1 Μελλοντικές Κατευθύνσεις στην Αυτονομία και την Υπευθυνότητα

Η διαφάνεια στην τεχνητή νοημοσύνη (TN) είναι ζωτικής σημασίας για την κατανόηση, τη διαχείριση και την εμπιστοσύνη των συστημάτων τεχνητής νοημοσύνης. Αναφέρεται στη σαφήνεια με την οποία οι αλγόριθμοι τεχνητής νοημοσύνης, οι διαδικασίες λήψης αποφάσεων και τα αποτελέσματα κοινοποιούνται και γίνονται κατανοητά στους χρήστες και στις διάφορες ομάδες που επηρεάζονται ή έχουν κάποιο συμφέρον από τη διαφάνεια στα συστήματα τεχνητής νοημοσύνης. Αυτή η πτυχή της ηθικής της τεχνητής νοημοσύνης αποκτά σημασία καθώς τα συστήματα τεχνητής νοημοσύνης γίνονται πιο διαδεδомμένα σε διάφορους τομείς, επηρεάζοντας σημαντικές αποφάσεις που επηρεάζουν τα άτομα και την κοινωνία. Η αναγκαιότητα για διαφάνεια τονίζεται από τον Rahwan και τους συναδέλφους του (2019), οι οποίοι τονίζουν τη σημασία του για την οικοδόμηση εμπιστοσύνης και τη διασφάλιση της υπεύθυνης χρήσης των τεχνολογιών TN. Η διαφάνεια αποτελεί μία από τις σημαντικότερες προϋποθέσεις για την υπεύθυνη χρήση της Τεχνητής Νοημοσύνης. Καθώς τα συστήματα TN συμμετέχουν ολοένα περισσότερο σε διαδικασίες λήψης αποφάσεων, οι χρήστες και τα άτομα που επηρεάζονται από αυτές τις αποφάσεις πρέπει να μπορούν να κατανοούν τον τρόπο λειτουργίας των συστημάτων.

Η προσέγγιση του AI Act αναδεικνύει τη σημασία της κατανόησης των δυνατοτήτων αλλά και των περιορισμών των συστημάτων TN. Η διαφάνεια δεν αφορά μόνο την τεχνική περιγραφή ενός μοντέλου, αλλά και τη δυνατότητα των ανθρώπων να ερμηνεύουν τα αποτελέσματα που παράγονται από αυτό.

3.4.2 Ξεπερνώντας τις Προκλήσεις της Διαφάνειας της Τεχνητής Νοημοσύνης

Η κύρια πρόκληση για την επίτευξη διαφάνειας στην τεχνητή νοημοσύνη είναι η εγγενής πολυπλοκότητα αυτών των συστημάτων, ιδιαίτερα εκείνων που βασίζονται σε προηγμένες τεχνικές

μηχανικής μάθησης, όπως η βαθιά μάθηση. Αυτά τα μοντέλα τεχνητής νοημοσύνης λειτουργούν συχνά ως «μαύρα κουτιά», με διαδικασίες λήψης αποφάσεων που δεν ερμηνεύονται εύκολα από τον άνθρωπο. Αυτή η ασάφεια δημιουργεί σημαντικά εμπόδια στη διασφάλιση της διαφάνειας, καθώς καθίσταται δύσκολος ο έλεγχος των αποφάσεων τεχνητής νοημοσύνης, εγείροντας ανησυχίες για πιθανές προκαταλήψεις και ζητήματα λογοδοσίας. Η εργασία των Jobin, Ienca και Vayena (2019) υπογραμμίζει την ανάγκη για σαφή και κατανοητά συστήματα τεχνητής νοημοσύνης για τη διασφάλιση της ηθικής συμμόρφωσης και την ενίσχυση της εμπιστοσύνης του κοινού.

3.4.3 Υπογράμμιση της Σημασίας της Εξηγήσιμης Τεχνητής Νοημοσύνης

Το Explainable AI (XAI) στοχεύει να αντιμετωπίσει την πρόκληση της διαφάνειας καθιστώντας τις διαδικασίες λήψης αποφάσεων τεχνητής νοημοσύνης πιο κατανοητές και ερμηνεύσιμες για τον άνθρωπο. Το XAI είναι ιδιαίτερα κρίσιμο σε τομείς υψηλού κινδύνου, όπως η υγειονομική περίθαλψη, η ποινική δικαιοσύνη και τα οικονομικά, όπου οι αποφάσεις τεχνητής νοημοσύνης μπορούν να επηρεάσουν σημαντικά τις ατομικές ζωές. Η κίνηση προς το XAI δεν είναι μόνο μια τεχνολογική αναγκαιότητα, αλλά και μια ηθική υποχρέωση, διασφαλίζοντας ότι τα συστήματα τεχνητής νοημοσύνης είναι υπόλογα και ότι οι αποφάσεις τους μπορούν να αξιολογηθούν.

Παράλληλα, η ύπαρξη κατάλληλης τεκμηρίωσης και η δυνατότητα παρακολούθησης των αποφάσεων ενός συστήματος συμβάλλουν στην αύξηση της λογοδοσίας. Όσο πιο διαφανές είναι ένα σύστημα, τόσο ευκολότερο είναι να εντοπιστούν πιθανά σφάλματα, προκαταλήψεις ή ανεπιθύμητες συνέπειες.

Η διαφάνεια λειτουργεί επίσης ως μηχανισμός ενίσχυσης της κοινωνικής εμπιστοσύνης απέναντι στην ΤΝ, καθώς επιτρέπει στους χρήστες να γνωρίζουν πότε αλληλεπιδρούν με ένα σύστημα ΤΝ και να αξιολογούν κριτικά τα αποτελέσματά του.

3.4.4 Η Αλληλεπίδραση της Διαφάνειας και της Ηθικής Τεχνητής Νοημοσύνης

Η διαφάνεια στην τεχνητή νοημοσύνη είναι βαθιά συνυφασμένη με ευρύτερα ηθικά ζητήματα, όπως η δικαιοσύνη, η λογοδοσία και η εμπιστοσύνη. Τα διαφανή συστήματα τεχνητής νοημοσύνης επιτρέπουν μεγαλύτερο έλεγχο των λειτουργιών τους, διευκολύνοντας τον εντοπισμό και τη διόρθωση προκαταλήψεων και αθέμιτων πρακτικών. Επιτρέπουν επίσης στους χρήστες και τους ενδιαφερόμενους να κατανοούν και να εμπιστεύονται τις αποφάσεις τεχνητής νοημοσύνης, κάτι που είναι απαραίτητο για την αποδοχή και την ηθική τους ενσωμάτωση στην κοινωνία.

3.4.5 Μελλοντικές Κατευθύνσεις για Διαφανή και Εξηγήσιμη Τεχνητή Νοημοσύνη

Προβλέποντας στο μέλλον, η ανάπτυξη διαφανών και εξηγήσιμων συστημάτων τεχνητής νοημοσύνης θα απαιτήσει συνεχείς προσπάθειες από τεχνολόγους, ηθικούς, υπεύθυνους χάραξης πολιτικής και άλλους ενδιαφερόμενους φορείς. Οι εξελίξεις στην τεχνολογία τεχνητής νοημοσύνης πρέπει να συμβαδίζουν με την ανάπτυξη πλαισίων και εργαλείων που ενισχύουν τη διαφάνεια και την επεξήγηση των συστημάτων τεχνητής νοημοσύνης. Αυτό περιλαμβάνει όχι μόνο τεχνικές καινοτομίες αλλά και ρυθμιστικά και μέτρα πολιτικής για να διασφαλιστεί ότι τα συστήματα τεχνητής νοημοσύνης αναπτύσσονται και χρησιμοποιούνται με ηθικό, διαφανή και υπεύθυνο τρόπο.

Η διαφάνεια και η επεξήγηση είναι θεμελιώδεις για την ηθική χρήση της τεχνητής νοημοσύνης, διασφαλίζοντας ότι τα συστήματα τεχνητής νοημοσύνης είναι κατανοητά, αξιόπιστα και υπεύθυνα. Καθώς η τεχνητή νοημοσύνη συνεχίζει να εξελίσσεται και να επηρεάζει διάφορες πτυχές της κοινωνίας, η έμφαση στη διαφανή και εξηγήσιμη τεχνητή νοημοσύνη θα είναι κρίσιμη για την προώθηση της υπεύθυνης και ηθικής ανάπτυξης και εξάπλωσης της τεχνητής νοημοσύνης. Η διασφάλιση ότι τα συστήματα τεχνητής νοημοσύνης είναι διαφανή και ότι οι λειτουργίες και οι αποφάσεις τους είναι εξηγήσιμες θα παραμείνει βασική πρόκληση και προτεραιότητα στον τομέα της ηθικής της τεχνητής νοημοσύνης.

3.5 Αλληλεπίδραση Ανθρώπου – Τεχνητής Νοημοσύνης και Κοινωνικές Επιπτώσεις

3.5.1 Δυναμική της Αλληλεπίδρασης Ανθρώπου – Τεχνητής Νοημοσύνης

Η αλληλεπίδραση ανθρώπου - τεχνητής νοημοσύνης έχει γίνει αναπόσπαστο μέρος της καθημερινής ζωής, επηρεάζοντας βαθιά τον τρόπο με τον οποίο τα άτομα αλληλοεπιδρούν με την τεχνολογία και μεταξύ τους. Αυτή η αλληλεπίδραση περιλαμβάνει διάφορες μορφές, από τη χρήση προσωπικών βοηθών με τεχνητή νοημοσύνη έως την αντιμετώπιση τεχνητής νοημοσύνης σε επαγγελματικά περιβάλλοντα. Η πολυπλοκότητα αυτών των αλληλεπιδράσεων εγείρει ηθικές ανησυχίες σχετικά με την εξάρτηση, την εμπιστοσύνη και τη μεταβαλλόμενη φύση των ανθρώπινων σχέσεων στο πλαίσιο της τεχνητής νοημοσύνης. Όπως ο Rahwan και οι συνάδελφοι του (2019) επισημαίνουν ότι η ενσωμάτωση της τεχνητής νοημοσύνης στα κοινωνικά πλαίσια απαιτεί τη βαθιά κατανόηση αυτών των αλληλεπιδράσεων για να διασφαλιστεί ότι ενισχύουν, αντί να μειώνουν, τις ανθρώπινες εμπειρίες και ικανότητες.

3.5.2 Κοινωνικός Αντίκτυπος της Τεχνητής Νοημοσύνης

Ο κοινωνικός αντίκτυπος της τεχνητής νοημοσύνης είναι εκτεταμένος, επηρεάζοντας τα πάντα, από την απασχόληση και την οικονομία μέχρι την ιδιωτικότητα και την ασφάλεια. Οι τεχνολογίες ΤΝ έχουν τη δυνατότητα να μεταμορφώσουν τις βιομηχανίες, να εξορθολογήσουν τις διαδικασίες και να δημιουργήσουν νέες ευκαιρίες. Ωστόσο, θέτουν επίσης προκλήσεις όπως η μετατόπιση εργασίας, η οικονομική ανισότητα και απειλές για την ιδιωτικότητα. Η εργασία της O'Neil (2017) με θέμα "Όπλα μαθηματικής καταστροφής" υπογραμμίζει πώς η τεχνητή νοημοσύνη, εάν δεν διαχειρίζεται προσεκτικά, μπορεί να επιδεινώσει τις κοινωνικές ανισότητες και να βλάψει ευάλωτους πληθυσμούς. Η κατανόηση και η αντιμετώπιση των κοινωνικών επιπτώσεων της τεχνητής νοημοσύνης είναι ζωτικής σημασίας για την αξιοποίηση των πλεονεκτημάτων της, ενώ ταυτόχρονα μετριάζονται οι κίνδυνοι της.

3.5.3 Ηθικές Θεωρήσεις στην Αλληλεπίδραση Ανθρώπου – Τεχνητής Νοημοσύνης

Οι ηθικοί προβληματισμοί στην αλληλεπίδραση ανθρώπου - τεχνητής νοημοσύνης περιλαμβάνουν θέματα αυτονομίας, συναίνεσης και τη δυνατότητα να επηρεάσει η τεχνητή νοημοσύνη τις ανθρώπινες αποφάσεις και συμπεριφορές. Καθώς τα συστήματα ΤΝ γίνονται πιο προηγμένα, υπάρχει μια αυξανόμενη ανησυχία για την ικανότητά τους να χειραγωγούν ή να επηρεάζουν αδικαιολόγητα τους χρήστες. Η διασφάλιση ότι οι αλληλεπιδράσεις ανθρώπου - ΤΝ σέβονται την ανθρώπινη αυτονομία και τη λήψη αποφάσεων αποτελεί βασικό ηθικό μέλημα.

3.5.4 Η Τεχνητή Νοημοσύνη και η Μεταβαλλόμενη Φύση της Εργασίας

Ο αντίκτυπος της τεχνητής νοημοσύνης στην αγορά εργασίας και η φύση της εργασίας είναι ένας κρίσιμος τομέας κοινωνικής ανησυχίας. Οι τεχνολογίες αυτοματισμού και τεχνητής νοημοσύνης έχουν τη δυνατότητα να εκτοπίσουν ορισμένους τύπους θέσεων εργασίας, ενώ δημιουργούν νέες. Αυτή η αλλαγή εγείρει σημαντικά ερωτήματα σχετικά με το μέλλον της εργασίας, την ανάπτυξη δεξιοτήτων και την οικονομική ασφάλεια. Η πρόκληση είναι να διασφαλιστεί ότι η μετάβαση σε μια οικονομία που βασίζεται στην τεχνητή νοημοσύνη είναι χωρίς αποκλεισμούς και δίκαιη, παρέχοντας ευκαιρίες σε όλα τα μέλη της κοινωνίας.

3.5.5 Πολιτιστικές και Κοινωνικές Επιπτώσεις της Τεχνητής Νοημοσύνης

Η τεχνητή νοημοσύνη έχει επίσης σημαντικές πολιτιστικές και κοινωνικές επιπτώσεις. Καθώς τα συστήματα τεχνητής νοημοσύνης σχεδιάζονται και εκπαιδεύονται με βάση συγκεκριμένα σύνολα δεδομένων και αξίες, μπορούν να αντανakλούν και να ενισχύσουν τις πολιτισμικές προκαταλήψεις. Αυτό εγείρει ανησυχίες σχετικά με την εκπροσώπηση και την ενσωμάτωση στην ανάπτυξη της ΤΝ. Οι Jobin, Ienca και Vayena (2019) τονίζουν τη σημασία της εξέτασης διαφορετικών προοπτικών και αξιών στη δημιουργία συστημάτων

τεχνητής νοημοσύνης για να διασφαλιστεί ότι είναι πολιτιστικά ευαίσθητα και κοινωνικά χωρίς αποκλεισμούς.

3.5.6 Μελλοντικές Προοπτικές για τη Συνύπαρξη Ανθρώπου – Τεχνητής Νοημοσύνης

Κοιτάζοντας το μέλλον, η συνύπαρξη ανθρώπων και τεχνητής νοημοσύνης παρουσιάζει ευκαιρίες και προκλήσεις. Το μέλλον της αλληλεπίδρασης ανθρώπου - τεχνητής νοημοσύνης πιθανότατα θα περιλαμβάνει πιο συνεργατικές και ολοκληρωμένες μορφές τεχνητής νοημοσύνης, που απαιτούν προσεκτική εξέταση των ηθικών, κοινωνικών και πολιτισμικών δυναμικών. Καθώς η τεχνητή νοημοσύνη συνεχίζει να προοδεύει, είναι επιτακτική η ενίσχυση μιας κοινωνίας που είναι ενημερωμένη, δεσμευμένη και προετοιμασμένη για αυτές τις αλλαγές. Αυτό περιλαμβάνει όχι μόνο τεχνολογικές καινοτομίες και καινοτομίες πολιτικής, αλλά και εκπαίδευση και συμμετοχή του κοινού για τη διαμόρφωση ενός μέλλοντος όπου η τεχνητή νοημοσύνη ενισχύει την ανθρώπινη κοινωνία.

Η αλληλεπίδραση μεταξύ ανθρώπων και τεχνητής νοημοσύνης και ο αντίκτυπός της στην κοινωνία είναι ένα πολύπλευρο ζήτημα που απαιτεί προσεκτική εξέταση και προληπτική διαχείριση. Καθώς η τεχνητή νοημοσύνη ενσωματώνεται περισσότερο στον κοινωνικό μας ιστό, η κατανόηση και η αντιμετώπιση των ηθικών και κοινωνικών συνεπειών της είναι ζωτικής σημασίας. Η διασφάλιση ότι οι αλληλεπιδράσεις ανθρώπου-τεχνητής νοημοσύνης είναι ωφέλιμες, ηθικές και εμπλουτισμένες θα είναι καθοριστικής σημασίας για την πλοήγηση στις προκλήσεις και τις ευκαιρίες που παρουσιάζει η τεχνητή νοημοσύνη τα επόμενα χρόνια.

3.6 Παγκόσμια Ηθική και Πολιτική για την Τεχνητή Νοημοσύνη

3.6.1 Αναγκαιότητα Παγκόσμιας Ηθικής στην Τεχνητή Νοημοσύνη

Η ανάπτυξη της Τεχνητής Νοημοσύνης (ΤΝ) δεν περιορίζεται από γεωγραφικά όρια, καθιστώντας την παγκόσμια ηθική στην τεχνητή νοημοσύνη αναγκαιότητα. Καθώς τα συστήματα τεχνητής νοημοσύνης χρησιμοποιούνται όλο και περισσότερο σε διαφορετικές χώρες και πολιτισμούς, η ανάγκη για ένα παγκοσμίως αποδεκτό σύνολο ηθικών κατευθυντήριων γραμμών καθίσταται επιτακτική. Οι Jobin, Ienca και Vayena (2019) τονίζουν τη σημασία της δημιουργίας ενός παγκόσμιου τοπίου κατευθυντήριων γραμμών ηθικής τεχνητής νοημοσύνης που μπορούν να καθοδηγήσουν την ανάπτυξη και χρήση της τεχνητής νοημοσύνης σε διάφορα πολιτιστικά και εθνικά πλαίσια. Αυτή η παγκόσμια ηθική είναι ζωτικής σημασίας για τη διασφάλιση ότι οι

τεχνολογίες τεχνητής νοημοσύνης αναπτύσσονται και εφαρμόζονται με τρόπους που σέβονται τα παγκόσμια ανθρώπινα δικαιώματα και αξίες.

3.6.2 Προκλήσεις στη Διαμόρφωση Παγκόσμιων Πολιτικών Τεχνητής Νοημοσύνης

Η διαμόρφωση παγκόσμιων πολιτικών για την τεχνητή νοημοσύνη παρουσιάζει πολλές προκλήσεις, κυρίως λόγω της ποικιλομορφίας των πολιτισμικών κανόνων, των νομικών συστημάτων και των οικονομικών συνθηκών μεταξύ των χωρών. Επιπλέον, ο γρήγορος ρυθμός της τεχνολογικής προόδου της τεχνητής νοημοσύνης συχνά ξεπερνά την ικανότητα των πλαισίων πολιτικής να συμβαδίζουν. Αυτή η ασυμφωνία μπορεί να οδηγήσει σε κενά στους κανονισμούς, αφήνοντας ορισμένα ηθικά και κοινωνικά ζητήματα αδιευκρίνιστα. Επιπλέον, η παγκόσμια συναίνεση σχετικά με την ηθική και τις πολιτικές της τεχνητής νοημοσύνης είναι δύσκολο να επιτευχθεί λόγω των ποικίλων προτεραιοτήτων και συμφερόντων διαφορετικών εθνών και ενδιαφερομένων.

3.6.3 Σημασία της Διεθνούς Συνεργασίας

Η διεθνής συνεργασία είναι απαραίτητη για την ανάπτυξη αποτελεσματικής παγκόσμιας ηθικής και πολιτικών για την τεχνητή νοημοσύνη. Αυτή η συνεργασία περιλαμβάνει όχι μόνο κυβερνήσεις αλλά και διεθνείς οργανισμούς, τον ιδιωτικό τομέα, την ακαδημαϊκή κοινότητα και την κοινωνία των πολιτών. Όπως τονίστηκε από τους Jobin, Ienca και Vayena (2019), απαιτούνται συνεργατικές προσπάθειες για τη δημιουργία εναρμονισμένων προτύπων και αρχών που μπορούν να καθοδηγήσουν την ηθική ανάπτυξη και χρήση της τεχνητής νοημοσύνης παγκοσμίως. Μια τέτοια συνεργασία μπορεί να βοηθήσει να διασφαλιστεί ότι οι τεχνολογίες τεχνητής νοημοσύνης ευθυγραμμίζονται με ευρύτερα παγκόσμια συμφέροντα, όπως η αειφόρος ανάπτυξη και η ανθρώπινη ευημερία.

3.6.4 Ανάπτυξη Πολιτικών για την Τεχνητή Νοημοσύνη χωρίς Αποκλεισμούς και Πολιτισμικής Ευαισθησίας

Η δημιουργία περιεκτικών και πολιτιστικά ευαίσθητων πολιτικών τεχνητής νοημοσύνης είναι ζωτικής σημασίας για την παγκόσμια αποδοχή και αποτελεσματικότητά τους. Οι πολιτικές πρέπει να λαμβάνουν υπόψη τα διαφορετικά πολιτιστικά πλαίσια και τα κοινωνικά πρότυπα στα οποία λειτουργούν τα συστήματα τεχνητής νοημοσύνης. Αυτή η περιεκτικότητα διασφαλίζει ότι οι τεχνολογίες τεχνητής νοημοσύνης δεν διακονίζουν ακούσια πολιτιστικές προκαταλήψεις ή ανισότητες. Η συμμετοχικότητα στην ανάπτυξη πολιτικής συνεπάγεται επίσης τη δέσμευση ενός

ευρέος φάσματος ενδιαφερομένων, διασφαλίζοντας ότι λαμβάνονται υπόψη διαφορετικές προοπτικές στη διατύπωση κατευθυντήριων γραμμών και κανονισμών τεχνητής νοημοσύνης.

3.6.5 Ο Ρόλος των Διεθνών Ρυθμιστικών Φορέων

Οι διεθνείς ρυθμιστικοί φορείς διαδραματίζουν κρίσιμο ρόλο στη διαμόρφωση της παγκόσμιας δεοντολογίας και πολιτικών για την τεχνητή νοημοσύνη. Αυτοί οι φορείς μπορούν να διευκολύνουν το διάλογο, να ανταλλάσσουν βέλτιστες πρακτικές και να αναπτύξουν πλαίσια που μπορούν να προσαρμοστούν από διαφορετικές χώρες στα συγκεκριμένα τους πλαίσια. Το έργο τέτοιων φορέων είναι καθοριστικό για τη γεφύρωση του χάσματος μεταξύ των τεχνολογικών προόδων και της ηθικής διακυβέρνησης, παρέχοντας μια πλατφόρμα για παγκόσμια συνεργασία σχετικά με την ηθική και την πολιτική της τεχνητής νοημοσύνης.

3.6.6 Μελλοντικές Κατευθύνσεις για την Παγκόσμια Ηθική και Πολιτική της Τεχνητής Νοημοσύνης

Το μέλλον της παγκόσμιας ηθικής και πολιτικής της τεχνητής νοημοσύνης πιθανότατα θα περιλαμβάνει συνεχείς προσαρμογές και βελτιώσεις ώστε να συμβαδίζουν με τις τεχνολογικές εξελίξεις. Θα απαιτήσει επίσης συνεχή προσπάθεια για την εξισορρόπηση της τεχνολογικής καινοτομίας με ηθικούς λόγους, διασφαλίζοντας ότι η ανάπτυξη της τεχνητής νοημοσύνης καθοδηγείται από αρχές που προάγουν το κοινό καλό. Η διατήρηση μιας προσαρμοστικής και ανταποκρινόμενης προσέγγισης στην παγκόσμια πολιτική τεχνητής νοημοσύνης θα είναι το κλειδί για την αντιμετώπιση των αναδυόμενων ηθικών προκλήσεων και την αξιοποίηση των δυνατοτήτων της τεχνητής νοημοσύνης για θετικό παγκόσμιο αντίκτυπο.

Η παγκόσμια ηθική και πολιτική για την τεχνητή νοημοσύνη είναι ουσιαστικής σημασίας για τη διαχείριση του εκτεταμένου αντίκτυπου των τεχνολογιών τεχνητής νοημοσύνης. Η ανάπτυξη και η εφαρμογή αυτών των κατευθυντήριων γραμμών απαιτεί διεθνή συνεργασία, συμπερίληψη και πολιτιστική ευαισθησία. Καθώς η τεχνητή νοημοσύνη συνεχίζει να προχωρά και να διεισδύει σε διάφορες πτυχές της παγκόσμιας κοινωνίας, η ανάγκη για συνολική και συνεκτική παγκόσμια ηθική και πολιτικές γίνεται όλο και πιο σημαντική για να διασφαλιστεί ότι η τεχνητή νοημοσύνη είναι μια δύναμη για το καλό στον κόσμο.

4

Περιπτώσιολογικές Μελέτες στην Ηθική της Τεχνητής Νοημοσύνης

4.1 Εισαγωγή στις Μελέτες Περίπτωσης Ηθικών Διλημμάτων

Το παρόν κεφάλαιο επιδιώκει να εφαρμόσει τα συμπεράσματα της θεωρητικής συζήτησης για την ηθική της Τεχνητής Νοημοσύνης σε συγκεκριμένες καταστάσεις όπου προκύπτουν σοβαρά ηθικά διλήμματα. Αφού στα προηγούμενα κεφάλαια εξετάστηκαν οι γενικές έννοιες, οι τεχνολογικές εξελίξεις και οι βασικές αρχές που διέπουν την ηθική της TN, στο παρόν κεφάλαιο η προσοχή στρέφεται σε πραγματικά ή υποθετικά σενάρια, τα οποία φέρνουν στο προσκήνιο τις πιο κρίσιμες ηθικές πτυχές της. Οι μελέτες περιπτώσεων επιλέχθηκαν ώστε να αντικατοπτρίζουν τη μεγάλη ποικιλία τομέων στους οποίους η TN έχει ήδη ή πρόκειται να έχει σημαντική επίδραση, όπως η επιβολή του νόμου (π.χ. προγνωστικά συστήματα εγκληματικότητας), τα αυτόνομα όπλα που επαναπροσδιορίζουν την έννοια της πολεμικής εμπλοκής, η χρήση αλγορίθμων στη διαδικασία ιατρικής διάγνωσης και η συλλογή και επεξεργασία προσωπικών δεδομένων σε κλίμακα (Bergold & Kovera, 2024, Birhane, 2021).

Επιδιώκεται, συνεπώς, να αναδειχθεί ο τρόπος με τον οποίο οι γενικές ηθικές αρχές εφαρμόζονται ή δοκιμάζονται όταν ο αλγόριθμος καλείται να αντιμετωπίσει σύνθετα ζητήματα. Μέσα από την ανάλυση αυτών των περιπτώσεων, θα επιχειρηθεί να φανεί πώς οι αποφάσεις και τα διλήμματα που ανακύπτουν μπορούν να διαχειριστούν με βάση τη βιβλιογραφία και τις ρυθμιστικές προτάσεις που έχουν ήδη παρουσιαστεί, παρέχοντας έτσι χρήσιμες κατευθύνσεις για μια υπεύθυνη εξέλιξη της TN (Birhane, 2021).

4.2 Τεχνητή Νοημοσύνη στην Επιβολή του Νόμου

Η χρήση της Τεχνητής Νοημοσύνης στην επιβολή του νόμου έχει ανοίξει νέους ορίζοντες στη συλλογή πληροφοριών και τη διαμόρφωση στρατηγικών για την αντιμετώπιση της

εγκληματικότητα, ωστόσο γεννά και μια σειρά από έντονα ηθικά διλήμματα. Συγκεκριμένα, τα συστήματα προληπτικής αστυνόμευσης (predictive policing) στηρίζονται σε αλγορίθμους που αναλύουν τεράστιες βάσεις δεδομένων προκειμένου να εντοπίσουν “υποψήφιος” περιοχές ή άτομα με αυξημένη πιθανότητα διάπραξης αδικημάτων. Παρότι αυτό θεωρητικά θα μπορούσε να οδηγήσει σε πιο αποτελεσματική κατανομή των αστυνομικών πόρων, οι ενδείξεις αλγοριθμικής προκατάληψης είναι ανησυχητικές. Ιστορικά, οι βάσεις δεδομένων που τροφοδοτούν τους αλγορίθμους μπορεί να αντικατοπτρίζουν προκατειλημμένες αστυνομικές πρακτικές, με αποτέλεσμα να αναπαράγουν διακρίσεις εις βάρος μειονοτήτων ή συγκεκριμένων γεωγραφικών περιοχών. Έτσι, η τεχνολογία κινδυνεύει να εδραιώσει συστημικές αδικίες αντί να εξαλείψει τα αίτια της εγκληματικότητας (Brayne & Christin, 2020).

Ένα άλλο χαρακτηριστικό παράδειγμα εφαρμογής της TN στην επιβολή του νόμου είναι η αναγνώριση προσώπου, που χρησιμοποιείται όλο και συχνότερα από τις αρχές για την ταυτοποίηση υπόπτων. Παρότι υπόσχεται ταχύτητα και ακρίβεια, η τεχνολογία αυτή έχει αποδειχθεί σε μελέτες πως παρουσιάζει μεγαλύτερα ποσοστά σφαλμάτων στους σκουρόχρωμους πληθυσμούς, γεγονός που εγείρει ζήτημα φυλετικών διακρίσεων και άδικων προσαγωγών. Επιπλέον, η εκτεταμένη χρήση συστημάτων αναγνώρισης προσώπου σε δημόσιους χώρους δημιουργεί ανησυχίες σχετικά με το απόρρητο, καθώς διευκολύνει την αδιάκριτη παρακολούθηση πολιτών.

Ένα πρόσθετο πεδίο προβληματισμού αφορά τους αλγορίθμους εκτίμησης επικινδυνότητας (risk assessment), που χρησιμοποιούνται σε κάποιες χώρες για την υπό όρους αποφυλάκιση ή την επιβολή ποινής. Περιπτώσεις όπως το σύστημα COMPAS στις ΗΠΑ, το οποίο επιχείρησε να προβλέψει την πιθανότητα επανάληψης εγκλήματος, καταδεικνύουν σοβαρά ζητήματα μεροληψίας εις βάρος μειονοτήτων, καθώς ο αλγόριθμος “βαθμολογούσε” δυσανάλογα αυστηρά ορισμένες κοινωνικές ομάδες. Η έλλειψη διαφάνειας σε τέτοιου είδους συστήματα δημιουργεί επιπρόσθετη σύγχυση, καθώς δικηγόροι, δικαστές και κατηγορούμενοι δεν είναι σε θέση να ελέγξουν πώς εξάγεται ένα κρίσιμο αποτέλεσμα που επηρεάζει καθοριστικά την ελευθερία ή την τιμωρία ενός ατόμου. Συνολικά, η TN στην επιβολή του νόμου προσφέρει μεν σημαντικές δυνατότητες, ταυτόχρονα όμως απαιτεί τη θέσπιση αυστηρών ηθικών και νομικών πλαισίων, με στόχο να διασφαλιστεί ότι η νέα αυτή τεχνολογία θα υπηρετεί αρχές δικαιοσύνης, διαφάνειας και σεβασμού των ανθρωπίνων δικαιωμάτων (Joh, 2018; Richardson et al 2019).

Πίνακας 3. Η χρήση της Τεχνητής Νοημοσύνης στην Επιβολή του Νόμου.

Πεδίο Εφαρμογής	Πλεονεκτήματα	Προβληματισμοί / Κίνδυνοι	Ενδεικτικά Παραδείγματα / Μελέτες
Προληπτική Αστυνόμευση (Predictive Policing)	- Αποτελεσματικότερη κατανομή αστυνομικών πόρων.	Προληπτική Αστυνόμευση (Predictive Policing)	- Αποτελεσματικότερη κατανομή αστυνομικών πόρων.
Αναγνώριση Προσώπου	- Ταχεία ταυτοποίηση υπόπτων.	Αναγνώριση Προσώπου	- Ταχεία ταυτοποίηση υπόπτων.
Αλγόριθμοι Εκτίμησης Επικινδυνότητας (Risk Assessment)	- Πρόβλεψη πιθανότητας επανάληψης εγκλήματος (π.χ. αποφυλάκιση υπό όρους).	Αλγόριθμοι Εκτίμησης Επικινδυνότητας (Risk Assessment)	- Πρόβλεψη πιθανότητας επανάληψης εγκλήματος (π.χ. αποφυλάκιση υπό όρους).
Ηθικά και Νομικά Ζητήματα	- Δυνατότητα ανάλυσης μεγάλων όγκων δεδομένων προς όφελος της δημόσιας ασφάλειας.	Ηθικά και Νομικά Ζητήματα	- Δυνατότητα ανάλυσης μεγάλων όγκων δεδομένων προς όφελος της δημόσιας ασφάλειας.

4.3 Αυτόνομες Οπλικές Τεχνολογίες

Τα αυτόνομα οπλικά συστήματα αποτελούν μία από τις πλέον αμφιλεγόμενες εφαρμογές της Τεχνητής Νοημοσύνης (TN) στον στρατιωτικό τομέα, καθώς επιτρέπουν σε ρομποτικές

πλατφόρμες να λαμβάνουν θανατηφόρες αποφάσεις χωρίς άμεση ανθρώπινη παρέμβαση (Horowitz, 2019). Η ικανότητα ενός συστήματος να αξιολογεί καταστάσεις και να λαμβάνει αποφάσεις που ενδέχεται να οδηγήσουν σε απώλεια ανθρώπινης ζωής εγείρει κρίσιμα ηθικά ερωτήματα. Κατ' αρχάς, τίθεται ζήτημα ανάληψης ευθύνης: ποιος λογοδοτεί σε περίπτωση λανθασμένου πλήγματος ή ακούσιας απώλειας αθώων ζωών; Επιπλέον, η απουσία ανθρώπινης κρίσης στον κύκλο λήψης αποφάσεων περιορίζει την ηθική κρίση που συνήθως συνοδεύει την ανθρώπινη συμμετοχή στον πόλεμο. Η ΤΝ, όταν καλείται να “κρίνει” αυτόνομα, δεν διαθέτει τις απαραίτητες ικανότητες κατανόησης των συμφραζομένων, ενσωμάτωσης πολιτισμικών παραμέτρων ή εκτίμησης του κινδύνου κλιμάκωσης της βίας, γεγονός που εγείρει σημαντικές ηθικές ανησυχίες (Roff, 2019).

Στο διεθνές επίπεδο, έχουν διατυπωθεί σοβαρές ανησυχίες σχετικά με το ενδεχόμενο κούρσας εξοπλισμών σε ό,τι αφορά τα αυτόνομα όπλα. Πολλοί ειδικοί φοβούνται ότι όσο αναπτύσσονται εξελιγμένα συστήματα ΤΝ για πολεμικούς σκοπούς, τόσο μειώνεται ο χρόνος διαθέσιμης αντίδρασης για τον ανθρώπινο παράγοντα, αυξάνοντας τον κίνδυνο ανεξέλεγκτων συγκρούσεων (de Neufville & Baum, 2021). Παράλληλα, η ευκολία χρήσης και η αναλογικά μειωμένη ανθρώπινη δέσμευση σε αυτόνομα όπλα θα μπορούσε να ενθαρρύνει τη συχνότερη προσφυγή σε στρατιωτικές λύσεις, αντί για διπλωματικές προσεγγίσεις. Η αρχή της διάκρισης και της αναλογικότητας, θεμελιώδεις έννοιες του Διεθνούς Ανθρωπιστικού Δικαίου, τίθενται υπό αμφισβήτηση όταν η τελική απόφαση για την επιλογή στόχων λαμβάνεται από αλγόριθμους που ίσως βασίζονται σε ελλιπή ή αλλοιωμένα δεδομένα.

Οι συζητήσεις στον Οργανισμό Ηνωμένων Εθνών (ΟΗΕ) σχετικά με τον έλεγχο ή ακόμα και την πλήρη απαγόρευση τέτοιων αυτόνομων συστημάτων υπογραμμίζουν τη σοβαρότητα των ηθικών διλημμάτων. Αρκετές χώρες, αλλά και μη κυβερνητικές οργανώσεις, υποστηρίζουν τη θέσπιση σαφούς νομικού πλαισίου για τη ρύθμιση της ΤΝ σε στρατιωτικές εφαρμογές, καθώς και τη διασφάλιση ότι πάντα θα υπάρχει ανθρώπινος παράγοντας στον κύκλο λήψης θανατηφόρων αποφάσεων (Horowitz, 2019). Άλλες όμως αντιτείνουν ότι η πλήρης απαγόρευση θα μπορούσε να επιβραδύνει τεχνολογικές καινοτομίες που θα ενισχύσουν τις αμυντικές δυνατότητες χωρών με πιο περιορισμένους πόρους. Σε κάθε περίπτωση, η διεθνής κοινότητα αναζητά μια ισορροπία μεταξύ καινοτομίας και ηθικής ευθύνης, τονίζοντας ότι η ανάπτυξη της ΤΝ για στρατιωτικούς σκοπούς πρέπει να διέπεται από διαφάνεια, λογοδοσία και σεβασμό των ανθρωπίνων δικαιωμάτων. Η ένταση αυτής της συζήτησης μαρτυρά πως το μέλλον των αυτόνομων οπλικών τεχνολογιών θα εξαρτηθεί σε μεγάλο βαθμό από τη βούληση των κρατών και των διεθνών οργανισμών να

υιοθετήσουν αποτελεσματικές ρυθμίσεις που θα προλαμβάνουν καταχρήσεις και θα διασφαλίζουν την ανθρώπινη αξιοπρέπεια (de Neufville & Baum, 2021).

Πίνακας 4: Ζητήματα Ηθικής, Νομικής και Διεθνούς Διάστασης των Αυτόνομων Οπλικών Τεχνολογιών.

Θεματική Κατηγορία	Βασικά Σημεία / Ζητήματα
Ορισμός και Περιγραφή	Οπλικά συστήματα που βασίζονται στην τεχνητή νοημοσύνη και λαμβάνουν αποφάσεις χρήσης θανατηφόρας βίας χωρίς άμεση ανθρώπινη παρέμβαση.
Ηθικά Ζητήματα	- Ανάλυση ευθύνης και λογοδοσία για λανθασμένες αποφάσεις. - Έλλειψη ανθρώπινης κρίσης και ηθικής διάστασης. - Αδυναμία των μηχανών να αξιολογήσουν συμφραζόμενα, πολιτισμικές παραμέτρους και κίνδυνο κλιμάκωσης βίας.
Διεθνείς Ανησυχίες	- Κίνδυνος κούρσας εξοπλισμών. - Μειωμένος χρόνος αντίδρασης του ανθρώπινου παράγοντα που μπορεί να οδηγήσει σε ανεξέλεγκτες συγκρούσεις. - Ευκολότερη προσφυγή στη στρατιωτική δράση αντί της διπλωματίας.
Νομικά και Ανθρωπιστικά Ζητήματα	- Αμφισβήτηση αρχών Διεθνούς Ανθρωπιστικού Δικαίου (διάκριση, αναλογικότητα). - Λήψη αποφάσεων από αλγόριθμους με πιθανώς ανεπαρκή ή αλλοιωμένα δεδομένα.
Στάση Διεθνούς Κοινότητας και Προτάσεις	- Συζητήσεις στον ΟΗΕ για έλεγχο ή πλήρη απαγόρευση.

	- Υποστήριξη θέσπισης σαφούς νομικού πλαισίου και απαίτησης ανθρώπινης παρέμβασης. - Αντίθετες απόψεις για την αναγκαιότητα τεχνολογικής καινοτομίας.
Βασικές Προτεινόμενες Αρχές Ρύθμισης	- Διαφάνεια και λογοδοσία κατά την ανάπτυξη και χρήση. - Σεβασμός ανθρωπίνων δικαιωμάτων. - Διατήρηση της ανθρώπινης αξιοπρέπειας και εμπλοκής στον κύκλο λήψης αποφάσεων.

4.4 Τεχνητή Νοημοσύνη στην Ιατρική Διάγνωση

Η εφαρμογή συστημάτων Τεχνητής Νοημοσύνης (TN) στον τομέα της ιατρικής διάγνωσης υπόσχεται σημαντικά οφέλη, όπως η επιτάχυνση και βελτίωση της ακρίβειας στις κλινικές αποφάσεις, η μείωση του ανθρώπινου σφάλματος και η εξοικονόμηση πολύτιμων πόρων (CONSORT-AI and SPIRIT-AI Steering Group (2019)). Ειδικότερα, τεχνολογίες όπως η βαθιά μάθηση έχουν χρησιμοποιηθεί για την ανίχνευση παθήσεων με υψηλά ποσοστά ακρίβειας, συγκρίσιμα ή και ανώτερα από αυτά των έμπειρων ιατρών (Xia et. al, 2025). Παρά τις αισιόδοξες προοπτικές, όμως, προκύπτουν καίρια ηθικά διλήμματα. Ένα πρωταρχικό ζήτημα αφορά την αξιοπιστία των αλγορίθμων: ενώ μπορούν να βελτιωθούν διαρκώς μέσω της επεξεργασίας μεγάλου όγκου δεδομένων, παραμένει πιθανό να παράσχουν λανθασμένη διάγνωση σε ορισμένες περιπτώσεις. Όταν η διάγνωση είναι λανθασμένη ή ατελής, τίθεται το ερώτημα της ευθύνης: πρέπει να αναλάβει το βάρος ο ιατρός που βασίστηκε στο σύστημα, ο κατασκευαστής της TN ή θεωρείται “αναπόφευκτο” τεχνολογικό ρίσκο;

Ενδεικτικά, ένα παράδειγμα που συχνά αναφέρεται είναι η χρήση ενός αλγορίθμου για την ανίχνευση δερματικών βλαβών, ο οποίος παρήγαγε ψευδώς θετικές διαγνώσεις με αποτέλεσμα να συστήνονται περιττές βιοψίες σε ορισμένους ασθενείς (Ngiam & Khor, 2019). Παρότι ο σκοπός της τεχνολογίας ήταν να θωρακίσει τον ιατρικό κλάδο από σφάλματα και να ενισχύσει τη διαγνωστική ακρίβεια, η υπερβολική εμπιστοσύνη στο αποτέλεσμα της TN και η έλλειψη επαρκούς ελέγχου από τον άνθρωπο οδήγησαν σε ανεπιθύμητες συνέπειες. Τέτοια περιστατικά υπογραμμίζουν την ανάγκη διαφάνειας και επεξηγησιμότητας (explainability) των αλγορίθμων, ώστε οι ιατροί να μπορούν να αντιληφθούν τους παράγοντες που συντελούν σε μια συγκεκριμένη εκτίμηση. Επιπλέον, η

δυνατότητα αναθεώρησης των διαγνωστικών συμπερασμάτων από τον κλινικό γιατρό είναι απαραίτητη για την τήρηση της αρχής “*primum non nocere*” (πρώτα απ’ όλα να μην προκαλείς βλάβη).

Η σημασία της ηθικής επίβλεψης στην υγειονομική περίθαλψη αναδεικνύεται, λοιπόν, ως θεμελιώδης. Τα πρωτόκολλα ελέγχου ποιότητας, η συμμόρφωση με τους κανονισμούς προστασίας προσωπικών δεδομένων και η σαφής καθοδήγηση ως προς την ευθύνη (liability) αποτελούν ζωτικά στοιχεία για την ασφαλή ενσωμάτωση της TN στην ιατρική πρακτική. Παράλληλα, η εκπαίδευση των επαγγελματιών υγείας στη χρήση και την αξιολόγηση των αλγοριθμικών εργαλείων είναι κρίσιμη, καθώς η ορθή εφαρμογή προϋποθέτει συνδυασμό της κλινικής εμπειρίας και της τεχνολογικής καινοτομίας. Εν τέλει, η TN μπορεί να γίνει ένα πανίσχυρο εργαλείο που ενισχύει τον γιατρό, όχι όμως να τον αντικαθιστά πλήρως. Η συστηματική ηθική εποπτεία και η ανάπτυξη υπεύθυνων προτύπων διασφαλίζουν ότι οι κλινικές αποφάσεις θα λαμβάνονται με τρόπο που προάγει την υγεία και διαφυλάσσει την ανθρώπινη αξιοπρέπεια (Xia et al., 2025).

4.5 Διαχείριση Προσωπικών Δεδομένων και Ιδιωτικότητα

Η ευρεία αξιοποίηση της Τεχνητής Νοημοσύνης (TN) για τη διαχείριση και ανάλυση μεγάλων ποσοτήτων προσωπικών δεδομένων έχει προκαλέσει έντονο προβληματισμό σχετικά με την ιδιωτικότητα και την προστασία της πληροφορίας. Ιδιαίτερα σε πλατφόρμες κοινωνικής δικτύωσης ή σε υπηρεσίες που συγκεντρώνουν δεδομένα χρηστών, οι αλγόριθμοι μηχανικής μάθησης μπορούν να αποκαλύψουν ευαίσθητες πτυχές της ζωής ενός ατόμου, δημιουργώντας ακριβή προφίλ για στοχευμένη διαφήμιση ή ακόμα και επιτήρηση (Taddeo & Floridi, 2018). Ένα βασικό ηθικό ζήτημα αφορά την ενημερωμένη συγκατάθεση των χρηστών: συχνά, οι όροι χρήσης είναι περίπλοκοι ή ασαφείς, με αποτέλεσμα οι περισσότεροι άνθρωποι να μη γνωρίζουν πώς ακριβώς αξιοποιούνται τα προσωπικά τους δεδομένα από τους αλγορίθμους. Επιπλέον, η συνεχής συλλογή πληροφοριών σε πραγματικό χρόνο αυξάνει τον κίνδυνο διαρροών, καθώς τυχόν κενά ασφαλείας ή επιθέσεις χάκερ μπορεί να εκθέσουν απόρρητα στοιχεία χιλιάδων ή και εκατομμυρίων χρηστών (Morley et. al, 2020).

Παραδείγματα περιστατικών που έχουν ξεσηκώσει συζητήσεις σε παγκόσμιο επίπεδο είναι τα σκάνδαλα αξιοποίησης δεδομένων για πολιτικές εκστρατείες, καθώς και η πιθανή χειραγώγηση γνώμης μέσω αλγορίθμων που προωθούν συγκεκριμένο περιεχόμενο. Η στοχευμένη διαφήμιση, για παράδειγμα, βασίζεται στην ανάλυση των προτιμήσεων και των διαδικτυακών συνηθειών των χρηστών, με αποτέλεσμα να δημιουργείται μια σχεδόν “προσωποποιημένη” εμπειρία, η οποία όμως παραβιάζει σε μεγάλο βαθμό την ιδιωτική σφαίρα (Shilton, 2017). Ειδικότερα, εταιρείες που

διαθέτουν εξελιγμένες τεχνολογίες ΤΝ έχουν τη δυνατότητα να συσχετίζουν δεδομένα από πολλαπλές πηγές, δημιουργώντας προγνωστικά μοντέλα συμπεριφοράς. Αυτή η πρακτική εγκυμονεί κινδύνους, καθώς οι χρήστες συχνά δεν αντιλαμβάνονται το εύρος της ανάλυσης και του διαμοιρασμού των προσωπικών τους πληροφοριών.

Η ανάγκη για αυστηρότερα νομικά και ηθικά πλαίσια καθίσταται επιτακτική. Στην Ευρωπαϊκή Ένωση, ο Γενικός Κανονισμός για την Προστασία Δεδομένων (GDPR) εισήγαγε αρχές όπως η ελαχιστοποίηση δεδομένων και το δικαίωμα στη λήθη, ωστόσο οι ραγδαίες εξελίξεις στην ΤΝ καθιστούν αβέβαιο κατά πόσο οι ισχύοντες κανονισμοί μπορούν να ανταποκριθούν στις καινούριες τεχνολογικές πραγματικότητες. Η διαφάνεια στον τρόπο επεξεργασίας των δεδομένων, η ύπαρξη μηχανισμών λογοδοσίας και ο σαφής καθορισμός ευθυνών είναι κρίσιμοι παράγοντες για να κερδηθεί η εμπιστοσύνη των χρηστών. Όπως υπογραμμίζεται σε πληθώρα πρόσφατων μελετών, η ηθική διάσταση της ΤΝ αναδεικνύεται πιο επιτακτική από ποτέ, απαιτώντας συνεργασία μεταξύ νομοθετών, τεχνολόγων, ερευνητών και πολιτών, ώστε να εξασφαλιστεί ότι η καινοτομία στην αξιοποίηση δεδομένων προχωράει παράλληλα με τον σεβασμό των δικαιωμάτων και των ελευθεριών των ατόμων (Taddeo & Floridi, 2018).

4.6 Συγκριτική Συζήτηση και Συμπεράσματα Κεφαλαίου

Οι περιπτώσεις που παρουσιάστηκαν, από την επιβολή του νόμου και τα αυτόνομα οπτικά συστήματα έως την ιατρική διάγνωση και τη διαχείριση προσωπικών δεδομένων, αναδεικνύουν κοινά σημεία που επαναλαμβάνονται στα ηθικά διλήμματα σχετικά με την Τεχνητή Νοημοσύνη. Η ανάγκη για διαφάνεια ως προς τους αλγόριθμους και τις αποφάσεις τους εμφανίζεται σταθερά σε κάθε τομέα, είτε πρόκειται για προβλέψεις εγκληματικότητας είτε για διάγνωση ασθενειών. Παράλληλα, η κατανομή της ευθύνης σε περίπτωση λανθασμένων ή επιβλαβών αποφάσεων δημιουργεί σοβαρούς προβληματισμούς, αφού οι επιλογές της ΤΝ επηρεάζουν άμεσα ανθρώπινες ζωές, αλλά ο <εντολέας> της τεχνολογίας αλλά συχνά παραμένει ασαφές ποιος φορέας ή ποιο πρόσωπο φέρει την τελική ευθύνη.

Επιπλέον, η προκατάληψη των αλγορίθμων (bias) αποτελεί διαρκή πηγή ανησυχίας, καθώς επηρεάζει τη δίκαιη αντιμετώπιση εύάλωτων ομάδων και την αξιοπιστία των συστημάτων σε κρίσιμα περιβάλλοντα (Morley et al., 2020). Συνολικά, οι μελέτες περιπτώσεων καταδεικνύουν ότι η ενσωμάτωση ηθικών αρχών δεν είναι απλώς θεωρητική αναγκαιότητα, αλλά ουσιαστική προϋπόθεση για την κοινωνικά δίκαιη ανάπτυξη και χρήση της ΤΝ (Eitel-Porter, 2020). Μέσα από αυτές τις εφαρμογές διαφαίνεται το εύρος των συνεπειών που δύνανται να προκύψουν, επιβεβαιώνοντας τη σημασία της διεθνούς συνεργασίας, της νομοθετικής ρύθμισης και της

ουσιαστικής εποπτείας. Αυτή η σύντομη συγκριτική ματιά ολοκληρώνει το κεφάλαιο, προετοιμάζοντας τη συζήτηση για τις αναδυόμενες τάσεις και καινοτομίες που θα εξεταστούν στο επόμενο στάδιο της ανάλυσης.

5

Αναδυόμενες Τάσεις και

Μελλοντικές Κατευθύνσεις στην

Ηθική της Τεχνητής

Νοημοσύνης

5.1 Εισαγωγή στο μέλλον της ηθικής της Τεχνητής Νοημοσύνης

Η τεχνολογία της Τεχνητής Νοημοσύνης (TN) εξελίσσεται ραγδαία, ανοίγοντας νέους ορίζοντες για την ανθρώπινη καινοτομία αλλά και προκαλώντας ολοένα και πιο σύνθετα ηθικά διλήμματα. Οι πρόσφατες πρόοδοι στην επεξεργασία μεγάλων δεδομένων, τη βαθιά μάθηση και την ενισχυτική μάθηση οδηγούν σε εφαρμογές που επηρεάζουν άμεσα ή έμμεσα την ανθρώπινη ζωή, από τη βελτιστοποίηση βιομηχανικών διαδικασιών, μέχρι την υποστήριξη κλινικών αποφάσεων (Jobin et. al, 2019). Ωστόσο, όσο οι δυνατότητες της TN διευρύνονται, τόσο εντείνονται οι ανησυχίες σχετικά με την ιδιωτικότητα, την ευθύνη και τη διαφάνεια. Στόχος αυτού του κεφαλαίου είναι η διερεύνηση των νεότερων τεχνολογικών εξελίξεων στην TN και των αντίστοιχων ηθικών προκλήσεων που εγείρουν. Παράλληλα, υπογραμμίζεται η ανάγκη υιοθέτησης μιας προληπτικής και προσαρμοστικής ηθικής προσέγγισης, ικανής να ανταποκρίνεται σε ένα συνεχώς μεταβαλλόμενο περιβάλλον (Eitel-Porter, 2021). Η συζήτηση αυτή καταδεικνύει ότι η ηθική της TN δεν μπορεί να περιορίζεται σε εκ των υστέρων ρυθμίσεις, αλλά απαιτεί στρατηγικό σχεδιασμό που προλαμβάνει ή ελαχιστοποιεί πιθανούς κινδύνους.

5.2 Νέες Εξελίξεις στην Τεχνητή Νοημοσύνη και οι Ηθικές τους Προκλήσεις

Η ταχύτατη ανάπτυξη της προηγμένης μηχανικής μάθησης σε συνδυασμό με την εκτεταμένη αξιοποίηση μεγάλων δεδομένων (Big Data) έχει αναδείξει νέες προοπτικές, αλλά και σύνθετα ηθικά ζητήματα. Η συλλογή και επεξεργασία τεράστιων όγκων πληροφορίας δίνει τη δυνατότητα σε

αλγόριθμους βαθιάς μάθησης να επιτύχουν υψηλή ακρίβεια στην πρόβλεψη και ταξινόμηση, συμβάλλοντας σημαντικά στην αυτοματοποίηση κρίσιμων διαδικασιών, από τη διάγνωση ασθενειών μέχρι τη χρηματοοικονομική αξιολόγηση (Morley et. al, 2020). Ωστόσο, η αποτελεσματικότητα των αλγορίθμων εξαρτάται σε μεγάλο βαθμό από τα δεδομένα με τα οποία εκπαιδεύονται. Όταν αυτά τα δεδομένα αντικατοπτρίζουν κοινωνικές ανισότητες ή στερεοτυπικές αντιλήψεις, οι αλγόριθμοι τείνουν να αναπαράγουν και να ενισχύουν αυτές τις προκαταλήψεις, οδηγώντας σε άδικες ή μεροληπτικές αποφάσεις εις βάρος συγκεκριμένων πληθυσμιακών ομάδων (Birhane, 2021).

Παράλληλα, η πολυπλοκότητα των σύγχρονων μοντέλων βαθιάς μάθησης συχνά τα μετατρέπει σε «μαύρα κουτιά», δυσχεραίνοντας την ερμηνεία των αποφάσεων που λαμβάνουν. Η έλλειψη επεξηγησιμότητας προκαλεί ηθικούς και νομικούς προβληματισμούς, καθώς οι ενδιαφερόμενοι, από τους χρήστες και τους πελάτες μέχρι τις αρχές ρύθμισης, αδυνατούν να κατανοήσουν ή να αμφισβητήσουν τα αποτελέσματα ενός αλγορίθμου. Σε ευαίσθητους τομείς όπως η υγεία, η δικαιοσύνη και η εύρεση εργασίας, η αδιαφανής λήψη αποφάσεων πλήττει την εμπιστοσύνη στο σύστημα και δυσχεραίνει την απόδοση ευθυνών σε περιπτώσεις σφάλματος (Birhane, 2021).

Επιπλέον, η συγκέντρωση δεδομένων σε τόσο μεγάλη κλίμακα αυξάνει τον κίνδυνο για παραβιάσεις ιδιωτικότητας, αφού κυβερνοεπιθέσεις και διαρροές δεδομένων μπορούν να εκθέσουν προσωπικές πληροφορίες εκατομμυρίων ανθρώπων. Ως αποτέλεσμα, αναδεικνύεται επιτακτική η ανάγκη για ηθικές κατευθυντήριες γραμμές και νομικά πλαίσια που να διασφαλίζουν τη δίκαιη, διαφανή και ασφαλή χρήση της προηγμένης μηχανικής μάθησης και των Big Data, με έμφαση στην υπευθυνότητα όλων των εμπλεκόμενων φορέων (Morley et. al, 2020).

Η ραγδαία εξέλιξη των δημιουργικών συστημάτων Τεχνητής Νοημοσύνης (Generative AI) έχει φέρει στο προσκήνιο ένα σύνολο ηθικών και κοινωνικών προκλήσεων. Ειδικότερα, οι αλγόριθμοι που παράγουν κείμενο, εικόνες, βίντεο ή ήχο με τρόπο που προσομοιάζει έντονα το ανθρώπινο έργο έχουν τη δυνατότητα να δημιουργήσουν ρεαλιστικά αλλά ψευδή ή παραπλανητικά περιεχόμενα, γνωστά ως deepfakes (συνθετικά παραπονημένα πολυμέσα) (Vaccari & Chadwick, 2020). Παρόμοια προηγμένα γλωσσικά μοντέλα μπορούν να παράγουν ευρύ φάσμα κειμένων, ενισχύοντας τη δημιουργικότητα και την καινοτομία, αλλά ταυτόχρονα αυξάνουν τον κίνδυνο διάδοσης παραπληροφόρησης και χειραγώγησης της κοινής γνώμης (Floridi & Chiriatti, 2020). Η δυνατότητα παρωδίας ή ακόμη και πλήρους παραποίησης της φωνής και της εικόνας ενός ατόμου εντείνει τις ανησυχίες για παραβίαση της ιδιωτικότητας, καθώς και για την πιθανή υπονόμευση της εμπιστοσύνης στην αυθεντικότητα ψηφιακών μέσων.

Παράλληλα, η δημιουργία περιεχομένου που βασίζεται σε δείγματα τα οποία ενδέχεται να προστατεύονται από πνευματικά δικαιώματα προκαλεί ερωτήματα σχετικά με την κυριότητα του παραγόμενου έργου και το κατά πόσο θα έπρεπε να αποδίδονται πνευματικά δικαιώματα στους προγραμματιστές, στους χρήστες ή ακόμη και στα ίδια τα συστήματα ΤΝ. Η ηθική και νομική συζήτηση καθίσταται ακόμη πιο περίπλοκη, δεδομένου ότι η τεχνολογία αυτή μπορεί να εφαρμοστεί είτε για καινοτόμους σκοπούς (π.χ. καλλιτεχνική δημιουργία, εκπαίδευση) είτε για κακόβουλες ενέργειες (Chesney & Citron, 2019). Συνολικά, η διαρκής βελτίωση και ευρεία διάδοση της Δημιουργικής ΤΝ επιβάλλει την υιοθέτηση αυστηρών ρυθμιστικών πλαισίων και κωδίκων δεοντολογίας, ώστε να περιοριστούν τα φαινόμενα παραπλάνησης και να προστατευθούν τα δικαιώματα τόσο των δημιουργών όσο και των πολιτών.

Η ένταξη της Τεχνητής Νοημοσύνης (ΤΝ) στο Διαδίκτυο των Πραγμάτων (IoT) έχει επιταχύνει τη μετάβαση σε έξυπνες πόλεις, «έξυπνα» οικιακά περιβάλλοντα και αυτόνομα οχήματα, δημιουργώντας νέες προκλήσεις για την ηθική και την ασφάλεια. Σε επίπεδο έξυπνων πόλεων, η συνεχής συλλογή και ανάλυση δεδομένων από αισθητήρες επιτρέπει την πιο αποτελεσματική διαχείριση πόρων, αλλά θέτει ερωτήματα σχετικά με τη διαφύλαξη της ιδιωτικότητας, καθώς κάμερες και συστήματα παρακολούθησης μπορούν να δημιουργήσουν ένα καθεστώς διαρκούς επιτήρησης (Kitchin, 2016). Οι οικιακές συσκευές που λειτουργούν με ΤΝ, όπως «έξυπνα» ψυγεία ή ηχεία με φωνητικούς βοηθούς, συλλέγουν συνεχώς πληροφορίες για τις καταναλωτικές συνήθειες ή τις προτιμήσεις των κατοίκων, γεγονός που εγείρει ανησυχίες ως προς την πιθανή εκμετάλλευση ή διαρροή των δεδομένων (Li et. al, 2016).

Εξίσου κρίσιμο ζήτημα αποτελεί η ασφάλεια αυτών των συστημάτων, ιδίως αν ληφθεί υπόψη η προοπτική χακαρίσματος που θα επέτρεπε σε κακόβουλους χρήστες να αποκτήσουν έλεγχο όχι μόνο σε προσωπικές πληροφορίες, αλλά και σε φυσικές υποδομές (π.χ. φωτεινούς σηματοδότες, ενεργειακά δίκτυα). Στην περίπτωση των αυτόνομων οχημάτων, η εξάρτηση από σύνθετους αλγορίθμους και αισθητήρες απαιτεί ενδελεχή έλεγχο αξιοπιστίας, ώστε να αποτραπούν ατυχήματα ή επικίνδυνες βλάβες (Kalisa & Extension, Kiu Publication, 2024). Την ίδια στιγμή, η μαζική υιοθέτηση αυτοματοποιημένων διαδικασιών και αποφάσεων αναδεικνύει τον κίνδυνο απώλειας του ανθρώπινου ελέγχου, καθώς οι πολίτες στηρίζονται ολοένα και περισσότερο σε συστήματα που λαμβάνουν πρωτοβουλίες για λογαριασμό τους. Σε ένα τέτοιο περιβάλλον, η ισορροπία ανάμεσα στα οφέλη και στους πιθανούς κινδύνους της ΤΝ στο IoT απαιτεί την ανάπτυξη ισχυρών ρυθμιστικών πλαισίων, κωδίκων δεοντολογίας και τεχνολογικών προτύπων που διασφαλίζουν τη διαφάνεια, την ιδιωτικότητα και τη λογοδοσία.

Πίνακας 5: Νέες Εξελίξεις στην Τεχνητή Νοημοσύνη και Κύριες Ηθικές Προκλήσεις.

Θεματικός Τομέας	Τεχνολογικές Εξελίξεις	Κύριες Ηθικές Προκλήσεις
Μηχανική Μάθηση και Big Data	Αλγόριθμοι βαθιάς μάθησης και μεγάλα σύνολα δεδομένων	Μεροληψία, αναπαραγωγή στερεοτύπων, έλλειψη επεξηγησιμότητας, ιδιωτικότητα δεδομένων
Generative AI (Δημιουργική TN)	Παραγωγή κειμένων, εικόνων, ήχου, βίντεο, deepfakes	Παραπληροφόρηση, παραβίαση ιδιωτικότητας, πνευματικά δικαιώματα, ζητήματα εμπιστοσύνης
Τεχνητή Νοημοσύνη και IoT	Έξυπνες πόλεις, έξυπνες συσκευές, αυτόνομα οχήματα	Παρακολούθηση και διαρκής επιτήρηση, ασφάλεια δεδομένων, κίνδυνος κυβερνοεπιθέσεων, απώλεια ανθρώπινου ελέγχου

5.3 Ρυθμιστικά Πλαίσια και Πολιτικές

5.3.1 Οι Πρωτοβουλίες στην Ευρωπαϊκή Ένωση

Οι προσπάθειες ρύθμισης της Τεχνητής Νοημοσύνης (TN) εντείνονται διεθνώς, καθώς κυβερνήσεις και υπερεθνικοί οργανισμοί αναγνωρίζουν την ανάγκη για ένα ισχυρό πλαίσιο που θα διασφαλίζει την ηθική και υπεύθυνη ανάπτυξη της τεχνολογίας. Στην Ευρωπαϊκή Ένωση, ένα χαρακτηριστικό παράδειγμα είναι ο προτεινόμενος Κανονισμός για την Τεχνητή Νοημοσύνη (EU AI Act), ο οποίος εισάγει μια πολυεπίπεδη προσέγγιση κινδύνου για τις εφαρμογές της TN (Floridi, 2020). Σύμφωνα με αυτόν, ορισμένες χρήσεις, όπως τα συστήματα «κοινωνικής βαθμολόγησης», προτείνεται να απαγορευτούν πλήρως, ενώ για τις εφαρμογές υψηλού κινδύνου, όπως τα ιατρικά διαγνωστικά ή τα αυτόνομα οχήματα, καθιερώνεται υποχρεωτική αξιολόγηση κινδύνου και αυστηρές απαιτήσεις διαφάνειας. Επιπλέον, τίθενται όρια στην παρακολούθηση σε δημόσιους χώρους και προβλέπεται ρύθμιση για την επεξηγησιμότητα των συστημάτων TN, ώστε να προστατεύονται τα θεμελιώδη δικαιώματα των πολιτών.

Η εξέλιξη της Τεχνητής Νοημοσύνης έχει οδηγήσει στη μετάβαση από γενικές ηθικές συστάσεις προς πιο οργανωμένα πλαίσια διακυβέρνησης. Η ανάγκη αυτή προέκυψε λόγω της αυξανόμενης επιρροής της TN σε κρίσιμους τομείς της κοινωνίας και της οικονομίας.

Η προσέγγιση που αποτυπώνεται στο AI Act βασίζεται στην ιδέα ότι η καινοτομία και η προστασία των πολιτών δεν αποτελούν αντικρουόμενους στόχους. Αντίθετα, η βιώσιμη ανάπτυξη της ΤΝ απαιτεί την ταυτόχρονη προώθηση της τεχνολογικής προόδου και τη διασφάλιση υψηλού επιπέδου εμπιστοσύνης.

5.3.2 Οι Κατευθυντήριες Γραμμές για Αξιόπιστη και Υπεύθυνη Τεχνητή Νοημοσύνη

Παράλληλα, άλλοι διεθνείς φορείς έχουν εκδώσει κατευθυντήριες γραμμές για αξιόπιστη και υπεύθυνη ΤΝ. Οι Αρχές του ΟΟΣΑ (OECD) για την ΤΝ (2019) προτείνουν, μεταξύ άλλων, τη λογοδοσία των υπευθύνων ανάπτυξης και χρήσης της τεχνολογίας, την προώθηση της διαφάνειας και την προστασία της ιδιωτικότητας. Σε παρόμοιο πνεύμα, η Ευρωπαϊκή Επιτροπή έχει αναπτύξει τις «Κατευθυντήριες Γραμμές για την Έμπιστη Τεχνητή Νοημοσύνη» (Ethics Guidelines for Trustworthy AI), που υπογραμμίζουν αρχές όπως ο σεβασμός της ανθρώπινης αυτονομίας, η πρόληψη βλάβης, η δικαιοσύνη και η δυνατότητα λογοδοσίας (Wirtz et al., 2018). Αυτές οι αρχές εξειδικεύονται σε πρακτικές πολιτικές, όπως οι μηχανισμοί για την παρακολούθηση και αξιολόγηση των αλγορίθμων, καθώς και οι κώδικες δεοντολογίας που καλούνται να υιοθετούν οι εταιρείες τεχνολογίας.

5.3.3 Ο Ρόλος Διεθνών Πλαισίων

Σε παγκόσμιο επίπεδο, ο ρόλος των διεθνών οργανισμών καθίσταται κρίσιμος. Ο Οργανισμός Ηνωμένων Εθνών (ΟΗΕ) φιλοξενεί τακτικές συζητήσεις γύρω από τα αυτόνομα όπλα και τα ανθρώπινα δικαιώματα στον ψηφιακό κόσμο, ενώ η UNESCO υιοθέτησε το 2021 μια Σύσταση για την Ηθική της ΤΝ, η οποία προτρέπει τα κράτη-μέλη να αναπτύξουν ισχυρούς θεσμούς ρύθμισης, να προωθήσουν την τεχνολογική ισότητα και να προστατέψουν τις ευάλωτες ομάδες (UNESCO, 2021). Παρότι οι διεθνείς αυτές πρωτοβουλίες δεν έχουν νομική δεσμευτική νομική ισχύ, αποτελούν ισχυρό πλαίσιο αναφοράς για τη θέσπιση δεσμευτικών κανόνων ή κωδίκων δεοντολογίας σε εθνικό επίπεδο.

Είναι εμφανές ότι οι σχετικές πολιτικές και νομικές εξελίξεις ακολουθούν, αλλά και επιδιώκουν να προλάβουν, τις τεχνολογικές καινοτομίες. Καθώς η ΤΝ εισχωρεί ολοένα και περισσότερο στην καθημερινή ζωή, από την ασφάλεια και την υγεία έως τα αυτόνομα οχήματα και τη βιομηχανική παραγωγή, τα ρυθμιστικά πλαίσια καλούνται να σταθμίσουν προσεκτικά τα οφέλη της καινοτομίας έναντι των κινδύνων που εγκυμονεί. Ο ιδιωτικός τομέας, η ακαδημαϊκή κοινότητα και οι κρατικοί θεσμοί χρειάζεται να συνεργαστούν στενά, ώστε να διαμορφώσουν ολοκληρωμένες πολιτικές που όχι μόνο θα προστατεύουν τα δικαιώματα των πολιτών, αλλά ταυτόχρονα θα επιτρέπουν τη βιώσιμη ανάπτυξη και αξιοποίηση της ΤΝ προς όφελος του κοινωνικού συνόλου. Η σταδιακή εναρμόνιση

των προσεγγίσεων σε διεθνές επίπεδο δείχνει ότι η ηθική της ΤΝ δεν είναι απλώς ζήτημα θεωρητικών αρχών, αλλά αναπόσπαστο κομμάτι της μελλοντικής τεχνολογικής και κοινωνικής προόδου (Wirtz et. al, 2018).

Το AI Act αναδεικνύει μια ευρωπαϊκή προσέγγιση σχεδίασης και ανάπτυξης συστημάτων ΤΝ που βασίζεται στις αρχές της αξιοπιστίας, της ασφάλειας, της διαφάνειας και της ανθρώπινης εποπτείας. Η σημασία του έγκειται όχι μόνο στο ρυθμιστικό του περιεχόμενο αλλά και στο γεγονός ότι αποτελεί μία από τις πρώτες ολοκληρωμένες προσπάθειες παγκόσμιας κλίμακας να μετατραπούν οι ηθικές αρχές της ΤΝ σε πρακτικές κατευθύνσεις ανάπτυξης και χρήσης.

Πίνακας 6: Ρυθμιστικά Πλαίσια, Πολιτικές και Διεθνείς Πρωτοβουλίες για την Υπεύθυνη Ανάπτυξη της Τεχνητής Νοημοσύνης.

Επίπεδο Ρύθμισης	Παραδείγματα πρωτοβουλιών	Κύριοι άξονες δράσης / αρχές
Ευρωπαϊκή Ένωση	Κανονισμός EU AI Act Κατευθυντήριες Γραμμές για Έμπιστη ΤΝ	Πολυεπίπεδη προσέγγιση κινδύνου, απαγορεύσεις, αξιολόγηση κινδύνου, διαφάνεια, προστασία θεμελιωδών δικαιωμάτων
Διεθνείς Οργανισμοί	ΟΟΣΑ – Αρχές για την ΤΝ UNESCO – Σύσταση για την Ηθική της ΤΝ ΟΗΕ – Συζητήσεις για αυτόνομα όπλα και ανθρώπινα δικαιώματα	Λογοδοσία, διαφάνεια, προστασία ιδιωτικότητας, ανθρώπινη αυτονομία, δικαιοσύνη, προστασία ευάλωτων ομάδων, τεχνολογική ισότητα
Εθνικό και Κοινωνικό Επίπεδο	Εθνικές νομοθεσίες και κώδικες δεοντολογίας	Συνεργασία ιδιωτικού τομέα, ακαδημαϊκής κοινότητας και κρατικών θεσμών για διασφάλιση βιώσιμης ανάπτυξης και προστασία δικαιωμάτων

5.4 Ηθικά πρότυπα και τεχνολογίες υπεύθυνης Σχεδίασης και Ανάπτυξης Συστημάτων Τεχνητής Νοημοσύνης

5.4.1 Ηθικές Κατευθυντήριες Αρχές

Οι πρωτοβουλίες της βιομηχανίας και της επιστημονικής κοινότητας για την προώθηση της ηθικής στην Τεχνητή Νοημοσύνη (TN) εστιάζουν, αφενός, στη διαμόρφωση ηθικών κατευθυντήριων αρχών και, αφετέρου, στη δημιουργία συγκεκριμένων τεχνολογικών εργαλείων που διευκολύνουν την υπεύθυνη υλοποίηση αυτών των αρχών. Ένα θεμελιώδες σημείο αφετηρίας είναι οι δηλώσεις αξιών, όπως οι «Google AI Principles», που αναφέρουν ρητά την επιδίωξη διαφάνειας, δικαιοσύνης, λογοδοσίας και σεβασμού της ιδιωτικότητας στην ανάπτυξη και τη διάθεση συστημάτων TN (Floridi & Cowls, 2019). Παρόμοιες αρχές υιοθετούνται από διάφορους οργανισμούς, μεταξύ των οποίων οι επαγγελματικές ενώσεις ACM και IEEE, οι οποίες παρέχουν κώδικες δεοντολογίας και κατευθυντήριες γραμμές για τους επαγγελματίες του κλάδου (Dignum, 2019).

5.4.2 Τεχνολογικά Εργαλεία για Υπεύθυνη Σχεδίαση και Ανάπτυξη Συστημάτων Τεχνητής Νοημοσύνης

Πέρα από τις υψηλού επιπέδου δηλώσεις αρχών, η βιομηχανία και η ακαδημαϊκή κοινότητα αναπτύσσουν τεχνολογικά εργαλεία για υπεύθυνη σχεδίαση και ανάπτυξη συστημάτων TN. Ενδεικτικό είναι το ρεύμα της Επεξηγήσιμης TN (Explainable AI (XAI)), που αποσκοπεί να καταστήσει πιο διαφανείς τους σύνθετους αλγορίθμους βαθιάς μάθησης, δίνοντας τη δυνατότητα στους χρήστες και στους φορείς λήψης αποφάσεων να κατανοούν τα κριτήρια που οδήγησαν ένα σύστημα σε κάποιο συγκεκριμένο αποτέλεσμα (Miller, 2017). Παράλληλα, πολλοί ερευνητές και εταιρείες αναπτύσσουν αλγόριθμους ανίχνευσης και μείωσης μεροληψίας (bias mitigation), ώστε να περιοριστεί ο κίνδυνος οι προϋπάρχουσες κοινωνικές ή θεσμικές ανισότητες να ενσωματώνονται σε αλγοριθμικές αποφάσεις. Μια ακόμα προσέγγιση είναι το «Ethics by Design», όπου η ηθική αξιολόγηση και οι αρχές δικαιοσύνης ενσωματώνονται στο στάδιο του σχεδιασμού του συστήματος, ώστε οι προγραμματιστές να εντοπίζουν έγκαιρα πιθανά προβλήματα και να αναπτύσσουν λύσεις ήδη από τα πρώτα βήματα (Hagendorff, 2020).

5.4.3 Αυτορρύθμιση και Συνεργατικές Προσπάθειες

Ιδιαίτερα σημαντική είναι και η διάσταση της αυτορρύθμισης και των συνεργατικών πρωτοβουλιών. Επαγγελματικές ενώσεις όπως η IEEE διαθέτουν Ομάδες Εργασίας για την Ηθική της ΤΝ, οι οποίες επιδιώκουν τη διαμόρφωση κοινών προτύπων και βέλτιστων πρακτικών. Στο ίδιο πλαίσιο, ο κλάδος της τεχνολογίας αναγνωρίζει όλο και περισσότερο ότι οι δεσμευτικές νομικές ρυθμίσεις δεν αρκούν από μόνες τους για να αντιμετωπίσουν όλες τις προκλήσεις· η εμπλοκή της κοινότητας των προγραμματιστών σε διαδικασίες αναστοχασμού και συνεχούς βελτίωσης είναι εξίσου καθοριστική (El-Deeb, 2024). Μέσα από συνέδρια, εργαστήρια και συλλογικά έργα ανοικτού κώδικα (open-source projects), προάγεται η ανταλλαγή γνώσεων και εμπειριών, ενισχύεται η διαφάνεια και δημιουργείται μια κουλτούρα συνεργασίας.

Εν τέλει, το μήνυμα που προκύπτει είναι ότι η προώθηση της ηθικής στην ΤΝ δεν μπορεί να στηρίζεται αποκλειστικά στη νομοθεσία. Απαιτείται μια πολυεπίπεδη προσέγγιση: αφενός, η υιοθέτηση ευρέως αναγνωρισμένων ηθικών αρχών από εταιρείες και ερευνητές· αφετέρου, η παροχή τεχνολογικών εργαλείων που θα διευκολύνουν την υπεύθυνη εφαρμογή και αξιολόγηση των συστημάτων σε όλο τον κύκλο ζωής τους. Ταυτόχρονα, η αυτορρύθμιση και η συστηματική συνεργασία μεταξύ βιομηχανίας και επιστήμης εμπλουτίζουν τη συζήτηση, επιτρέποντας τη γρήγορη προσαρμογή των ηθικών προτύπων και κριτηρίων στις εκάστοτε τεχνολογικές εξελίξεις. Αυτό το δυναμικό περιβάλλον δείχνει ότι, για να επιτευχθεί μια πραγματικά υπεύθυνη ΤΝ, απαιτείται συνεχής διάλογος και ενεργή δέσμευση όλων των ενδιαφερόμενων μερών (El-Deeb, 2024).

Μία από τις σημαντικότερες εξελίξεις στον χώρο της τεχνητής νοημοσύνης είναι η μετάβαση από γενικές ηθικές αρχές σε ολοκληρωμένα συστήματα διακυβέρνησης της ΤΝ. Η σύγχρονη προσέγγιση δεν περιορίζεται στη διατύπωση αρχών, αλλά απαιτεί την ενσωμάτωση διαδικασιών διαχείρισης κινδύνων, αξιολόγησης επιπτώσεων, παρακολούθησης της απόδοσης και συνεχούς βελτίωσης. Η ανάγκη αυτή προκύπτει από τα ιδιαίτερα χαρακτηριστικά των συστημάτων ΤΝ, όπως η δυνατότητα συνεχούς μάθησης, η περιορισμένη διαφάνεια ορισμένων μοντέλων και οι πιθανές κοινωνικές συνέπειες των αποφάσεών τους. Ως αποτέλεσμα, η υπεύθυνη ανάπτυξη και χρήση της τεχνητής νοημοσύνης αντιμετωπίζεται πλέον ως οργανωτική διαδικασία που πρέπει να ενσωματώνεται στη συνολική στρατηγική και λειτουργία ενός οργανισμού.

6

Μελέτη Περίπτωσης και

Συγκριτική Ανάλυση Κύκλων

Ζωής Λογισμικού, Λειτουργικών

Συστημάτων και Συστημάτων

Τεχνητής Νοημοσύνης

6.1 Σκοπός, Ερευνητική Στόχευση και Ορισμός της Μελέτης Περίπτωσης

Η παρούσα ενότητα λειτουργεί ως «γέφυρα» ανάμεσα στο θεωρητικό μέρος της εργασίας και στο πρακτικό, συγκριτικό κεφάλαιο της διπλωματικής εργασίας. Σκοπός της είναι να ορίσει καθαρά τι ακριβώς εξετάζουμε, με ποια μεθοδολογική προσέγγιση εξετάζεται και πώς δομούμε τη μελέτη περίπτωσης (case study) έτσι ώστε να έχει ακαδημαϊκή συνέπεια, πρακτική αξία και να απαντά στη βασική απαίτηση της εργασίας: να παρουσιαστεί, με τρόπο συγκεκριμένο και τεκμηριωμένο, η σύγκριση της ανάπτυξης κλασικού λογισμικού με την ανάπτυξη συστημάτων τεχνητής νοημοσύνης, και επιπλέον να συνδεθούν οι διαφορές αυτές με τον κύκλο ζωής ενός λειτουργικού συστήματος και τον κύκλο ζωής της τεχνητής νοημοσύνης.

Η επιλογή «μελέτης περίπτωσης» ως μορφής ανάπτυξης δεν είναι απλώς αφηγηματική. Μας επιτρέπει να αποτυπώσουμε την ανάπτυξη συστημάτων ως πραγματική διαδικασία λήψης αποφάσεων, με ρόλους, παραδοτέα, ελέγχους και σημεία όπου παράγονται ρίσκα ή/και εγγυήσεις. Αυτό είναι ιδιαίτερα κρίσιμο όταν μιλάμε για τεχνητή νοημοσύνη, γιατί η διεθνής συζήτηση μετατοπίζεται από γενικές «αρχές ηθικής» σε πιο συγκεκριμένες απαιτήσεις: διακυβέρνηση (governance), διαχείριση κινδύνου (risk management), διαφάνεια (transparency) και λογοδοσία (accountability). Τέτοιες προσεγγίσεις εμφανίζονται πλέον ως κεντρικές τόσο σε οργανωτικά

πλαίσια διακυβέρνησης της TN όσο και σε επίσημα διεθνή/θεσμικά εργαλεία αξιολόγησης και κατευθυντήριες γραμμές (Mäntymäki et al., 2022; National Institute of Standards and Technology [NIST], 2023; UNESCO, 2022, 2023; United Nations, 2024).

Σκοπός της ενότητας του κεφαλαίου

Ο άμεσος σκοπός του κεφαλαίου είναι διπλός. Πρώτον, να παρουσιάσει πώς “χτίζεται” ένα σύστημα όταν αυτό είναι κλασικό λογισμικό από τη σύλληψη απαιτήσεων μέχρι τη συντήρηση, αναδεικνύοντας με σαφήνεια τον διακριτό ρόλο του αναλυτή και του προγραμματιστή/μηχανικού λογισμικού (developer). Δεύτερον, να δείξει πώς η σχεδίαση και ανάπτυξη συστημάτων τεχνητής νοημοσύνης διαφοροποιείται, όχι μόνο τεχνικά αλλά και μεθοδολογικά και ηθικό οργανωτικά, καθώς εισάγει την εξάρτηση από δεδομένα, την πιθανοκρατική απόδοση, την πιθανότητα υποβάθμισης στον χρόνο και, κυρίως, νέες απαιτήσεις αξιολόγησης ασφάλειας και κοινωνικού αντίκτυπου (Weidinger et al., 2023; NIST, 2023).

Σε αυτό το πλαίσιο, η «σύγκριση» δεν αντιμετωπίζεται ως απλή αντιπαράθεση δύο τεχνολογιών. Αντιμετωπίζεται ως σύγκριση δύο διαφορετικών λογικών παραγωγής αξιοπιστίας. Στο κλασικό λογισμικό, η αξιοπιστία προκύπτει κυρίως από προδιαγραφές, έλεγχο υλοποίησης και δοκιμές. Στην τεχνητή νοημοσύνη, η αξιοπιστία προκύπτει επίσης από αυτά, αλλά κυρίως από το τρίπτυχο δεδομένα–μοντέλο–συνεχής παρακολούθηση, το οποίο χρειάζεται διαρκή επιτήρηση, διακυβέρνηση και διαχείριση κινδύνων σε όλο τον κύκλο ζωής (NIST, 2023; UNESCO, 2023). Αυτή η μετατόπιση εξηγεί γιατί η συζήτηση για «υπεύθυνη TN» οργανώνεται σήμερα γύρω από μηχανισμούς πολιτικής και διακυβέρνησης και όχι μόνο γύρω από γενικές ηθικές διακηρύξεις (UNESCO, 2022; United Nations, 2024).

Ερευνητική στόχευση

Η ερευνητική στόχευση της ενότητας 6.1 είναι να ορίσει τη μελέτη περίπτωσης ως ένα εργαλείο κατανόησης ρόλων, ευθυνών και παραδοτέων μέσα σε δύο διαφορετικά περιβάλλοντα ανάπτυξης:

1. Ανάπτυξη λογισμικού: πώς μετασχηματίζονται οι ανάγκες ενός οργανισμού/χρήστη σε απαιτήσεις και στη συνέχεια σε λειτουργικό σύστημα, με έμφαση στη σαφή διάκριση των παραδοτέων του αναλυτή (π.χ. απαιτήσεις, κριτήρια αποδοχής, επιχειρησιακή λογική) και του developer (π.χ. αρχιτεκτονική, κώδικας, δοκιμές, εγκατάσταση).
2. Σχεδίαση και ανάπτυξη συστημάτων τεχνητής νοημοσύνης: πώς μετασχηματίζεται ένα πρόβλημα σε πρόβλημα μάθησης, πώς καθορίζονται μετρικές επιτυχίας, πώς σχεδιάζεται η αξιολόγηση κινδύνου, και πώς ο αναλυτής και ο developer/ML engineer μοιράζονται ρόλους σε μια διαδικασία όπου η «συμπεριφορά» του συστήματος δεν είναι πλήρως

ντετερμινιστική αλλά προκύπτει από δεδομένα και εκπαίδευση (NIST, 2023; Weidinger et. al, 2023).

Αυτή η στόχευση συνδέεται άμεσα με ένα ευρύτερο συμπέρασμα της πρόσφατης βιβλιογραφίας: η TN δεν αποτελεί απλώς νέα «λειτουργικότητα» αλλά τεχνολογία με ευρείες κοινωνικές επιπτώσεις, που απαιτεί οργανωτική ετοιμότητα, θεσμική κατανόηση και εργαλεία πολιτικής, ειδικά όταν εφαρμόζεται σε ευαίσθητους τομείς (π.χ. δημόσια διοίκηση, επιτήρηση, υγεία, δικαιοσύνη) (Fosso Wamba et al., 2021; United Nations, 2024). Ταυτόχρονα, η εκρηκτική ανάπτυξη της γενετικής TN (Generative AI) και η ένταση της ερευνητικής δραστηριότητας καθιστούν ακόμη πιο αναγκαία μια «συστηματική» γλώσσα αξιολόγησης κινδύνων και επιπτώσεων, ώστε να μην μείνουμε σε περιγραφές υψηλού επιπέδου (Dwivedi & Elluri, 2024; Masle, 2024).

Ορισμός της «μελέτης περίπτωσης» και τι θεωρούμε “case”

Στο παρόν κεφάλαιο, ως «μελέτη περίπτωσης (case study)» ορίζεται μια συγκροτημένη, ρεαλιστική αναπαράσταση ενός έργου ανάπτυξης που περιγράφεται με τρόπο ώστε να είναι σαφή:

- ποιο είναι το ζητούμενο (στόχος συστήματος),
- ποιοι είναι οι ρόλοι και οι ευθύνες (αναλυτής, developer, stakeholders),
- ποια είναι τα παραδοτέα ανά φάση,
- ποια είναι τα σημεία ελέγχου (quality gates),
- ποιοι κίνδυνοι ανακύπτουν και πώς αντιμετωπίζονται.

Το “case” δεν είναι απαραίτητα ένα πραγματικό έργο με ευαίσθητα δεδομένα ή εταιρικές λεπτομέρειες. Είναι ένα διδακτικό αλλά ρεαλιστικό σενάριο που επιτρέπει να φανούν οι κρίσιμες διαφορές ανάμεσα σε κλασικό λογισμικό και TN, χωρίς να χαθεί η τεχνική ακρίβεια. Η επιλογή αυτού του «ενδιάμεσου» τρόπου (ούτε αφηρημένη θεωρία, ούτε κλειστή εταιρική μελέτη) ευθυγραμμίζεται με τον τρόπο που πολλά σύγχρονα πλαίσια αντιμετωπίζουν την υπεύθυνη TN: ως ζήτημα διαδικασιών, ελέγχων και τεκμηρίωσης που πρέπει να είναι επαναλήψιμα και ελέγξιμα (auditability), όχι ως μεμονωμένη “ηθική πρόθεση” (NIST, 2023; UNESCO, 2023; Mäntymäki et. al, 2022).

Με αυτή τη βάση, η ενότητα 6.1 «κλειδώνει» το τι θα ακολουθήσει: στα επόμενα υποκεφάλαια, θα παρουσιάσουμε δύο παράλληλες αναπτύξεις (λογισμικό και TN) ως cases, θα τα συγκρίνουμε στο επίπεδο ρόλων/παραδοτέων/ρίσκων, και στη συνέχεια θα τα συνδέσουμε με τους κύκλους ζωής λειτουργικού συστήματος και TN, ώστε να καταλήξουμε σε μια καθαρή εικόνα του, τι σημαίνουν στην πράξη η διαφάνεια, η μεροληψία και οι ηθικές απαιτήσεις σε «συστήματα που παράγουν αποφάσεις» (UNESCO, 2022; United Nations, 2024; Weidinger et. al, 2023).

6.2 Περίπτωση Α: Κύκλος Ζωής, Ρόλοι, Διαδικασία, Παραδοτέα

Για να είναι η σύγκριση με την ανάπτυξη συστημάτων τεχνητής νοημοσύνης πραγματικά «χειροπιαστή», επιλέγεται ως μελέτη περίπτωσης ένα έργο κλασικής ανάπτυξης λογισμικού που είναι ρεαλιστικό, συνηθισμένο σε δημόσιους και ιδιωτικούς οργανισμούς και έχει σαφείς απαιτήσεις συμμόρφωσης: ένα “Σύστημα Ηλεκτρονικής Εξυπηρέτησης Πολιτών”, δηλαδή μια διαδικτυακή πλατφόρμα όπου ο πολίτης υποβάλλει αιτήματα (π.χ. βεβαιώσεις, αιτήματα τεχνικών υπηρεσιών, ραντεβού), επισυνάπτει δικαιολογητικά, παρακολουθεί την πορεία διεκπεραίωσης και λαμβάνει απαντήσεις/έγγραφα ηλεκτρονικά. Το έργο αυτό είναι κατάλληλο ως case γιατί, παρότι δεν είναι «τεχνητή νοημοσύνη», αγγίζει θέματα που αργότερα γίνονται ακόμη πιο απαιτητικά στην ΤΝ: επεξεργασία προσωπικών δεδομένων, ιχνηλασιμότητα ενεργειών, λογοδοσία, διαφάνεια διαδικασιών και τεκμηρίωση αποφάσεων/ενεργειών του συστήματος (UNESCO, 2022; United Nations, 2024; Lu, 2022).

Στην κλασική ανάπτυξη λογισμικού, η βασική ιδέα είναι ότι το σύστημα πρέπει να συμπεριφέρεται ντετερμινιστικά όταν δώσουμε τα ίδια δεδομένα/εντολές, να παράγει σταθερά το ίδιο αποτέλεσμα. Αυτό δεν σημαίνει ότι είναι «εύκολο», αλλά σημαίνει ότι η ποιότητα και η αξιοπιστία του λογισμικού χτίζονται κυρίως μέσα από σωστά καθορισμένες απαιτήσεις, σωστή σχεδίαση, ελεγχόμενη υλοποίηση και δοκιμές. Η τεκμηρίωση εδώ δεν είναι τυπικότητα είναι ο μηχανισμός με τον οποίο εξασφαλίζεται ότι το σύστημα κάνει αυτό που συμφωνήθηκε και ότι μπορεί να ελεγχθεί σε περίπτωση σφάλματος, διαρροής ή αμφισβήτησης (UNESCO, 2023; United Nations, 2024).

6.2.1 Έναρξη Έργου και Αποσαφήνιση Αναγκών: Ποιο Πρόβλημα Επιλύεται

Η εκκίνηση του έργου γίνεται σχεδόν πάντα με ένα φαινομενικά απλό αίτημα: «θέλουμε ένα σύστημα ηλεκτρονικών αιτημάτων». Στην πράξη, όμως, αυτό είναι μόνο ο τίτλος. Το κρίσιμο σημείο είναι να μεταφραστεί η γενική πρόθεση σε συγκεκριμένη λειτουργικότητα: ποιος θεωρείται χρήστης (πολίτης, υπάλληλος, προϊστάμενος), τι δικαιώματα έχει ο καθένας, ποια αιτήματα υποστηρίζονται, ποια δικαιολογητικά απαιτούνται, ποια βήματα έγκρισης ισχύουν, πότε αποστέλλονται ειδοποιήσεις, πώς γίνεται η ταυτοποίηση και τι θεωρείται επιτυχής διεκπεραίωση.

Εδώ, ο ρόλος του αναλυτή (αναλυτής επιχειρησιακών απαιτήσεων/αναλυτής συστημάτων) είναι καθοριστικός: δεν «γράφει κώδικα», αλλά οργανώνει το πρόβλημα. Χαρτογραφεί τη ροή εργασίας (workflow), εντοπίζει εμπλεκόμενους (stakeholders), ορίζει κανόνες, αποσαφηνίζει εξαιρέσεις και μειώνει την ασάφεια. Σε έργα που αγγίζουν προσωπικά δεδομένα, ο αναλυτής οφείλει ήδη από την αρχή να εντάξει την προστασία δικαιωμάτων και τη λογοδοσία ως απαιτήσεις, όχι ως «κάτι που θα δούμε μετά». Αυτό ευθυγραμμίζεται με τη σύγχρονη θεώρηση ότι η τεχνολογία πρέπει να

ενσωματώνει κοινωνικές αξίες και προστασίες από τον σχεδιασμό της, ειδικά όταν επεξεργάζεται δεδομένα πολιτών (UNESCO, 2022; United Nations, 2024; Lu, 2022).

6.2.2 Ρόλοι: Τι κάνει ο αναλυτής και τι κάνει ο developer στην πράξη

Στη μελέτη περίπτωσης, ο αναλυτής λειτουργεί ως «μεταφραστής» μεταξύ πραγματικότητας και συστήματος. Παίρνει τις ανάγκες του οργανισμού και τις κάνει απαιτήσεις που μπορούν να υλοποιηθούν και να ελεγχθούν. Το πρακτικό του αποτύπωμα είναι ότι παράγει συγκεκριμένα παραδοτέα: περιγραφές λειτουργιών, κανόνες, σενάρια χρήσης, κριτήρια αποδοχής, καθώς και έναν μηχανισμό ιχνηλασιμότητας που δείχνει ποια απαίτηση υλοποιείται πού και πώς ελέγχεται. Σε ευαίσθητα συστήματα δημόσιας εξυπηρέτησης, ο αναλυτής πρέπει να ορίζει και απαιτήσεις διαφάνειας: ποια γεγονότα καταγράφονται, ποιος βλέπει τι, ποιο είναι το «ίχνος» μιας ενέργειας, ποια είναι η ελάχιστη αναγκαία πρόσβαση και πώς τεκμηριώνεται η ορθότητα της διαδικασίας (UNESCO, 2023; United Nations, 2024).

Ο developer (προγραμματιστής/μηχανικός λογισμικού) παίρνει αυτές τις προδιαγραφές και τις μετατρέπει σε αρχιτεκτονική, κώδικα και λειτουργικό σύστημα. Εδώ το βασικό ζητούμενο είναι η τεχνική υλοποίηση να είναι συνεπής, ασφαλής και συντηρήσιμη: να υπάρχουν καθαρά διαχωρισμένα υποσυστήματα (π.χ. διαχείριση χρηστών, διαχείριση αιτημάτων, αρχειοθέτηση, ειδοποιήσεις), έλεγχοι πρόσβασης, μηχανισμοί καταγραφής (logging), διαδικασίες εγκατάστασης και ενημερώσεων, και πολύ πρακτικά να μπορεί μια ομάδα να υποστηρίξει το σύστημα στον χρόνο χωρίς κάθε αλλαγή να «σπάει» τα πάντα. Σε αυτό το σημείο η έννοια της λογοδοσίας αποκτά μηχανική διάσταση: δεν είναι μόνο «ποιος φταίει», αλλά «πώς αποδεικνύεται τι έγινε» μέσω καταγραφών, εκδόσεων, αλλαγών και τεστ (UNESCO, 2023; Lu, 2022).

6.2.3 Διαδικασία Ανάπτυξης: Από Απαιτήσεις σε Παραγωγή

Μετά την αποσαφήνιση των απαιτήσεων, το έργο περνά σε φάσεις που, ανεξάρτητα από το αν ο οργανισμός χρησιμοποιεί πιο κλασικό κύκλο ανάπτυξης ή πιο ευέλικτη μέθοδο (agile), περιλαμβάνουν σταθερά βήματα.

Πρώτα έρχεται η ανάλυση απαιτήσεων: ο αναλυτής καταγράφει λειτουργικές απαιτήσεις (τι κάνει το σύστημα) και μη λειτουργικές (ασφάλεια, απόδοση, διαθεσιμότητα, καταγραφή ενεργειών, συμμόρφωση). Στη συνέχεια περνάμε στη σχεδίαση (design), όπου ο developer σε συνεργασία με τον αναλυτή ορίζει την τεχνική δομή: βάση δεδομένων, ρόλους χρηστών, διεπαφές (UI), υπηρεσίες (APIs), και ροές. Έπειτα ακολουθεί η υλοποίηση (implementation), όπου το σύστημα «γράφεται» και ενσωματώνεται σε ενιαίο σύνολο. Μετά έρχονται οι δοκιμές (testing): δοκιμές λειτουργικότητας, δοκιμές ασφάλειας, δοκιμές φόρτου όπου χρειάζεται. Τέλος, το σύστημα περνά

σε παραγωγική λειτουργία (deployment) και μπαίνει σε περίοδο συντήρησης (maintenance), όπου διορθώνονται σφάλματα, γίνονται ενημερώσεις και προστίθενται αλλαγές.

Σε συστήματα που διαχειρίζονται προσωπικά δεδομένα, η διαδικασία αυτή οφείλει να έχει ενσωματωμένους ελέγχους για ιδιωτικότητα και λογοδοσία, γιατί η «αδιαφάνεια» ενός συστήματος δεν είναι απλώς κακό UX είναι πιθανό νομικό/ηθικό ρίσκο όταν αφορά δικαιώματα πολιτών (Lu, 2022; UNESCO, 2022). Αυτή η λογική, αν και αναδεικνύεται εντονότερα στην TN, είναι πλήρως εφαρμόσιμη και στο κλασικό λογισμικό: ο σχεδιασμός πρέπει να παράγει τεκμήρια, όχι μόνο λειτουργίες (UNESCO, 2023; United Nations, 2024).

6.2.4 Παραδοτέα: Τι «Παράγεται» σε Κάθε Στάδιο

Σε ένα σοβαρό έργο ανάπτυξης λογισμικού, δεν «παράγεται» μόνο κώδικας. Παράγεται ένα σύνολο παραδοτέων που, μαζί, αποτελούν τον φάκελο τεκμηρίωσης και λειτουργίας του συστήματος.

Στο στάδιο ανάλυσης, ο αναλυτής παραδίδει τεκμηριωμένες απαιτήσεις, ροές εργασίας, ορισμούς ρόλων και δικαιωμάτων, περιγραφές λειτουργιών και κριτήρια αποδοχής. Στο στάδιο σχεδίασης, ο developer παραδίδει τεχνική αρχιτεκτονική, σχεδίαση δεδομένων, προδιαγραφές διεπαφών, καθώς και μηχανισμούς καταγραφής και ελέγχου. Στην υλοποίηση, παραδίδεται ο κώδικας σε αποθετήριο (repository), μαζί με τεχνική τεκμηρίωση, οδηγίες εγκατάστασης και ρυθμίσεις. Στις δοκιμές, παραδίδονται σχέδια δοκιμών (test plans), αποτελέσματα και διορθώσεις. Στην παραγωγή, παραδίδονται διαδικασίες λειτουργίας, παρακολούθησης και αντιμετώπισης συμβάντων (incident handling), ώστε να υπάρχει πρακτική λογοδοσία: το σύστημα να μπορεί να ελεγχθεί, να αποκατασταθεί και να βελτιωθεί όταν κάτι πάει στραβά (UNESCO, 2023; United Nations, 2024).

Ειδικά για την παρούσα εργασία, αξίζει να τονιστεί ότι σε ένα σύστημα δημόσιας εξυπηρέτησης τα παραδοτέα που σχετίζονται με διαφάνεια δεν είναι «έξτρα»: είναι λειτουργική απαίτηση. Ένα καλό σύστημα πρέπει να μπορεί να εξηγήσει ποιο βήμα έγινε, από ποιον, πότε, με ποιο αποτέλεσμα, και με ποια δικαιοδοσία όχι για να «εξηγήσει ένα μοντέλο», αλλά για να αποδείξει τη νομιμότητα και την ορθότητα της διαδικασίας (Lu, 2022; UNESCO, 2022). Αυτή η λογική προετοιμάζει άμεσα το έδαφος για το αντίστοιχο κομμάτι στην TN, όπου η απαίτηση διαφάνειας γίνεται ακόμη πιο δύσκολη λόγω της πολυπλοκότητας και της πιθανοκρατικής φύσης των μοντέλων (Weidinger et al., 2023).

6.2.5 Ένα μικρό παράδειγμα μέσα στο case: “Αίτημα με έγκριση”

Ας θεωρήσουμε ένα πολύ συγκεκριμένο υποσενάριο: ο πολίτης υποβάλλει αίτημα «χορήγησης βεβαίωσης», επισυνάπτει δικαιολογητικά, και το αίτημα πρέπει να περάσει από έλεγχο πληρότητας και έγκριση από προϊστάμενο. Ο αναλυτής ορίζει τι σημαίνει «πλήρης αίτηση», ποια

δικαιολογητικά είναι υποχρεωτικά, ποιος έχει δικαίωμα να ζητήσει συμπληρωματικά στοιχεία, ποια είναι η προθεσμία απάντησης και πότε το αίτημα χαρακτηρίζεται «ολοκληρωμένο». Ο developer υλοποιεί τα βήματα αυτά ως ροή στο σύστημα, σχεδιάζει την αποθήκευση, δημιουργεί τις οθόνες, τις ειδοποιήσεις και κρίσιμα τη σωστή καταγραφή όλων των ενεργειών. Έτσι, αν υπάρξει αμφισβήτηση ή παράπονο, δεν θα χρειάζεται «να θυμηθούμε τι έγινε», αλλά θα υπάρχει ίχνος ενεργειών, που αποτελεί πρακτικό μηχανισμό λογοδοσίας και διαφάνειας (UNESCO, 2023; United Nations, 2024).

Η Περίπτωση Α δείχνει το βασικό πρότυπο της κλασικής ανάπτυξης λογισμικού: η ποιότητα παράγεται μέσω σαφών απαιτήσεων, ελέγξιμων παραδοτέων και δοκιμών. Ταυτόχρονα, δείχνει ότι έννοιες όπως διαφάνεια, λογοδοσία και ηθική λειτουργία δεν είναι αποκλειστικές της τεχνητής νοημοσύνης. Είναι ήδη κρίσιμες στο κλασικό λογισμικό όταν αυτό επηρεάζει δικαιώματα, ιδιωτικότητα και πρόσβαση σε υπηρεσίες, επειδή η «αδιαφάνεια» ή η κακή διακυβέρνηση δεδομένων μπορεί να έχει πραγματικές κοινωνικές επιπτώσεις (UNESCO, 2022; United Nations, 2024; Lu, 2022). Με αυτή τη βάση, η μετάβαση στην Περίπτωση Β (ανάπτυξη τεχνητής νοημοσύνης) θα καταδείξει όχι ότι οι αρχές αλλάζουν, αλλά ότι η πολυπλοκότητα, η αβεβαιότητα και οι νέοι κίνδυνοι αυξάνονται, απαιτώντας ακόμη πιο δομημένους μηχανισμούς αξιολόγησης και ασφάλειας (Weidinger et. al, 2023).

6.3 Περίπτωση Β: Κύκλος ζωής συστήματος Τεχνητής

Νοημοσύνης ρόλοι, διαδικασίες και παραδοτέα

Στην Περίπτωση Β χρησιμοποιείται ακριβώς το ίδιο επιχειρησιακό πλαίσιο με την Περίπτωση Α το Σύστημα Ηλεκτρονικής Εξυπηρέτησης Πολιτών με μία κρίσιμη διαφοροποίηση: ενσωματώνεται ένα υποσύστημα τεχνητής νοημοσύνης που αναλαμβάνει αυτόματη κατηγοριοποίηση και προτεραιοποίηση αιτημάτων (π.χ. διάκριση επείγοντων από μη επείγοντα, αυτόματη δρομολόγηση σε αρμόδια υπηρεσία). Η επιλογή αυτή δεν αλλάζει τον σκοπό του συστήματος αλλάζει τον τρόπο με τον οποίο λαμβάνονται αποφάσεις εντός του συστήματος. Έτσι, η σύγκριση γίνεται «μήλο με μήλο»: ίδιος οργανισμός, ίδιοι χρήστες, ίδιες νομικές/ηθικές υποχρεώσεις, αλλά διαφορετικός μηχανισμός απόφασης.

Η προσθήκη τεχνητής νοημοσύνης μετατοπίζει το κέντρο βάρους από τη σταθερή, ντετερμινιστική λογική κανόνων (rules based logic) σε μια πιθανοκρατική λογική μάθησης από δεδομένα. Αυτή η μετατόπιση είναι ο λόγος που διεθνώς η ανάπτυξη ΤΝ συνοδεύεται από νέα πλαίσια ασφάλειας, διακυβέρνησης και αξιολόγησης κοινωνικών επιπτώσεων, τα οποία δεν ήταν απαραίτητα στον ίδιο

βαθμό στο κλασικό λογισμικό (National Institute of Standards and Technology [NIST], 2023; Weidinger et. al, 2023; UNESCO, 2022).

6.3.1 Τι Αλλάζει όταν Εισάγεται Τεχνητή Νοημοσύνη στο Ίδιο Πληροφοριακό Σύστημα

Στο κλασικό σύστημα της Περίπτωσης Α, η κατηγοριοποίηση αιτημάτων γίνεται με ρητούς κανόνες: συγκεκριμένες επιλογές φόρμας, προκαθορισμένες κατηγορίες, χειροκίνητος έλεγχος. Στην παρούσα περίπτωση, η κατηγοριοποίηση γίνεται από ένα μοντέλο μηχανικής μάθησης που «μαθαίνει» από ιστορικά δεδομένα αιτημάτων και αποφάσεων υπαλλήλων. Το σύστημα πλέον δεν εγγυάται απόλυτη ορθότητα σε κάθε περίπτωση, αλλά αποδίδει με κάποιο ποσοστό επιτυχίας, το οποίο πρέπει να οριστεί, να μετρηθεί και να παρακολουθείται.

Αυτή η αλλαγή έχει δύο άμεσες συνέπειες. Πρώτον, η έννοια του «σφάλματος» αλλάζει: δεν είναι πλέον ένα καθαρό bug που διορθώνεται με αλλαγή κώδικα, αλλά μια λανθασμένη πρόβλεψη που μπορεί να προκύψει ακόμα και αν το σύστημα λειτουργεί «όπως σχεδιάστηκε». Δεύτερον, η ευθύνη δεν εξαντλείται στην ορθότητα της υλοποίησης, αλλά επεκτείνεται στην καταλληλότητα των δεδομένων, στην επιλογή μετρικών και στη συνεχή επιτήρηση της συμπεριφοράς του μοντέλου (NIST, 2023; Weidinger et. al, 2023).

6.3.2 Ο Ρόλος του Αναλυτή στη Σχεδίαση Συστημάτων Τεχνητής Νοημοσύνης

Στην σχεδίαση συστημάτων TN, ο ρόλος του αναλυτή διευρύνεται ουσιαστικά. Δεν αρκεί να ορίσει τι «πρέπει να κάνει» το σύστημα· πρέπει να ορίσει τι σημαίνει “καλή απόδοση” και ποιοι κίνδυνοι είναι αποδεκτοί. Στο συγκεκριμένο case, ο αναλυτής καλείται να απαντήσει σε ερωτήματα όπως: με ποια κριτήρια ένα αίτημα θεωρείται «επείγον»; ποια λάθη είναι πιο σοβαρά (π.χ. να χαρακτηριστεί επείγον ένα μη επείγον ή το αντίστροφο); ποια ποσοστά σφάλματος είναι ανεκτά; και ποιος πρέπει να ελέγχει ή να διορθώνει το αποτέλεσμα της TN.

Ο αναλυτής, επιπλέον, ορίζει μετρικές αξιολόγησης (π.χ. ακρίβεια, ανάκληση, ποσοστό ψευδώς αρνητικών), αλλά και μη τεχνικές απαιτήσεις, όπως η δυνατότητα ανθρώπινης παρέμβασης, η καταγραφή των αποφάσεων του μοντέλου και η τεκμηρίωση των δεδομένων εκπαίδευσης. Στο πλαίσιο διεθνών κατευθυντήριων γραμμών, αυτός ο ρόλος θεωρείται κρίσιμος για τη μετάβαση από αφηρημένες ηθικές αρχές σε εφαρμόσιμη διακυβέρνηση της TN (UNESCO, 2022; UNESCO, 2023; United Nations, 2024).

6.3.3 Ο Ρόλος του developer/ML Engineer στην Υλοποίηση Συστημάτων TN

Ο developer, ο οποίος εδώ λειτουργεί συχνά ως μηχανικός μηχανικής μάθησης (ML engineer), δεν υλοποιεί απλώς λειτουργικότητα αλλά ολόκληρη τη ροή μάθησης. Στο case, αυτό σημαίνει:

συλλογή και καθαρισμό ιστορικών δεδομένων αιτημάτων, επιλογή χαρακτηριστικών (features), εκπαίδευση μοντέλου, αξιολόγηση, και ενσωμάτωση του μοντέλου στο παραγωγικό σύστημα.

Κρίσιμο στοιχείο είναι ότι ο developer δεν «κλειδώνει» τη συμπεριφορά του συστήματος μία φορά. Αντίθετα, σχεδιάζει μηχανισμούς συνεχούς παρακολούθησης (monitoring), ώστε να εντοπίζονται φαινόμενα υποβάθμισης (model drift), αλλαγές στα δεδομένα ή αυξανόμενη μεροληψία. Επιπλέον, υλοποιεί μηχανισμούς καταγραφής που επιτρέπουν να αναλυθεί εκ των υστέρων γιατί ένα αίτημα ταξινομήθηκε με συγκεκριμένο τρόπο. Αυτή η τεχνική υποδομή αποτελεί προϋπόθεση για λογοδοσία και αξιολόγηση κινδύνου, όπως αναγνωρίζεται σε σύγχρονα πλαίσια ασφάλειας TN (NIST, 2023; Weidinger et. al, 2023).

6.3.4 Διαδικασία Σχεδίασης και Ανάπτυξης Συστημάτων TN: Από τον Ορισμό του Προβλήματος έως την Τελική Αξιολόγηση του Μοντέλου

Η διαδικασία σχεδίασης και ανάπτυξης συστημάτων TN ξεκινά με διατύπωση προβλήματος (problem framing): τι προβλέπουμε και γιατί. Ακολουθεί η διαχείριση δεδομένων, που περιλαμβάνει επιλογή, καθαρισμό και τεκμηρίωση. Στη συνέχεια έρχεται η εκπαίδευση μοντέλου, η αξιολόγηση με βάση τις μετρικές που έχουν οριστεί, και η ενσωμάτωση στο σύστημα. Σε αντίθεση με το κλασικό λογισμικό, η διαδικασία αυτή δεν ολοκληρώνεται με την εγκατάσταση· συνεχίζεται με επιτήρηση, επανεκπαίδευση και αναθεώρηση.

Στο συγκεκριμένο case, αυτό σημαίνει ότι το σύστημα πρέπει να ελέγχεται περιοδικά: αν αλλάξει ο τύπος αιτημάτων που υποβάλλονται, αν εμφανιστούν νέες κατηγορίες, ή αν παρατηρηθεί συστηματική υποεκπροσώπηση συγκεκριμένων ομάδων. Αυτή η συνεχής διαδικασία επιβεβαιώνει τη θέση ότι η TN δεν είναι «προϊόν» αλλά διαρκής υπηρεσία υπό αξιολόγηση, κάτι που τονίζεται σε σύγχρονες αναλύσεις ασφάλειας και κοινωνικού αντίκτυπου της γενετικής και εφαρμοσμένης TN (Weidinger et. al, 2023; Masle, 2024).

6.3.5 Παραδοτέα κατά την Υλοποίηση Συστημάτων Τεχνητής Νοημοσύνης

Σε επίπεδο παραδοτέων, κατά την υλοποίηση συστημάτων TN παράγει διαφορετικό σύνολο τεκμηρίων σε σχέση με το κλασικό λογισμικό. Εκτός από κώδικα, παραδίδονται τεκμηριώσεις δεδομένων, περιγραφές μοντέλου, αποτελέσματα αξιολόγησης, καταγραφές αποφάσεων και διαδικασίες παρακολούθησης. Αυτά τα παραδοτέα δεν έχουν μόνο τεχνική αξία· αποτελούν εργαλεία ηθικής τεκμηρίωσης, καθώς επιτρέπουν σε έναν οργανισμό να αποδείξει ότι έχει λάβει μέτρα για διαφάνεια, μείωση μεροληψίας και διαχείριση κινδύνου (UNESCO, 2023; NIST, 2023).

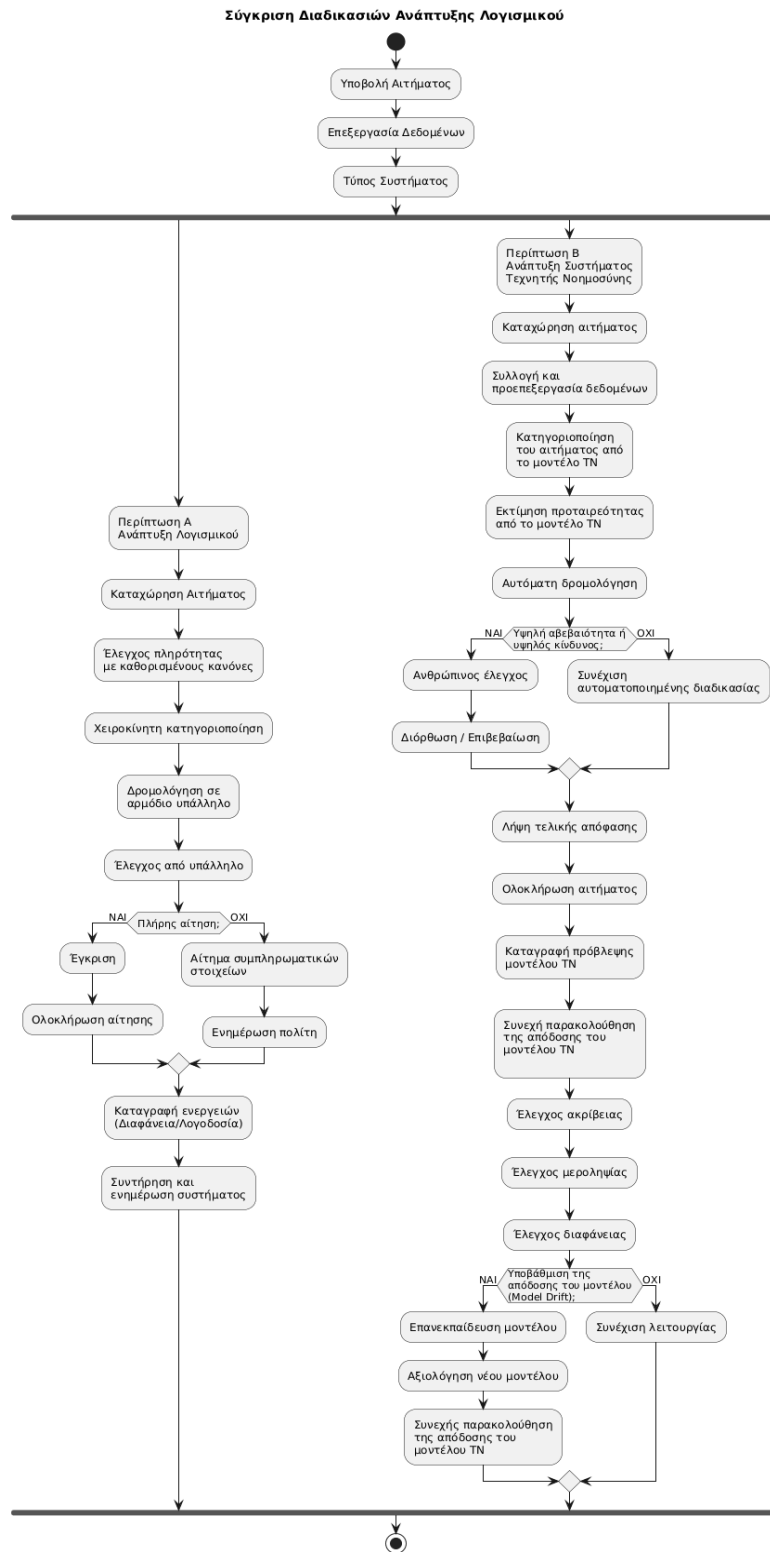
Ιδιαίτερα σημαντικό είναι ότι τα παραδοτέα αυτά απευθύνονται όχι μόνο σε τεχνικούς, αλλά και σε μη τεχνικούς φορείς (διοίκηση, ελεγκτικές αρχές), κάτι που αντικατοπτρίζει τη μετατόπιση της TN

από τεχνικό εργαλείο σε κοινωνικό - τεχνικό σύστημα (Mäntymäki et. al, 2022; United Nations, 2024).

6.3.6 Τι Αναδεικνύει η Περίπτωση Β

Η Περίπτωση Β δείχνει καθαρά ότι η ενσωμάτωση τεχνητής νοημοσύνης στο ίδιο σύστημα δεν αλλάζει μόνο την τεχνολογία, αλλά αλλάζει τον τρόπο με τον οποίο ορίζονται οι ρόλοι, η ευθύνη και η ποιότητα. Ο αναλυτής καλείται να ενσωματώσει ηθικές και κοινωνικές παραμέτρους ως απαιτήσεις, ενώ ο developer καλείται να σχεδιάσει συστήματα που δεν είναι απλώς λειτουργικά αλλά και ελέγξιμα, διαφανή και ασφαλή στον χρόνο. Η σύγκριση με την Περίπτωση Α προετοιμάζει έτσι το έδαφος για την επόμενη ενότητα, όπου θα εξεταστούν οι κύκλοι ζωής λειτουργικών συστημάτων και τεχνητής νοημοσύνης, και θα φανεί γιατί η TN απαιτεί πιο σύνθετες και διαρκείς μορφές διακυβέρνησης (NIST, 2023; UNESCO, 2022; Weidinger et. al, 2023).

Διάγραμμα 2. «Διάγραμμα Σύγκρισης Διαδικασιών Ανάπτυξης Λογισμικού και Ανάπτυξης Συστημάτων Τεχνητής Νοημοσύνης»



6.4 Συγκριτική Ανάλυση των Δύο Μελετών Περίπτωσης (Ανάπτυξη Λογισμικού έναντι Ανάπτυξης Λογισμικού με Χρήση Τεχνητής Νοημοσύνης)

Η συγκριτική σύνθεση των δύο μελετών περίπτωσης αναδεικνύει ότι η διαφορά μεταξύ ανάπτυξης κλασικού λογισμικού και ανάπτυξης συστημάτων τεχνητής νοημοσύνης δεν είναι απλώς τεχνολογική. Πρόκειται για διαφορά παραδείγματος ως προς τον τρόπο που ορίζονται η ποιότητα, η ευθύνη, ο έλεγχος και η ηθική λειτουργία ενός συστήματος. Παρότι και στις δύο περιπτώσεις συναντούμε κοινά στάδια (ανάλυση, σχεδίαση, υλοποίηση, έλεγχος, λειτουργία), η λογική με την οποία παράγεται η «αξιοπιστία» διαφέρει ουσιαστικά.

Στο κλασικό λογισμικό, όπως παρουσιάστηκε στην Περίπτωση Α, η αξιοπιστία θεμελιώνεται πρωτίστως στη σαφήνεια των απαιτήσεων και στη ντετερμινιστική συμπεριφορά του συστήματος. Ο αναλυτής μετατρέπει τις ανάγκες του οργανισμού σε ρητές προδιαγραφές και κανόνες, ενώ ο developer υλοποιεί αυτές τις προδιαγραφές με τρόπο ώστε, εφόσον το σύστημα λειτουργεί σωστά, το αποτέλεσμα να είναι προβλέψιμο. Ένα σφάλμα θεωρείται απόκλιση από τις προδιαγραφές και αντιμετωπίζεται ως τεχνικό πρόβλημα που μπορεί να εντοπιστεί και να διορθωθεί μέσω δοκιμών και ενημερώσεων. Η λογοδοσία, σε αυτή την περίπτωση, είναι σχετικά γραμμική: μπορούμε να ελέγξουμε τον κώδικα, τα αρχεία καταγραφής και τις εκδόσεις για να εντοπίσουμε πού και γιατί κάτι πήγε λάθος (Lu, 2022; UNESCO, 2023).

Αντίθετα, στην Περίπτωση Β, όπου στο ίδιο σύστημα ενσωματώνεται τεχνητή νοημοσύνη, η αξιοπιστία δεν εξαντλείται στη σωστή υλοποίηση. Το σύστημα μπορεί να λειτουργεί ακριβώς όπως σχεδιάστηκε και παρ' όλα αυτά να παράγει ανεπιθύμητα ή άδικα αποτελέσματα, επειδή η συμπεριφορά του προκύπτει από δεδομένα και στατιστικά μοτίβα. Το σφάλμα εδώ δεν είναι πάντα «bug», αλλά λανθασμένη πρόβλεψη ή συστηματική απόκλιση (π.χ. μεροληψία υπέρ ή κατά συγκεκριμένων κατηγοριών αιτημάτων). Αυτός ο μετασχηματισμός της έννοιας του σφάλματος αποτελεί βασικό σημείο διαφοροποίησης και εξηγεί γιατί η ανάπτυξη συστημάτων ΤΝ συνοδεύεται από έντονη συζήτηση γύρω από διαχείριση κινδύνου και συνεχή επιτήρηση (National Institute of Standards and Technology [NIST], 2023; Weidinger et. al, 2023).

Ένα δεύτερο κρίσιμο σημείο σύγκρισης αφορά τον ρόλο του αναλυτή. Στο κλασικό λογισμικό, ο αναλυτής εστιάζει κυρίως στο «τι πρέπει να κάνει» το σύστημα και τότε θεωρείται επιτυχής μια λειτουργία. Στην ΤΝ, ο ρόλος αυτός διευρύνεται: ο αναλυτής καλείται να ορίσει τι σημαίνει «αποδεκτή απόδοση», ποια σφάλματα είναι πιο επιβλαβή από άλλα και ποιοι κίνδυνοι είναι κοινωνικά ή οργανωτικά ανεκτοί. Με άλλα λόγια, ο αναλυτής γίνεται συν-διαμορφωτής του ηθικού

προφίλ του συστήματος, καθώς οι επιλογές του επηρεάζουν άμεσα ποιοι πολίτες θα εξυπηρετηθούν ταχύτερα, ποιοι θα καθυστερήσουν και πώς κατανέμονται οι πόροι. Αυτή η μετατόπιση αντανακλά τη σύγχρονη αντίληψη ότι η ηθική της TN δεν είναι προσθήκη στο τέλος, αλλά ενσωματώνεται στον σχεδιασμό μέσω απαιτήσεων και μετρικών (UNESCO, 2022; UNESCO, 2023).

Ανάλογη διαφοροποίηση παρατηρείται και στον ρόλο του developer. Στην ανάπτυξη λογισμικού, ο developer διαχειρίζεται κυρίως κώδικα, αρχιτεκτονική και τεχνικές εξαρτήσεις. Στην TN, ο developer (ή ML engineer) διαχειρίζεται επιπλέον δεδομένα, μοντέλα και διαδικασίες εκπαίδευσης και επανεκπαίδευσης. Η ευθύνη του επεκτείνεται πέρα από την «ορθή λειτουργία» στην επίβλεψη της συμπεριφοράς του συστήματος στον χρόνο, μέσω μηχανισμών παρακολούθησης, καταγραφής και αξιολόγησης. Αυτή η διαρκής τεχνική φροντίδα δεν είναι απλώς καλή πρακτική· αποτελεί προϋπόθεση για λογοδοσία και ασφάλεια, όπως υπογραμμίζεται σε σύγχρονα πλαίσια διακυβέρνησης της TN (NIST, 2023; Mäntymäki et. al, 2022).

Ιδιαίτερη σημασία έχει και η σύγκριση ως προς τη διαφάνεια (transparency). Στο κλασικό λογισμικό, η διαφάνεια συνδέεται κυρίως με τη δυνατότητα ανασύνθεσης της λογικής μιας διαδικασίας: ποιο βήμα εκτελέστηκε, ποιος κανόνας εφαρμόστηκε, ποιος χρήστης ενήργησε. Στην TN, η διαφάνεια γίνεται πιο σύνθετη, καθώς δεν αρκεί να γνωρίζουμε «τι έγινε», αλλά ανακύπτει το ερώτημα «γιατί το μοντέλο κατέληξε σε αυτή την απόφαση». Η ανάγκη για τεκμηρίωση, καταγραφή και επεξηγησιμότητα (explainability) προκύπτει ακριβώς από αυτή τη διαφορά και συνδέεται άμεσα με τις ανησυχίες για κοινωνική εμπιστοσύνη και δικαιώματα πολιτών (UNESCO, 2022; United Nations, 2024).

Τέλος, η σύγκριση των δύο cases αναδεικνύει διαφορετική ηθική δυναμική. Στο κλασικό λογισμικό, η ηθική διάσταση αφορά κυρίως τη σωστή χρήση δεδομένων, την προστασία ιδιωτικότητας και τη δίκαιη εφαρμογή κανόνων. Στην TN, η ηθική αποκτά επιπλέον διαστάσεις: μεροληψία, άνιση μεταχείριση, αδιαφάνεια αποφάσεων και δυσκολία απόδοσης ευθύνης. Αυτές οι διαστάσεις εξηγούν γιατί διεθνείς οργανισμοί και θεσμοί αντιμετωπίζουν την TN ως κοινωνικό τεχνικό σύστημα που απαιτεί ειδικά εργαλεία διακυβέρνησης και συνεχή αξιολόγηση, πέρα από τις κλασικές πρακτικές ανάπτυξης λογισμικού (UNESCO, 2022; United Nations, 2024; Weidinger et. al, 2023).

Η συγκριτική ανάλυση δείχνει ότι η μετάβαση από το λογισμικό στην τεχνητή νοημοσύνη δεν καταργεί τις βασικές αρχές ανάπτυξης συστημάτων, αλλά τις επαναπροσδιορίζει. Οι ρόλοι παραμένουν, όμως αποκτούν νέο περιεχόμενο· οι διαδικασίες διατηρούνται, αλλά γίνονται κυκλικές και διαρκείς· και η ηθική, από υποστηρικτική αρχή, μετατρέπεται σε κεντρικό κριτήριο αξιολόγησης. Αυτή η σύνθεση προετοιμάζει το έδαφος για την επόμενη ενότητα, όπου θα

εξεταστούν αναλυτικά οι κύκλοι ζωής ενός λειτουργικού συστήματος και της τεχνητής νοημοσύνης, και θα φανεί πώς οι παραπάνω διαφορές αποκτούν πρακτική σημασία στον χρόνο.

6.5 Κύκλος ζωής Λειτουργικού Συστήματος

Ο κύκλος ζωής ενός λειτουργικού συστήματος (operating system lifecycle) αποτελεί ένα χαρακτηριστικό παράδειγμα ώριμης, θεσμοθετημένης τεχνολογικής διαδικασίας, όπου η αξιοπιστία, η ασφάλεια και η λογοδοσία έχουν ενσωματωθεί εδώ και δεκαετίες στον σχεδιασμό και στη λειτουργία. Σε αντίθεση με την τεχνητή νοημοσύνη, το λειτουργικό σύστημα δεν «μαθαίνει» από δεδομένα ούτε τροποποιεί τη συμπεριφορά του αυτόνομα. Παρ' όλα αυτά, επειδή λειτουργεί ως θεμέλιο πάνω στο οποίο εκτελούνται κρίσιμες εφαρμογές και διαχειρίζονται ευαίσθητα δεδομένα, ο κύκλος ζωής του είναι αυστηρά δομημένος και συνοδεύεται από σαφείς διαδικασίες ελέγχου, ενημέρωσης και απόσυρσης.

Η ανάλυση του κύκλου ζωής ενός λειτουργικού συστήματος έχει ιδιαίτερη σημασία για την παρούσα εργασία, διότι προσφέρει ένα σημείο αναφοράς (baseline) δείχνει πώς μια σύνθετη τεχνολογία μπορεί να εξελιχθεί με τρόπο προβλέψιμο, ελεγχόμενο και θεσμικά αποδεκτό. Αυτό το σημείο αναφοράς θα αξιοποιηθεί στη συνέχεια για να φανεί γιατί ο κύκλος ζωής της τεχνητής νοημοσύνης είναι πιο ρευστός και απαιτεί διαφορετικά εργαλεία διακυβέρνησης (National Institute of Standards and Technology [NIST], 2023; United Nations, 2024).

6.5.1 Σχεδίαση και Αρχική Ανάπτυξη

Ο κύκλος ζωής ενός λειτουργικού συστήματος ξεκινά με τη φάση σχεδίασης και αρχικής ανάπτυξης. Σε αυτό το στάδιο καθορίζονται οι βασικές αρχές του συστήματος: αρχιτεκτονική πυρήνα, διαχείριση διεργασιών, μνήμης και συσκευών, μηχανισμοί ασφάλειας και διαχείρισης χρηστών. Οι αποφάσεις αυτές είναι ιδιαίτερα κρίσιμες, καθώς επηρεάζουν το σύστημα για πολλά χρόνια.

Η σχεδίαση δεν είναι μόνο τεχνική. Περιλαμβάνει και κανονιστικές και ηθικές παραδοχές, όπως το πώς προστατεύονται τα δεδομένα χρηστών, πώς εφαρμόζεται ο έλεγχος πρόσβασης και πώς καταγράφονται οι ενέργειες για λόγους λογοδοσίας. Αν και αυτές οι έννοιες συζητούνται σήμερα έντονα στο πλαίσιο της TN, στο λειτουργικό σύστημα αποτελούν εδώ και καιρό θεμελιώδη στοιχεία καλής πρακτικής, καθώς η παραβίασή τους μπορεί να έχει άμεσες επιπτώσεις στην ιδιωτικότητα και στα δικαιώματα των χρηστών (Lu, 2022; UNESCO, 2022).

6.5.2 Κυκλοφορία και Παραγωγική Λειτουργία

Μετά την ολοκλήρωση της αρχικής ανάπτυξης, το λειτουργικό σύστημα περνά στη φάση κυκλοφορίας (release) και τίθεται σε παραγωγική λειτουργία. Σε αυτό το στάδιο, το σύστημα διανέμεται στους χρήστες και αρχίζει να χρησιμοποιείται σε πραγματικές συνθήκες. Η βασική υπόθεση εδώ είναι ότι το λειτουργικό σύστημα θα παραμείνει σταθερό ως προς τη συμπεριφορά του: οι ίδιες εντολές και ρυθμίσεις θα παράγουν τα ίδια αποτελέσματα, εκτός αν υπάρξει συνειδητή τροποποίηση μέσω ενημέρωσης.

Η παραγωγική λειτουργία συνοδεύεται από μηχανισμούς υποστήριξης, τεκμηρίωσης και καταγραφής συμβάντων. Τα αρχεία καταγραφής (logs) και οι μηχανισμοί αναφοράς σφαλμάτων αποτελούν βασικά εργαλεία για τη διαχείριση της ασφάλειας και της αξιοπιστίας. Σε αυτό το σημείο, η λογοδοσία είναι σαφώς οριοθετημένη, ο κατασκευαστής/διαχειριστής του λειτουργικού συστήματος φέρει την ευθύνη για την ορθή λειτουργία και την αντιμετώπιση κρίσιμων προβλημάτων (UNESCO, 2023; United Nations, 2024).

6.5.3 Ενημερώσεις, Διορθώσεις και Διαχείριση Ευπαθειών

Ένα από τα πιο χαρακτηριστικά στοιχεία του κύκλου ζωής ενός λειτουργικού συστήματος είναι η συνεχής παροχή ενημερώσεων (updates). Οι ενημερώσεις αυτές μπορεί να αφορούν διορθώσεις σφαλμάτων, βελτιώσεις απόδοσης ή, κυρίως, επιδιορθώσεις ασφάλειας (security patches). Η ύπαρξη αυτής της φάσης αναγνωρίζει ότι κανένα σύστημα δεν είναι τέλει από την αρχή και ότι η ασφάλεια είναι μια διαρκής διαδικασία.

Από ηθική και θεσμική σκοπιά, οι ενημερώσεις αποτελούν μηχανισμό προστασίας των χρηστών. Η μη έγκαιρη διόρθωση γνωστών ευπαθειών μπορεί να εκθέσει δεδομένα ή να επιτρέψει κακόβουλες ενέργειες, γεγονός που μεταφράζεται σε αποτυχία εκπλήρωσης θεμελιωδών υποχρεώσεων απέναντι στην ιδιωτικότητα και την ασφάλεια (Lu, 2022; UNESCO, 2022). Σε αυτό το πλαίσιο, η διαφάνεια είναι κρίσιμη: οι χρήστες πρέπει να γνωρίζουν τι διορθώνεται, γιατί και με ποιες επιπτώσεις, ώστε να μπορούν να αξιολογήσουν τον κίνδυνο και να λάβουν αποφάσεις.

6.5.4 Μακροχρόνια Συντήρηση και Σταθερότητα

Καθώς το λειτουργικό σύστημα ωριμάζει, εισέρχεται σε μια φάση μακροχρόνιας συντήρησης. Σε αυτή τη φάση, η έμφαση μετατοπίζεται από την εισαγωγή νέων χαρακτηριστικών στη διατήρηση της σταθερότητας και της συμβατότητας. Πολλοί οργανισμοί βασίζουν κρίσιμες λειτουργίες σε συγκεκριμένες εκδόσεις λειτουργικών συστημάτων, γεγονός που καθιστά την προβλεψιμότητα και τη σταθερότητα βασικές αξίες.

Αυτή η φάση αναδεικνύει ένα σημαντικό στοιχείο: το λειτουργικό σύστημα αντιμετωπίζεται ως υποδομή και όχι ως πειραματική τεχνολογία. Η ηθική διάσταση εδώ συνδέεται με την ευθύνη απέναντι σε οργανισμούς και κοινωνίες που εξαρτώνται από τη συνεχή λειτουργία του συστήματος. Η αιφνίδια ή ανεξέλεγκτη αλλαγή θα μπορούσε να προκαλέσει σοβαρές κοινωνικές και οικονομικές επιπτώσεις, κάτι που αναγνωρίζεται και στη διεθνή συζήτηση για υπεύθυνη τεχνολογική διακυβέρνηση (United Nations, 2024).

6.5.5 Τέλος Κύκλου Ζωής και Απόσυρση (End-Of-Life)

Ο κύκλος ζωής ενός λειτουργικού συστήματος ολοκληρώνεται με τη φάση απόσυρσης (End-Of-Life). Σε αυτό το στάδιο, ο κατασκευαστής ανακοινώνει ότι το σύστημα δεν θα λαμβάνει πλέον ενημερώσεις ή υποστήριξη. Η ανακοίνωση αυτή δεν είναι απλώς τεχνική πληροφορία έχει άμεσες επιπτώσεις στην ασφάλεια και στη συμμόρφωση οργανισμών και χρηστών.

Η απόσυρση συνοδεύεται συνήθως από σαφή χρονοδιαγράμματα και προτάσεις μετάβασης σε νεότερες εκδόσεις. Αυτή η πρακτική ενσωματώνει μια βασική αρχή υπεύθυνης τεχνολογίας: οι χρήστες δεν αφήνονται εκτεθειμένοι χωρίς προειδοποίηση. Η ύπαρξη σαφούς τέλους κύκλου ζωής επιτρέπει τον προγραμματισμό, τη μείωση κινδύνων και τη διασφάλιση της συνέχειας λειτουργίας κρίσιμων συστημάτων (UNESCO, 2023; United).

6.5.6 Τι μας Δείχνει ο Κύκλος Ζωής του Λειτουργικού Συστήματος

Συνολικά, ο κύκλος ζωής ενός λειτουργικού συστήματος δείχνει πώς μια σύνθετη τεχνολογία μπορεί να ενταχθεί σε σταθερό θεσμικό και οργανωτικό πλαίσιο. Κάθε φάση από τη σχεδίαση έως την απόσυρση είναι προβλέψιμη, τεκμηριωμένη και συνοδεύεται από σαφείς ρόλους και ευθύνες. Η διαφάνεια, η λογοδοσία και η προστασία των χρηστών δεν αποτελούν αφηρημένες αρχές, αλλά υλοποιούνται μέσω συγκεκριμένων διαδικασιών και παραδοτέων.

Αυτή η εικόνα λειτουργεί ως αντίστιξη στον κύκλο ζωής της τεχνητής νοημοσύνης που θα αναλυθεί στην επόμενη ενότητα. Εκεί, θα φανεί ότι η απουσία σταθερής συμπεριφοράς, η εξάρτηση από δεδομένα και η ανάγκη συνεχούς επανεκπαίδευσης καθιστούν ανεπαρκές το κλασικό μοντέλο κύκλου ζωής. Η σύγκριση αυτή αναδεικνύει γιατί η TN απαιτεί νέες μορφές διακυβέρνησης και ηθικής εποπτείας, πέρα από τα καθιερωμένα πρότυπα που λειτούργησαν αποτελεσματικά για το λειτουργικό σύστημα (NIST, 2023; Weidinger et. al, 2023).

6.6 Κύκλος Ζωής Μοντέλου Συστημάτων Τεχνητής Νοημοσύνης (Model Lifecycle)

Ο κύκλος ζωής της τεχνητής νοημοσύνης διαφέρει ριζικά από τον κύκλο ζωής ενός λειτουργικού συστήματος, όχι επειδή είναι «νεότερη» τεχνολογία, αλλά επειδή η ίδια της η φύση είναι δυναμική, πιθανοκρατική και εξαρτημένη από δεδομένα. Ενώ το λειτουργικό σύστημα εξελίσσεται μέσω συνειδητών τεχνικών παρεμβάσεων (ενημερώσεις, διορθώσεις, εκδόσεις), ένα σύστημα τεχνητής νοημοσύνης μπορεί να αλλάξει τη συμπεριφορά του ακόμη και χωρίς καμία αλλαγή στον κώδικα, απλώς επειδή αλλάζουν τα δεδομένα ή το περιβάλλον χρήσης. Αυτή η ιδιαιτερότητα καθιστά τον κύκλο ζωής της TN λιγότερο γραμμικό και περισσότερο κυκλικό και επαναληπτικό, με έντονη έμφαση στη συνεχή επιτήρηση, αξιολόγηση και αναθεώρηση (National Institute of Standards and Technology [NIST], 2023; Weidinger et. al, 2023).

Η ανάλυση του κύκλου ζωής της TN στην παρούσα εργασία δεν γίνεται μόνο για τεχνικούς λόγους. Γίνεται κυρίως για να αναδειχθεί πώς η ηθική, η διαφάνεια και η λογοδοσία δεν μπορούν να αντιμετωπιστούν ως μεμονωμένα στάδια, αλλά πρέπει να διατρέχουν ολόκληρη τη διάρκεια ζωής του συστήματος. Αυτός ακριβώς είναι και ο λόγος που διεθνείς οργανισμοί μιλούν πλέον για “Governance Across the AI Lifecycle” και όχι απλώς για «ηθικό σχεδιασμό» στην αρχή ενός έργου (UNESCO, 2022; United Nations, 2024).

6.6.1 Διατύπωση Προβλήματος και Σκοπού του Μοντέλου

Ο κύκλος ζωής της τεχνητής νοημοσύνης ξεκινά με τη διατύπωση του προβλήματος (problem framing). Σε αντίθεση με το κλασικό λογισμικό, όπου το πρόβλημα συχνά ορίζεται ως μια ακολουθία λειτουργιών, στην TN το πρόβλημα πρέπει να μετατραπεί σε πρόβλημα πρόβλεψης ή ταξινόμησης. Στο case της εργασίας, αυτό σημαίνει ότι η ανάγκη να κατηγοριοποιούνται και να προτεραιοποιούνται αιτήματα πολιτών μετασχηματίζεται σε ερώτημα του τύπου: «με ποια πιθανότητα ένα αίτημα ανήκει σε συγκεκριμένη κατηγορία ή επίπεδο προτεραιότητας».

Η φάση αυτή είναι κρίσιμη από ηθική σκοπιά, διότι εδώ λαμβάνονται αποφάσεις που επηρεάζουν έμμεσα την κοινωνική πραγματικότητα. Το πώς ορίζεται η «επιτυχία» του μοντέλου, ποιες μετρικές χρησιμοποιούνται και ποια σφάλματα θεωρούνται πιο σοβαρά από άλλα δεν είναι ουδέτερες επιλογές. Αντικατοπτρίζουν αξίες, προτεραιότητες και αντιλήψεις δικαιοσύνης, γι’ αυτό και η διεθνής βιβλιογραφία τονίζει ότι η ηθική της TN ξεκινά ήδη από τη διατύπωση του προβλήματος και όχι από την τελική εφαρμογή (UNESCO, 2022; UNESCO, 2023).

6.6.2 Συλλογή, Επιλογή και Τεκμηρίωση Δεδομένων

Ακολουθεί η φάση των δεδομένων, η οποία αποτελεί τη ραχοκοκαλιά κάθε συστήματος τεχνητής νοημοσύνης.

Σε αυτή τη φάση συλλέγονται ιστορικά δεδομένα, επιλέγονται μεταβλητές και γίνεται καθαρισμός και προετοιμασία. Σε αντίθεση με το λειτουργικό σύστημα, όπου τα δεδομένα είναι είσοδος αλλά όχι καθοριστικός παράγοντας συμπεριφοράς, στην TN τα δεδομένα διαμορφώνουν άμεσα τον τρόπο που “σκέφτεται” το σύστημα.

Η ποιότητα και η αντιπροσωπευτικότητα των δεδομένων συνδέονται άμεσα με ζητήματα μεροληψίας (bias) και άνισης μεταχείρισης. Αν τα δεδομένα αντανakλούν παλαιότερες πρακτικές, κοινωνικές ανισότητες ή ελλείψεις, το μοντέλο κινδυνεύει να τις αναπαράγει. Για τον λόγο αυτό, σύγχρονα ηθικά και κανονιστικά πλαίσια απαιτούν όχι μόνο τεχνικό έλεγχο των δεδομένων, αλλά και τεκμηρίωση της προέλευσής τους, των περιορισμών τους και των πιθανών κοινωνικών επιπτώσεων της χρήσης τους (UNESCO, 2023; United Nations, 2024; Odilla, 2024).

6.6.3 Εκπαίδευση και Αξιολόγηση του Μοντέλου

Η επόμενη φάση είναι η εκπαίδευση του μοντέλου (training) και η αρχική του αξιολόγηση.

Εδώ το σύστημα «μαθαίνει» συσχετίσεις και μοτίβα από τα δεδομένα και αποκτά τη δυνατότητα να παράγει προβλέψεις. Η αξιολόγηση γίνεται με στατιστικές μετρικές, όμως αυτές από μόνες τους δεν αρκούν για να κριθεί αν το σύστημα είναι κατάλληλο για χρήση.

Στο πλαίσιο της υπεύθυνης TN, η αξιολόγηση πρέπει να περιλαμβάνει και ποιοτικά και ηθικά κριτήρια, όπως η επίδραση διαφορετικών τύπων σφαλμάτων, η πιθανή μεροληψία έναντι συγκεκριμένων ομάδων και η σταθερότητα των αποτελεσμάτων. Όπως επισημαίνεται στη σύγχρονη βιβλιογραφία, ένα μοντέλο μπορεί να είναι «τεχνικά αποδεκτό» αλλά κοινωνικά προβληματικό, γεγονός που καθιστά αναγκαία τη διεπιστημονική αξιολόγηση πριν από την παραγωγική χρήση (Weidinger et al., 2023; UNESCO, 2022).

6.6.4 Ενσωμάτωση και Παραγωγική Λειτουργία

Μετά την αρχική αξιολόγηση, το μοντέλο ενσωματώνεται στο πληροφοριακό σύστημα και τίθεται σε παραγωγική λειτουργία. Σε αυτό το στάδιο, η TN αρχίζει να επηρεάζει πραγματικές διαδικασίες και ανθρώπινες εμπειρίες. Στο case της εργασίας, η προτεραιοποίηση αιτημάτων από το μοντέλο επηρεάζει τον χρόνο εξυπηρέτησης πολιτών και, έμμεσα, την αντίληψή τους για τη δικαιοσύνη και την αποτελεσματικότητα της διοίκησης.

Η παραγωγική λειτουργία δεν σημαίνει σταθερότητα. Αντιθέτως, αποτελεί την αρχή μιας φάσης συνεχούς παρατήρησης, όπου συλλέγονται στοιχεία για την πραγματική απόδοση του μοντέλου. Η ανάγκη αυτή έχει οδηγήσει στην ανάπτυξη πλαισίων διαχείρισης κινδύνου που αντιμετωπίζουν την TN ως σύστημα που πρέπει να παρακολουθείται διαρκώς, ακριβώς επειδή οι συνέπειες των αποφάσεων του εκτείνονται πέρα από το τεχνικό επίπεδο (NIST, 2023; United Nations, 2024).

6.6.5 Συνεχής Παρακολούθηση, Υποβάθμιση και Επανεκπαίδευση

Ένα από τα πιο χαρακτηριστικά στοιχεία του κύκλου ζωής της TN είναι η υποβάθμιση του μοντέλου στον χρόνο (model drift). Καθώς αλλάζουν τα δεδομένα εισόδου, οι κοινωνικές συνθήκες ή οι πρακτικές των χρηστών, το μοντέλο μπορεί να γίνει λιγότερο ακριβές ή πιο μεροληπτικό. Αυτό σημαίνει ότι η ποιότητα της TN δεν διασφαλίζεται μία φορά, αλλά απαιτεί συνεχή επανεκτίμηση και, ενδεχομένως, επανεκπαίδευση.

Η φάση αυτή είναι κρίσιμη από ηθική άποψη, διότι ένα σύστημα που δεν επιτηρείται μπορεί να συνεχίσει να λειτουργεί «νόμιμα» αλλά να παράγει άδικα αποτελέσματα. Η διεθνής βιβλιογραφία τονίζει ότι η απουσία μηχανισμών παρακολούθησης και αναθεώρησης αποτελεί μία από τις σοβαρότερες πηγές κινδύνου στην εφαρμοσμένη TN (Weidinger et al., 2023; NIST, 2023).

6.6.6 Απόσυρση, Αντικατάσταση και Ηθικό Τέλος Κύκλου Ζωής

Τέλος, όπως και στο λειτουργικό σύστημα, έτσι και στην TN υπάρχει η φάση απόσυρσης του μοντέλου. Ωστόσο, εδώ το τέλος κύκλου ζωής είναι λιγότερο σαφές. Ένα μοντέλο μπορεί να αντικατασταθεί από νεότερο, να τροποποιηθεί ριζικά ή να πάψει να χρησιμοποιείται επειδή κρίθηκε αναξιόπιστο ή ηθικά προβληματικό. Η απόφαση αυτή δεν είναι μόνο τεχνική είναι και ηθική και πολιτική, καθώς αφορά το αν ένα σύστημα συνεχίζει να εξυπηρετεί το δημόσιο συμφέρον.

Η ύπαρξη σαφών διαδικασιών απόσυρσης και τεκμηρίωσης θεωρείται πλέον αναγκαία για τη διαφύλαξη της εμπιστοσύνης του κοινού και τη διασφάλιση λογοδοσίας. Ένα σύστημα TN δεν πρέπει απλώς να «σβήνεται», αλλά να αφήνει πίσω του ίχνη που επιτρέπουν έλεγχο, αξιολόγηση και μάθηση για μελλοντικές εφαρμογές (UNESCO, 2023; United Nations, 2024).

6.6.7 Τι Αποκαλύπτει ο Κύκλος Ζωής της Τεχνητής Νοημοσύνης

Συνολικά, ο κύκλος ζωής της τεχνητής νοημοσύνης αναδεικνύει ότι πρόκειται για τεχνολογία σε μόνιμη εξέλιξη, όπου η ευθύνη δεν σταματά ποτέ πραγματικά. Σε αντίθεση με το λειτουργικό σύστημα, που μπορεί να θεωρηθεί «σταθερό προϊόν με ημερομηνία λήξης», η TN είναι μια διαδικασία που απαιτεί συνεχή φροντίδα, επανεξέταση και ηθική εγρήγορση. Αυτή η διαφορά

εξηγεί γιατί η TN αντιμετωπίζεται πλέον διεθνώς ως κοινωνικό-τεχνικό σύστημα που χρειάζεται ειδικά εργαλεία διακυβέρνησης, αξιολόγησης και πολιτικής, και όχι απλώς ως ένα ακόμη τεχνικό artefact (UNESCO, 2022; United Nations, 2024; Weidinger et. al, 2023).

Η κατανόηση αυτού του κύκλου ζωής αποτελεί το αναγκαίο υπόβαθρο για την επόμενη ενότητα, όπου θα συζητηθούν συγκεντρωτικά η διαφάνεια, η μεροληψία και οι ηθικές απαιτήσεις που πρέπει να πληροί ένα σύστημα ώστε να μπορεί να επεξεργάζεται δεδομένα και να λαμβάνει αποφάσεις με τρόπο υπεύθυνο και κοινωνικά αποδεκτό.

6.7 Ηθικές και Λειτουργικές Απαιτήσεις Συστημάτων Τεχνητής Νοημοσύνης

Η ενσωμάτωση τεχνητής νοημοσύνης σε πληροφοριακά συστήματα καθιστά αναγκαία μια σαφή μετάβαση από την αφηρημένη συζήτηση περί «ηθικών αρχών» σε συγκεκριμένες ηθικές και λειτουργικές απαιτήσεις, οι οποίες μπορούν να ενσωματωθούν, να ελεγχθούν και να τεκμηριωθούν σε όλο τον κύκλο ζωής του συστήματος. Στο πλαίσιο της παρούσας εργασίας, οι έννοιες της διαφάνειας, της μεροληψίας και της ηθικής εκ σχεδιασμού δεν αντιμετωπίζονται ως θεωρητικά ιδεώδη, αλλά ως πρακτικές προδιαγραφές που καθορίζουν αν ένα σύστημα είναι κοινωνικά αποδεκτό, θεσμικά αξιόπιστο και λειτουργικά υπεύθυνο (UNESCO, 2022; United Nations, 2024; National Institute of Standards and Technology [NIST], 2023).

Η σημασία αυτών των απαιτήσεων αναδεικνύεται ιδιαίτερα όταν η TN χρησιμοποιείται σε περιβάλλοντα που επηρεάζουν δικαιώματα, πρόσβαση σε υπηρεσίες ή κατανομή πόρων. Σε τέτοια περιβάλλοντα, ένα σύστημα μπορεί να είναι τεχνικά αποδοτικό αλλά ταυτόχρονα ηθικά ανεπαρκές, εάν δεν πληροί ελάχιστες προϋποθέσεις διαφάνειας, δικαιοσύνης και λογοδοσίας (Lu, 2022; Odilla, 2024).

6.7.1 Διαφάνεια (transparency): Από την Τεχνική Ορατότητα στη Θεσμική Κατανόηση

Η διαφάνεια αποτελεί μία από τις πιο συχνά αναφερόμενες, αλλά και πιο παρεξηγημένες, απαιτήσεις στην ηθική της τεχνητής νοημοσύνης. Στο κλασικό λογισμικό, η διαφάνεια συνδέεται κυρίως με την ιχνηλασιμότητα διαδικασιών: ποιος χρήστης έκανε ποια ενέργεια, ποιος κανόνας εφαρμόστηκε και ποιο ήταν το αποτέλεσμα. Στην TN, όμως, η διαφάνεια αποκτά βαθύτερη σημασία, καθώς αφορά όχι μόνο το *τι* συνέβη, αλλά και το *πώς* και *γιατί* προέκυψε μια απόφαση.

Σύγχρονα ηθικά και κανονιστικά πλαίσια τονίζουν ότι η διαφάνεια δεν ταυτίζεται απαραίτητα με την πλήρη αποκάλυψη του εσωτερικού μηχανισμού ενός μοντέλου, αλλά με τη δυνατότητα παροχής

νοηματικά επαρκών εξηγήσεων σε διαφορετικά ακροατήρια: τεχνικούς, διοίκηση, εποπτικές αρχές και πολίτες (UNESCO, 2022; UNESCO, 2023). Στο case της εργασίας, αυτό σημαίνει ότι το σύστημα οφείλει να μπορεί να τεκμηριώσει γιατί ένα αίτημα προτεραιοποιήθηκε ή κατηγοριοποιήθηκε με συγκεκριμένο τρόπο, ακόμη και αν η απόφαση προήλθε από μοντέλο μηχανικής μάθησης.

Η διαφάνεια, επομένως, λειτουργεί ως προϋπόθεση εμπιστοσύνης. Χωρίς αυτήν, οι αποφάσεις της TN γίνονται αντιληπτές ως αυθαίρετες ή «μαύρο κουτί», γεγονός που υπονομεύει τη νομιμοποίηση της τεχνολογίας στη δημόσια σφαίρα (United Nations, 2024; Weidinger et. al, 2023).

Η σχεδίαση και ανάπτυξη συστημάτων Τεχνητής Νοημοσύνης απαιτεί πλέον μια προσέγγιση βασισμένη στην αξιολόγηση κινδύνου. Σύμφωνα με τη φιλοσοφία του AI Act, δεν παρουσιάζουν όλα τα συστήματα TN τον ίδιο βαθμό πιθανών επιπτώσεων. Ως αποτέλεσμα, οι απαιτήσεις που πρέπει να ικανοποιούνται διαφοροποιούνται ανάλογα με το επίπεδο κινδύνου που ενδέχεται να δημιουργεί κάθε εφαρμογή.

6.7.2 Μεροληψία (Bias): Δομικό Πρόβλημα και όχι Τεχνική Αστοχία

Η μεροληψία στα συστήματα τεχνητής νοημοσύνης δεν αποτελεί απλή τεχνική ατέλεια, αλλά δομικό χαρακτηριστικό που προκύπτει από τον τρόπο με τον οποίο συλλέγονται, επιλέγονται και ερμηνεύονται τα δεδομένα. Σε αντίθεση με το κλασικό λογισμικό, όπου ένας κανόνας εφαρμόζεται το ίδιο για όλους, η TN μαθαίνει από ιστορικά δεδομένα που συχνά αντανακλούν κοινωνικές ανισότητες, διοικητικές πρακτικές ή παλαιότερες αποφάσεις με ενσωματωμένες προκαταλήψεις.

Η διεθνής βιβλιογραφία έχει δείξει ότι η μεροληψία μπορεί να εμφανιστεί σε πολλαπλά επίπεδα: στα δεδομένα εκπαίδευσης, στην επιλογή χαρακτηριστικών, στις μετρικές αξιολόγησης ή ακόμη και στον τρόπο που ερμηνεύονται τα αποτελέσματα (Odilla, 2024; Yen & Hung, 2021). Στο πλαίσιο της παρούσας μελέτης περίπτωσης, ένα σύστημα που προτεραιοποιεί αιτήματα πολιτών μπορεί, για παράδειγμα, να ευνοεί συγκεκριμένους τύπους αιτημάτων ή ομάδες χρηστών, όχι από πρόθεση, αλλά επειδή αυτό έμαθε από το παρελθόν.

Για τον λόγο αυτό, η αντιμετώπιση της μεροληψίας δεν μπορεί να περιορίζεται σε τεχνικές διορθώσεις. Απαιτεί συνεχή αξιολόγηση, τεκμηρίωση και κοινωνική επίγνωση, καθώς και σαφείς αποφάσεις για το τι θεωρείται δίκαιο ή αποδεκτό σε ένα συγκεκριμένο θεσμικό πλαίσιο (UNESCO, 2022; United Nations, 2024).

Η συνεχής παρακολούθηση, η αξιολόγηση επιδόσεων, η τεκμηρίωση και η ανθρώπινη εποπτεία αποτελούν βασικά στοιχεία μιας υπεύθυνης διαδικασίας σχεδίασης και ανάπτυξης συστημάτων TN.

Με τον τρόπο αυτό επιδιώκεται η δημιουργία συστημάτων που δεν είναι μόνο τεχνικά αποτελεσματικά αλλά και κοινωνικά αποδεκτά και αξιόπιστα.

6.7.3 Ηθική εκ Σχεδιασμού (Ethics-By-Design): Από τις Αρχές στις Απαιτήσεις

Η έννοια της ηθικής εκ σχεδιασμού (ethics-by-design) εκφράζει τη μετάβαση από τη ρητορική της «υπεύθυνης χρήσης» στη συστηματική ενσωμάτωση ηθικών αξιών στον σχεδιασμό και στην υλοποίηση των συστημάτων. Αντί η ηθική να εξετάζεται εκ των υστέρων, ενσωματώνεται από τα πρώτα στάδια της ανάλυσης απαιτήσεων και της αρχιτεκτονικής.

Σύμφωνα με διεθνείς κατευθυντήριες γραμμές, η ethics-by-design προσλαμβάνει πρακτική μορφή μέσω απαιτήσεων όπως: ελάχιστη συλλογή δεδομένων, σαφής ορισμός σκοπού, δυνατότητα ανθρώπινης παρέμβασης, μηχανισμοί ελέγχου και τεκμηρίωσης, καθώς και διαδικασίες διαχείρισης συμβάντων και παραπόνων (UNESCO, 2022; NIST, 2023). Στο case της εργασίας, αυτό σημαίνει ότι το σύστημα δεν αρκεί να «λειτουργεί σωστά», αλλά πρέπει να έχει ενσωματωμένους μηχανισμούς που επιτρέπουν έλεγχο, αμφισβήτηση και διόρθωση αποφάσεων.

Η ethics-by-design συνδέεται άμεσα με τη θεσμική διάσταση της TN. Όπως επισημαίνεται στη σύγχρονη συζήτηση για τη διακυβέρνηση της τεχνητής νοημοσύνης, η ευθύνη δεν βαρύνει μόνο τους χρήστες ή τους προγραμματιστές, αλλά και τον ίδιο τον τρόπο με τον οποίο σχεδιάζεται και προσφέρεται το σύστημα (Mäntymäki et al., 2022; United Nations, 2024).

Η προσέγγιση αυτή ενισχύει την έννοια του ethics-by-design, καθώς οι πιθανοί κίνδυνοι λαμβάνονται υπόψη ήδη από τα πρώτα στάδια ανάπτυξης του συστήματος. Παράλληλα, η διαχείριση κινδύνου δεν αντιμετωπίζεται ως μία εφάπαξ διαδικασία αλλά ως συνεχής δραστηριότητα καθ' όλη τη διάρκεια του κύκλου ζωής του συστήματος.

6.7.4 Λειτουργικές Απαιτήσεις για Υπεύθυνη Επεξεργασία και Λήψη Αποφάσεων

Οι ηθικές αρχές αποκτούν πρακτική αξία μόνο όταν μεταφράζονται σε λειτουργικές απαιτήσεις. Ένα σύστημα τεχνητής νοημοσύνης που επεξεργάζεται δεδομένα και λαμβάνει αποφάσεις οφείλει να παράγει συγκεκριμένα «ίχνη» λειτουργίας: καταγραφές εισόδων και εξόδων, τεκμηρίωση αποφάσεων, εκτιμήσεις αβεβαιότητας και δυνατότητα αναδρομικού ελέγχου. Χωρίς αυτά, η λογοδοσία παραμένει θεωρητική.

Σύγχρονα πλαίσια διαχείρισης κινδύνου τονίζουν ότι η υπεύθυνη TN απαιτεί όχι μόνο τεχνική ακρίβεια, αλλά και θεσμική αναγνωσιμότητα: οι αποφάσεις πρέπει να μπορούν να παρουσιαστούν, να εξηγηθούν και, αν χρειαστεί, να αμφισβητηθούν από μη τεχνικούς φορείς (NIST, 2023; UNESCO, 2023). Αυτό μετατρέπει τη διαφάνεια, τη μεροληψία και την ηθική από αφηρημένες αξίες σε μετρήσιμα και ελέγξιμα χαρακτηριστικά του συστήματος.

6.7.5 Συνολική Αποτίμηση

Συνολικά, η ανάλυση των ηθικών και λειτουργικών απαιτήσεων δείχνει ότι η υπεύθυνη εξέλιξη συστημάτων τεχνητής νοημοσύνης δεν εξαρτάται από μία μεμονωμένη τεχνική λύση ή έναν «ηθικό αλγόριθμο». Εξαρτάται από τη συνεκτική ενσωμάτωση διαφάνειας, αντιμετώπισης μεροληψίας και *ethics-by-design* σε όλο τον κύκλο ζωής του συστήματος. Αυτή η προσέγγιση επιτρέπει τη μετάβαση από τη θεωρία στην πράξη και συνδέει άμεσα την τεχνολογία με τις κοινωνικές αξίες και τα ανθρώπινα δικαιώματα που η παρούσα εργασία έχει θέσει στο επίκεντρό της (UNESCO, 2022; United Nations, 2024; Weidinger et. al, 2023).

Η προσέγγιση *Ethics by Design* αποκτά ουσιαστικό περιεχόμενο όταν συνδυάζεται με διαδικασίες αξιολόγησης κινδύνων και αξιολόγησης επιπτώσεων των συστημάτων τεχνητής νοημοσύνης. Η υπεύθυνη ανάπτυξη δεν περιορίζεται στην ενσωμάτωση ηθικών αρχών κατά τον σχεδιασμό ενός συστήματος, αλλά περιλαμβάνει και τη συστηματική εκτίμηση των πιθανών συνεπειών που μπορεί να προκύψουν για άτομα, κοινωνικές ομάδες και το ευρύτερο κοινωνικό σύνολο. Η αξιολόγηση αυτή πρέπει να λαμβάνει υπόψη τόσο την προβλεπόμενη χρήση όσο και πιθανές περιπτώσεις κακής χρήσης ενός συστήματος ΤΝ. Παράλληλα, τα αποτελέσματα της αξιολόγησης των επιπτώσεων πρέπει να αξιοποιούνται κατά τη διαδικασία εκτίμησης κινδύνου, ώστε να σχεδιάζονται κατάλληλα μέτρα ελέγχου και βελτίωσης. Με αυτόν τον τρόπο η ηθική μετατρέπεται από θεωρητική αρχή σε πρακτική διαδικασία διαχείρισης καθ' όλη τη διάρκεια του κύκλου ζωής ενός συστήματος τεχνητής νοημοσύνης.

6.8 Συμπεράσματα Κεφαλαίου

Το παρόν κεφάλαιο ανέδειξε με συστηματικό τρόπο ότι η σύγκριση μεταξύ κλασικού λογισμικού και τεχνητής νοημοσύνης δεν αποτελεί απλώς τεχνική άσκηση, αλλά εννοιολογικό και ηθικό σημείο καμπής για τον τρόπο με τον οποίο σχεδιάζονται, υλοποιούνται και αξιολογούνται σύγχρονα πληροφοριακά συστήματα. Μέσα από την ανάλυση των δύο *cases*, κατέστη σαφές ότι, ενώ το παραδοσιακό λογισμικό στηρίζεται κυρίως σε προκαθορισμένους κανόνες και σαφώς οριοθετημένες ευθύνες, η τεχνητή νοημοσύνη εισάγει ένα διαφορετικό παράδειγμα, όπου η συμπεριφορά του συστήματος αναδύεται από δεδομένα και πιθανοκρατικά μοντέλα και δεν μπορεί να προβλεφθεί πλήρως εκ των προτέρων (Weidinger et. al, 2023; NIST, 2023).

Η ανάλυση των ρόλων ανέδειξε ότι τόσο στο λογισμικό όσο και στην ΤΝ ο αναλυτής και ο *developer* παραμένουν κρίσιμοι, αλλά ο ρόλος τους μετασχηματίζεται. Στην περίπτωση της τεχνητής νοημοσύνης, ο αναλυτής καλείται να λειτουργήσει όχι μόνο ως τεχνικός μεταφραστής απαιτήσεων, αλλά και ως διαμεσολαβητής αξιών, ορίζοντας τι σημαίνει επιτυχία, δικαιοσύνη και αποδεκτό σφάλμα σε ένα συγκεκριμένο κοινωνικό και θεσμικό πλαίσιο. Αντίστοιχα, ο *developer* δεν περιορίζεται στην υλοποίηση αλγορίθμων, αλλά συμμετέχει ενεργά στη διαχείριση δεδομένων,

στη συνεχή επιτήρηση του μοντέλου και στη διασφάλιση ότι το σύστημα παραμένει αξιόπιστο και ελέγξιμο στον χρόνο (Mäntymäki et. al, 2022; UNESCO, 2022).

Η σύγκριση των κύκλων ζωής ανέδειξε μία από τις πιο ουσιώδεις διαφορές μεταξύ των δύο τεχνολογικών προσεγγίσεων. Ο κύκλος ζωής ενός λειτουργικού συστήματος χαρακτηρίζεται από σχετική σταθερότητα και σαφείς φάσεις ανάπτυξης, διάθεσης, συντήρησης και απόσυρσης. Αντιθέτως, ο κύκλος ζωής της τεχνητής νοημοσύνης είναι διαρκώς επαναληπτικός, καθώς η απόδοση και η συμπεριφορά του συστήματος εξαρτώνται από μεταβαλλόμενα δεδομένα και κοινωνικές συνθήκες. Η ανάγκη συνεχούς επανεκπαίδευσης, αξιολόγησης και ενδεχόμενης απόσυρσης καθιστά την ευθύνη μόνιμη και όχι περιορισμένη χρονικά (NIST, 2023; United Nations, 2024).

Ιδιαίτερη σημασία είχε η εστίαση στις ηθικές και λειτουργικές απαιτήσεις. Η διαφάνεια, η αντιμετώπιση της μεροληψίας και η ηθική εκ σχεδιασμού δεν προέκυψαν ως αφηρημένες αρχές, αλλά ως απαραίτητες προϋποθέσεις λειτουργικότητας και θεσμικής νομιμοποίησης. Ένα σύστημα τεχνητής νοημοσύνης που δεν μπορεί να εξηγήσει τις αποφάσεις του, να ελεγχθεί ως προς πιθανές μεροληψίες και να ενσωματώσει μηχανισμούς λογοδοσίας, δεν αποτυγχάνει μόνο ηθικά, αλλά και πρακτικά, καθώς υπονομεύει την εμπιστοσύνη των χρηστών και της κοινωνίας (UNESCO, 2023; Lu, 2022; Odilla, 2024).

Το κεφάλαιο κατέδειξε ότι η υπεύθυνη ανάπτυξη τεχνητής νοημοσύνης δεν μπορεί να αντιμετωπιστεί ως επέκταση των πρακτικών του κλασικού λογισμικού. Απαιτεί ένα διευρυμένο πλαίσιο σκέψης, όπου η τεχνική αρτιότητα συνυπάρχει με τη θεσμική ευθύνη, τη συνεχή επιτήρηση και τη ρητή ενσωμάτωση κοινωνικών αξιών. Αυτή η διαπίστωση λειτουργεί ως φυσική γέφυρα προς το τελικό κεφάλαιο της εργασίας, στο οποίο η τεχνητή νοημοσύνη προσεγγίζεται όχι μόνο ως τεχνολογικό εργαλείο, αλλά ως κοινωνικό φαινόμενο που απαιτεί διαρκή ηθική εγρήγορση, συλλογική ρύθμιση και υπεύθυνη πολιτική καθοδήγηση (UNESCO, 2022; United Nations, 2024).

7

Σύνθεση και Συμπεράσματα

7.1 Ανακεφαλαίωση των Βασικών Σημείων

Η παρούσα εργασία εστίασε στη συστηματική εξέταση της ηθικής διάστασης της Τεχνητής Νοημοσύνης (TN), ξεκινώντας από το ιστορικό και θεωρητικό υπόβαθρο που θεμελίωσε τη σύγχρονη συζήτηση και φτάνοντας έως τις αναδυόμενες τάσεις που προδιαγράφουν το μέλλον της. Μέσα από την ιστορική επισκόπηση, αναδείχθηκαν οι βασικοί σταθμοί στην ανάπτυξη συστημάτων TN: από τις θεωρητικές συζητήσεις του μεσοπολέμου και τη διατύπωση του Τεστ Τούρινγκ, ως τις πρώτες πρακτικές εφαρμογές και τη μετέπειτα εξέλιξη των αυτόνομων συστημάτων. Σε αυτό το πλαίσιο, η παρούσα εργασία ανέδειξε πώς η επανάσταση των δεδομένων (Big Data) και οι πρωτοποριακές τεχνικές βαθιάς μάθησης εκτόξευσαν τις ικανότητες των αλγορίθμων και, ταυτόχρονα, έφεραν στο προσκήνιο μια σειρά από ηθικά διλήμματα (Jobin et al., 2019).

Καθοριστικός άξονας της ανάλυσης υπήρξε η εννοιολόγηση και η πρακτική των ηθικών αρχών στην TN, όπως η διαφάνεια, η λογοδοσία, η δικαιοσύνη και ο σεβασμός στην ανθρώπινη αυτονομία. Τα ζητήματα διαφάνειας – ιδίως η έλλειψη επεξηγησιμότητας των αλγορίθμων «μαύρου κουτιού» – τονίστηκαν ως κεντρικά για τη διατήρηση της εμπιστοσύνης των χρηστών, αλλά και για τη δυνατότητα άσκησης ελέγχου σε περίπτωση μεροληπτικών ή βλαπτικών αποφάσεων (Birhane, 2021). Ένα ακόμη κρίσιμο σημείο αποτέλεσε η προβληματική της αλγοριθμικής προκατάληψης (bias): όταν ιστορικά άνισα δεδομένα ενσωματώνονται στα μοντέλα, αυτά αναπαράγουν διακρίσεις σε ευαίσθητους τομείς, όπως η απασχόληση, οι τραπεζικές χορηγήσεις ή η ποινική δικαιοσύνη (Morley et al., 2020). Οι επιπτώσεις τέτοιων μεροληψιών μπορεί να είναι εξαιρετικά σοβαρές, ενισχύοντας ανισότητες στις κοινωνικές δομές αντί να τις αμβλύνουν.

Στη συνέχεια, μέσα από συγκεκριμένες μελέτες περιπτώσεων, η εργασία έδειξε ότι οι ίδιες αρχές και προβληματισμοί αναφύονται σε πολλούς κλάδους: στην επιβολή του νόμου (predictive policing), όπου η TN κινδυνεύει να ενισχύσει διακρίσεις εις βάρος μειονοτήτων στα αυτόνομα όπλα, όπου τίθενται επιτακτικά ερωτήματα σχετικά με την απόφαση λήψης ζωής χωρίς ανθρώπινη παρέμβαση, στην ιατρική διάγνωση, όπου η αναποτελεσματική ή μεροληπτική αλγοριθμική πρόβλεψη μπορεί να θέσει σε κίνδυνο την υγεία των ασθενών και στη διαχείριση προσωπικών δεδομένων, όπου οι ανησυχίες για μαζική παρακολούθηση και αδιαφανή χρήση δεδομένων γίνονται

ολοένα εντονότερες (Brayne & Christin, 2020; Horowitz, 2019; Ngiam & Khor, 2019; Taddeo & Floridi, 2018). Κοινός παρονομαστής αναδεικνύεται η ανάγκη ισχυρών ρυθμιστικών και δεοντολογικών πλαισίων, που όχι μόνο θα αποτρέπουν καταχρήσεις, αλλά θα προωθούν τη δικαιοσύνη και τον σεβασμό των δικαιωμάτων.

Η συγκριτική ανάλυση ανέδειξε ότι τα συστήματα τεχνητής νοημοσύνης διαφέρουν ουσιαστικά από το κλασικό λογισμικό, καθώς βασίζονται σε δεδομένα και εμφανίζουν μη ντετερμινιστική συμπεριφορά. Η αξιοπιστία τους είναι δυναμική και απαιτεί συνεχή αξιολόγηση και επανεκπαίδευση, σε αντίθεση με τα παραδοσιακά συστήματα που στοχεύουν στην σταθερότητα. Παράλληλα, προκύπτουν σημαντικοί κίνδυνοι, όπως η μεροληψία και τα ζητήματα λογοδοσίας, γεγονός που ενισχύει περαιτέρω την ανάγκη για ισχυρά πλαίσια διακυβέρνησης και ηθικής εποπτείας.

Παράλληλα, επισημάνθηκε ότι η ηθική της TN δεν μπορεί να μείνει στάσιμη μπροστά στην ταχύτητα των τεχνολογικών εξελίξεων. Οι νέες τάσεις, όπως οι προηγμένοι αλγόριθμοι βαθιάς μάθησης και η «δημιουργική TN» (Generative AI), γεννούν περαιτέρω προκλήσεις για την αυθεντικότητα του ψηφιακού περιεχομένου, για τα πνευματικά δικαιώματα και για την παραπληροφόρηση (Chesney & Citron, 2019; Floridi & Chiriatti, 2020). Επίσης, η εδραίωση της TN σε “έξυπνες πόλεις” και σε συνδεδεμένες συσκευές (IoT) τονίζει την ανάγκη για προστασία της ιδιωτικότητας, θωράκιση της ασφάλειας υποδομών και διατήρηση του ανθρώπινου ελέγχου.

Τέλος, αναδείχθηκε η σημασία των πρωτοβουλιών για ρύθμιση και αυτορρύθμιση. Σε διεθνές επίπεδο, θεσμοί όπως η ΕΕ, ο ΟΟΣΑ και η UNESCO προωθούν κατευθυντήριες για αξιόπιστη και υπεύθυνη TN, ενώ κορυφαίες εταιρείες τεχνολογίας υιοθετούν ηθικές αρχές και αναπτύσσουν εργαλεία (Explainable AI, bias mitigation) που διευκολύνουν τη δίκαιη υλοποίηση των αλγορίθμων (Eitel-Porter, 2020). Όπως έγινε σαφές, όμως, τα νομικά πλαίσια πρέπει να συνοδεύονται από τη διαρκή συμμετοχή και τη συνεργασία όλων των ενδιαφερομένων – επιστημόνων, πολιτικών, επαγγελματιών, αλλά και της κοινωνίας των πολιτών.

Συνοψίζοντας, η εργασία κατέδειξε ότι η TN φέρνει αναμφισβήτητα οφέλη στην καινοτομία, εντούτοις ενέχει πολλαπλές ηθικές προκλήσεις. Η ισορροπία ανάμεσα στην τεχνολογική πρόοδο και την προστασία θεμελιωδών αξιών, όπως η δικαιοσύνη και τα ανθρώπινα δικαιώματα, είναι ένας δυναμικός στόχος που απαιτεί συνεχή εγρήγορση. Οι προβληματισμοί για διαφάνεια, λογοδοσία, προκατάληψη και ιδιωτικότητα απασχολούν κάθε πεδίο εφαρμογής της TN και αναμένεται να γίνουν εντονότεροι όσο η τεχνολογία ωριμάζει. Επομένως, η ρύθμιση της TN πρέπει να συνδυάζει νομικά, ηθικά και τεχνολογικά εργαλεία, οικοδομώντας ένα πλέγμα συνεργασιών που θα

εξασφαλίζει ότι η TN θα συμβάλλει θετικά στο μέλλον της ανθρωπότητας, χωρίς να υπονομεύει τις δημοκρατικές αρχές και τον ανθρώπινο σεβασμό.

7.2 Συμπεράσματα για τη Σημασία της Ηθικής στην Ανάπτυξη της Τεχνητής Νοημοσύνης

7.2.1 Αποφυγή Βλαβών και Άδικων Αποτελεσμάτων

Η σημασία της ηθικής στη διαδικασία ανάπτυξης και εφαρμογής της Τεχνητής Νοημοσύνης (TN) έχει καταστεί εμφανής μέσα από τα συμπεράσματα που προέκυψαν σε όλα τα προηγούμενα κεφάλαια. Όπως δείχνει η ερευνητική βιβλιογραφία και οι περιπτώσιολογικές μελέτες, η ενσωμάτωση ηθικών αρχών δεν συνιστά απλώς ένα θεωρητικό ιδεώδες, αλλά αποτελεί βασική προϋπόθεση για την πρόληψη βλαβών, τη ενίσχυση και διατήρηση της εμπιστοσύνης του κοινού και την προώθηση του δημόσιου συμφέροντος και του κοινού καλού. Από τους αλγορίθμους προληπτικής αστυνόμευσης έως τις εφαρμογές στην ιατρική διάγνωση, η TN επηρεάζει άμεσα ή έμμεσα την καθημερινότητα, δημιουργώντας ένα πεδίο ηθικών προκλήσεων που απαιτούν συστηματική αντιμετώπιση.

Αρχικά, η ηθική διάσταση της TN είναι ουσιώδης για την αποτροπή άδικων ή επιβλαβών αποτελεσμάτων. Τα παραδείγματα αλγοριθμικής μεροληψίας έχουν αναδείξει πόσο εύκολα μπορούν να διαιωνιστούν προϋπάρχουσες κοινωνικές διακρίσεις μέσω της επεξεργασίας δεδομένων. Εάν τα δεδομένα εκπαίδευσης ενός συστήματος αναπαράγουν στερεότυπα ή δομικές ανισότητες, οι αποφάσεις που λαμβάνει το σύστημα ενδέχεται να βλάπτουν ορισμένες κοινωνικές ομάδες, εμποδίζοντάς τες να έχουν δίκαιη πρόσβαση σε εργασία, ασφάλιση ή δικαιοσύνη. Από την άλλη πλευρά, η υγεία των ασθενών μπορεί να τεθεί σε κίνδυνο όταν η TN σε συστήματα ιατρικής διάγνωσης λειτουργεί με περιορισμένα ή ελλιπή δεδομένα, ή όταν οι υπεύθυνοι δεν έχουν μεριμνήσει για ελέγχους ποιότητας και εξακρίβωσης. Σε όλα αυτά, η παρουσία ηθικού ελέγχου, η διασφάλιση ότι τα σύνολα δεδομένων είναι αντιπροσωπευτικά και η απόδοση ευθύνης για τις αποφάσεις ενός αλγορίθμου, αποτελούν ζωτικά βήματα για να αποφευχθούν τέτοιου είδους βλάβες. Η εγκαθίδρυση μηχανισμών λογοδοσίας (accountability), που απαιτούν από τους σχεδιαστές και τους αναπτυξιακούς φορείς της TN να τεκμηριώνουν και να αιτιολογούν τις αποφάσεις των συστημάτων, συμβάλλει στην πρόληψη άδικων αποτελεσμάτων και στην προστασία των πολιτών από πιθανή βλάβη.

7.2.2 Ενίσχυση και διατήρηση της εμπιστοσύνης του κοινού

Δεύτερον, η ενίσχυση και διατήρηση της εμπιστοσύνης του κοινού αποδεικνύεται καθοριστική για την ευδοκίμηση των τεχνολογιών ΤΝ. Εάν η κοινωνία θεωρήσει τα συστήματα αλγοριθμικής λήψης αποφάσεων αδιαφανή, μεροληπτικά ή επικίνδυνα για την ιδιωτικότητα, είναι πιθανό να γεννηθεί αντίδραση, καινοτομία να επιβραδυνθεί και ενδεχομένως να αναπτυχθεί ένα κλίμα καχυποψίας. Επιπλέον, η διαφάνεια και η επεξηγησιμότητα των αλγορίθμων δεν αποτελούν μόνο ηθικά ζητούμενα· λειτουργούν και ως μηχανισμοί ενίσχυσης της εμπιστοσύνης των τελικών χρηστών. Μια ΤΝ που παρέχει, για παράδειγμα, επεξηγήσεις για το πώς έφτασε σε ένα ιατρικό συμπέρασμα ή σε μια απόφαση χρηματοδότησης, αυξάνει την πιθανότητα αποδοχής της από το κοινό. Αντιθέτως, όσο περισσότερο ενισχύεται η εικόνα ενός «μαύρου κουτιού» που παράγει αποτελέσματα χωρίς διαφανή λογική, τόσο αυξάνεται ο κίνδυνος αρνητικής διάθεσης εκ μέρους των πολιτών. Με άλλα λόγια, η κερδισμένη εμπιστοσύνη λειτουργεί ως πολλαπλασιαστής αξιοποίησης της τεχνολογίας, δημιουργώντας ένα περιβάλλον όπου οι άνθρωποι μπορούν να βασίζονται στην ΤΝ για κρίσιμες αποφάσεις, γνωρίζοντας ότι τηρούνται κάποιοι ηθικοί και ποιοτικοί κανόνες.

7.2.3 Προώθηση του δημόσιου συμφέροντος και του κοινού καλού

Η προώθηση του δημόσιου συμφέροντος και του κοινού καλού αναδεικνύεται ως ο υπέρτατος σκοπός που επιτρέπει στην ΤΝ να λειτουργήσει ως δύναμη θετικής αλλαγής. Η ΤΝ προσφέρει τη δυνατότητα αποτελεσματικότερης διαχείρισης πόρων, βελτίωσης της ποιότητας ζωής και ενίσχυσης της καινοτομίας. Όταν όμως παραβλέπονται οι ηθικές αρχές, τα οφέλη αυτά υπονομεύονται και διολισθαίνουν σε καταστάσεις βλάβης ή αδικίας. Για παράδειγμα, η ιδιωτικότητα δεν αποτελεί απλώς «πολυτέλεια»: η αλόγιστη συλλογή και επεξεργασία δεδομένων ενδέχεται να απειλήσει βασικές ελευθερίες, υπονομεύοντας έτσι την κοινωνική συνοχή. Αντιθέτως, όταν η ηθική ενσωματώνεται σε όλα τα στάδια σχεδιασμού και χρήσης ενός συστήματος ΤΝ, τότε η τεχνολογία αυτή μπορεί να συμβάλει στην κοινωνική πρόοδο λ.χ. μέσα από δίκαια και αξιόπιστα συστήματα, όπως μια πλατφόρμα υγείας που προστατεύει τα ευαίσθητα δεδομένα των ασθενών, προσφέροντας ταυτόχρονα ακριβείς και έγκαιρες διαγνώσεις. Σε αυτό το πλαίσιο, η ηθική δεν αντιμετωπίζεται ως «εμπόδιο» στην ανάπτυξη, αλλά ως καταλύτης που εξασφαλίζει τη βιωσιμότητα και τη μακροπρόθεσμη αποδοχή της ΤΝ.

Αυτά τα συμπεράσματα υπαγορεύουν ότι η ηθική διάσταση πρέπει να γίνεται αντιληπτή ως αναπόσπαστο στοιχείο του κύκλου ζωής της ΤΝ, ξεκινώντας από τον αρχικό σχεδιασμό και συνεχίζοντας σε όλα τα στάδια υλοποίησης, δοκιμής, διάθεσης και χρήσης. Η προσέγγιση του «Ethics-By-Design» που έχει αναδειχθεί στη σχετική βιβλιογραφία, επιμένει στη συμμετοχή

ηθικολόγων, κοινωνιολόγων και νομικών ήδη από την αρχική φάση ανάπτυξης ενός συστήματος, αντί να αναζητούνται λύσεις εκ των υστέρων, όταν τα προβλήματα έχουν ενσωματωθεί στον πυρήνα της τεχνολογίας. Με αυτόν τον τρόπο, οι ηθικές αρχές μετατρέπονται σε λειτουργικές προδιαγραφές λόγου χάρη, ο σχεδιασμός μιας βάσης δεδομένων που προφυλάσσει επαρκώς την ανωνυμία των χρηστών ή η υλοποίηση αλγορίθμων μηχανικής μάθησης που περιορίζουν την προκατάληψη μέσω τακτικών ελέγχων και ρυθμίσεων.

Περαιτέρω, η διάσταση της ηθικής ευθύνης δεν μπορεί να βαρύνει αποκλειστικά τους προγραμματιστές ή τις εταιρείες τεχνολογίας. Απαιτείται πολυεπίπεδη συνεργασία και ανοιχτός διάλογος που θα περιλαμβάνει κυβερνήσεις, ερευνητικά ιδρύματα, ΜΚΟ, ομάδες πολιτών και διεθνείς οργανισμούς. Οι συζητήσεις σε επίπεδο Ευρωπαϊκής Ένωσης ή Ηνωμένων Εθνών για τη ρύθμιση της ΤΝ, οι κατευθυντήριες γραμμές για «έμπιστη» ΤΝ (Trustworthy AI) και οι συστάσεις σχετικά με τα ανθρώπινα δικαιώματα καταδεικνύουν ότι, χωρίς ένα νομικό και ηθικό πλαίσιο, ακόμα και οι πιο πολύτιμες τεχνολογικές καινοτομίες μπορεί να αποτύχουν ή και να βλάψουν. Η κοινωνία οφείλει να απαιτήσει ηθική εγρήγορση, επειδή οι αποφάσεις των αλγορίθμων όλο και συχνότερα αγγίζουν ευαίσθητες σφαίρες της ζωής: από τη χρηματοδότηση και την εργασία, μέχρι την επιβολή του νόμου και τη δημόσια υγεία.

Συνοψίζοντας, η ηθική δεν είναι κάτι που απλώς «προστίθεται» στο τέλος της διαδικασίας, ούτε μια προαιρετική μέριμνα για τις εταιρείες που επιθυμούν να προβάλλουν κοινωνική υπευθυνότητα. Αντιθέτως, είναι η ραχοκοκαλιά μιας βιώσιμης και αξιόπιστης ΤΝ, που μπορεί να προωθήσει ταυτόχρονα την καινοτομία και τον σεβασμό των ανθρωπίνων αξιών. Είτε πρόκειται για την αποφυγή μεροληψίας, την προστασία της ιδιωτικότητας, τη διατήρηση της εμπιστοσύνης ή τη διευκόλυνση του κοινού καλού, τα ευρήματα της έρευνας αναδεικνύουν ότι η ηθική θωράκιση κάθε βήματος στον κύκλο ζωής της ΤΝ είναι προαπαιτούμενο επιτυχίας. Μόνο υπό αυτές τις συνθήκες, η ΤΝ θα διατηρήσει τα χαρακτηριστικά κοινωνικής αποδοχής, δίκαιης λειτουργίας και αξιόπιστης προσφοράς, επιβεβαιώνοντας τον ρόλο της ως τεχνολογίας που υπηρετεί το ανθρώπινο όφελος και όχι το αντίστροφο.

7.3 Τελικές Παρατηρήσεις και Προτάσεις για Υπεύθυνη

Αξιοποίηση της Τεχνητής Νοημοσύνης

Μέσα από όλα τα συμπεράσματα που αναδείχθηκαν, καθίσταται σαφές ότι η υπεύθυνη αξιοποίηση της Τεχνητής Νοημοσύνης (ΤΝ) απαιτεί μια ολιστική προσέγγιση, η οποία δεν θα περιορίζεται αποκλειστικά σε τεχνικές βελτιστοποιήσεις ή νομικές ρυθμίσεις. Αντιθέτως, προϋποθέτει τη διαρκή εγρήγορση όλων των εμπλεκόμενων φορέων: από τους πολιτικούς και τις ρυθμιστικές αρχές έως

τους δημιουργούς της τεχνολογίας, τους ερευνητές, αλλά και το ευρύ κοινό που ως τελικός χρήστης έχει κεντρικό ρόλο στη διαμόρφωση των κοινωνικών προτεραιοτήτων. Στο πλαίσιο αυτό, ορισμένες συγκεκριμένες προτάσεις μπορούν να αποτελέσουν πυξίδα για ένα μέλλον όπου η ΤΝ θα ενισχύει την ανθρώπινη ευημερία χωρίς να υπονομεύει θεμελιώδεις ηθικές αξίες.

Οι φορείς λήψης αποφάσεων (υπουργεία, ανεξάρτητες αρχές, διεθνείς οργανισμοί) καλούνται να θεσπίσουν διαδικασίες ελέγχου και αξιολόγησης της ηθικής διάστασης κάθε νέας εφαρμογής ΤΝ. Μία πρόταση είναι η δημιουργία επιτροπών ηθικής αξιολόγησης, παρόμοιων με εκείνες που επιβλέπουν βιοηθικά ζητήματα στην ιατρική έρευνα. Αυτές οι επιτροπές θα μπορούν να αξιολογούν αν τα σύνολα δεδομένων είναι επαρκώς αντιπροσωπευτικά και απαλλαγμένα από μεροληψία, αν οι αλγόριθμοι διαθέτουν μηχανισμούς διαφάνειας και λογοδοσίας και αν έχει προηγηθεί διαβούλευση με τις ομάδες πολιτών που μπορεί να επηρεαστούν. Επιπλέον, θα ήταν ωφέλιμη η υιοθέτηση προτύπων πιστοποίησης ηθικής τεχνολογίας, αντίστοιχων με τα υπάρχοντα πρότυπα ποιότητας (π.χ. ISO), τα οποία να περιλαμβάνουν συγκεκριμένα κριτήρια για τη φάση σχεδιασμού, δοκιμής και διάθεσης στην αγορά.

Η βιομηχανία της τεχνολογίας, από την άλλη πλευρά, μπορεί να ενισχύσει σημαντικά τις ηθικές προδιαγραφές των προϊόντων της μέσα από την ενσωμάτωση checks and balances σε κάθε στάδιο ανάπτυξης. Αυτό σημαίνει ότι οι ομάδες σχεδιασμού των αλγορίθμων οφείλουν να εφαρμόζουν από νωρίς τεχνικές mitigation για αλγοριθμική μεροληψία, να αναπτύσσουν μηχανισμούς επεξηγησιμότητας και να θέτουν ασφαλιστικές δικλίδες σε περίπτωση που το σύστημα πάρει μια λανθασμένη ή επιβλαβή απόφαση. Στο εσωτερικό των εταιρειών, θα μπορούσαν να δημιουργηθούν διεπιστημονικές ομάδες αποτελούμενες από μηχανικούς λογισμικού, νομικούς, κοινωνιολόγους και ηθικολόγους που θα συνεργάζονται συστηματικά στην ανάπτυξη νέων προϊόντων ΤΝ, ώστε να εντοπίζονται έγκαιρα οι πιθανές ηθικές παγίδες.

Έπειτα, αναδεικνύεται επιτακτικά η ανάγκη για ηθική παιδεία και ευαισθητοποίηση σε όλα τα επίπεδα. Οι μηχανικοί λογισμικού, οι αναλυτές δεδομένων και οι επιστήμονες πληροφορικής χρειάζεται να έχουν εμπεριστατωμένη γνώση των ηθικών πτυχών της ΤΝ, όπως και των αρχών προστασίας προσωπικών δεδομένων ή αντιμετώπισης της αλγοριθμικής προκατάληψης. Η εκπαίδευση αυτή θα μπορούσε να γίνεται είτε στο πλαίσιο προγραμμάτων πανεπιστημιακών σπουδών είτε μέσα από πιστοποιημένα σεμινάρια συνεχιζόμενης επαγγελματικής κατάρτισης. Αντίστοιχα, οι τελικοί χρήστες από ιδιώτες έως μικρές επιχειρήσεις και δημόσιους φορείς – θα πρέπει να πληροφορούνται πώς λειτουργούν οι βασικοί αλγόριθμοι, ποια δικαιώματα έχουν και τι κινδύνους εγκυμονεί η ανεπαρκής ρύθμιση. Μια πιο ενημερωμένη κοινωνία θα μπορεί να απαιτεί

και να εκτιμά τα ηθικά χαρακτηριστικά των συστημάτων ΤΝ, πιέζοντας έτσι για διαφάνεια και λογοδοσία.

Τέλος, είναι σημαντικό να αναγνωριστεί ότι η εξέλιξη της ΤΝ και της ηθικής είναι μια δυναμική διαδικασία, που δεν επιδέχεται οριστικές και αιώνιες λύσεις. Νέα μοντέλα μηχανικής μάθησης και αναδυόμενες τεχνολογίες, όπως η Γενική Τεχνητή Νοημοσύνη (AGI) ή τα προηγμένα δημιουργικά συστήματα (Generative AI), φέρνουν στο προσκήνιο φρέσκα ερωτήματα και προκλήσεις. Χρειάζεται λοιπόν συνεχής παρακολούθηση των εξελίξεων και συνεργασία μεταξύ ερευνητών, πολιτικών και κοινωνικών εταίρων, ώστε η ηθική πλαισίωση να επικαιροποιείται και να αντανακλά τις πραγματικές συνθήκες κάθε εποχής. Ο ανοιχτός δημόσιος διάλογος σε εθνικό και διεθνές επίπεδο είναι καίριος: μόνον όταν οι πολίτες έχουν φωνή στις αποφάσεις που αφορούν την ΤΝ, μπορούν να διασφαλιστούν οι αξίες της κοινωνικής δικαιοσύνης και του σεβασμού των ανθρωπίνων δικαιωμάτων.

Συνοψίζοντας, η υπεύθυνη αξιοποίηση της ΤΝ δεν μπορεί να προσεγγιστεί ως ένας τελικός προορισμός, αλλά ως μια συνεχιζόμενη δέσμευση προσαρμογής και εγρήγορσης. Απαιτεί βαθύ σεβασμό των ηθικών αρχών, εφαρμογή ποιοτικών και πιστοποιημένων στάνταρντ, έμπρακτη συνεργασία μεταξύ δημόσιων και ιδιωτικών φορέων, καθώς και ουσιαστική συμμετοχή των πολιτών στον διάλογο. Αν υιοθετηθεί αυτό το μοντέλο, η ΤΝ μπορεί να εδραιωθεί ως ένα πραγματικά καινοτόμο και ταυτόχρονα ανθρώπινο εργαλείο, που θα βελτιώνει τη ζωή μας και θα ενισχύει την κοινωνική πρόοδο, αντί να επιτείνει τις υπάρχουσες ανισότητες.

Βιβλιογραφία

- Bergold, A. N., & Kovera, M. B. (2024). *The contribution of facial recognition technology to wrongful arrests and trauma. Psychological Trauma: Theory, Research, Practice, and Policy*. Advance online publication. <https://doi.org/10.1037/tra0001783>
- Birhane, A. (2021). *Algorithmic injustice: a relational ethics approach*. Patterns (New York, N.Y.), 2(2), 100205. <https://doi.org/10.1016/j.patter.2021.100205>
- Brayne, S. & Christin, A. (2020). *Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts*. Social Problems. 68. 10.1093/socpro/spaa004.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). “Semantics derived automatically from language corpora contain human-like biases.” Science, 356(6334), 183-186.
- Chesney, R., & Citron, D. (2019). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. California Law Review, 107, 1753-1819. <https://doi.org/10.2139/ssrn.3213954>
- Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.
- CONSORT-AI and SPIRIT-AI Steering Group (2019). *Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed*. Nature medicine, 25(10), 1467–1468. <https://doi.org/10.1038/s41591-019-0603-3>
- Crawford, K., & Calo, R. (2016). “There is a blind spot in AI research.” Nature, 538, 311-313.
- de Neufville, R., & Baum, S. D. (2021). *Collective action on artificial intelligence: A primer and review*. Technology in Society, 66, 101649. <https://doi.org/10.1016/j.techsoc.2021.101649>
- Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer.
- Eitel-Porter, R. (2020). *Beyond the promise: implementing ethical AI*. AI and Ethics. 1. 1-8. 10.1007/s43681-020-00011-6.
- El-Deeb, S., Jahankhani, H., Hussien, O., Hussien, M., & Sandakelum, I. & Will Arachchige, I. S. (2024). *Towards Responsible AI: Exploring AI Frameworks, Ethical Dimensions and Regulations*. 10.1108/978-1-83549-001-320241007.
- Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.
- Floridi, L. & Chiriatti, M. (2020). *GPT-3: Its Nature, Scope, Limits, and Consequences*. Minds and Machines. 30. 1-14. 10.1007/s11023-020-09548-1.

- Floridi, L. & Cows, J. (2019). *A Unified Framework of Five Principles for AI in Society*. Harvard Data Science Review. 10.1162/99608f92.8cd550d1.
- Hagendorff, T. (2020). *The Ethics of AI Ethics: An Evaluation of Guidelines*. Minds & Machines 30, 99–120 (2020). <https://doi.org/10.1007/s11023-020-09517-8>
- Horowitz, M. (2019). *When speed kills: Lethal autonomous weapon systems, deterrence and stability*. Journal of Strategic Studies. 42. 764-788. 10.1080/01402390.2019.1621174.
- Jobin, A., Ienca, M., & Vayena, E. (2019). “*The global landscape of AI ethics guidelines*.” Nature Machine Intelligence, 1(9), 389-399. 1. 10.1038/s42256-019-0088-2.
- Joh, E. E., (2018). *Artificial Intelligence and Policing: Hints in the Carpenter Decision* (August 24, 2018). __ Ohio State Journal of Criminal Law __, 2018, Available at SSRN: <https://ssrn.com/abstract=3238212>
- Kalisa, M. & Extension, Kiu Publication. (2024). *Ethical Considerations in the Development and Implementation of Artificial Intelligence for Autonomous Vehicles*. 3. 68-73.
- Kitchin R. (2016). *The ethics of smart cities and urban science*. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 374(2083), 20160115. <https://doi.org/10.1098/rsta.2016.0115>
- Li, Shancang & Tryfonas, Theo & Li, Honglei. (2016). *The Internet of Things: a security point of view*. Internet Research. 26. 337-359. 10.1108/IntR-07-2014-0173.
- Matthias, A. (2004). “*The responsibility gap: Ascribing responsibility for the actions of learning automata*.” Ethics and Information Technology, 6(3), 175-183.
- Miller, T. (2017). *Explanation in Artificial Intelligence: Insights from the Social Sciences*. Artificial Intelligence. 267. 10.1016/j.artint.2018.07.007.
- Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). *From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices*. Science and engineering ethics, 26(4), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
- Ngiam, K. Y., & Khor, I. W. (2019). *Big data and machine learning algorithms for health-care delivery*. The Lancet. Oncology, 20(5), e262–e273. [https://doi.org/10.1016/S1470-2045\(19\)30149-4](https://doi.org/10.1016/S1470-2045(19)30149-4)
- OECD. (2019). *OECD Principles on AI*. OECD Publishing. <https://www.oecd.org/sti/ieconomy/ai/>
- O’Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books.
- Richardson, R., Schultz, J. & Crawford, K. (2019). *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice* (February 13,

- 2019). 94 N.Y.U. L. REV. ONLINE 192 (2019), Available at SSRN: <https://ssrn.com/abstract=3333423>
- Roff, H. M. (2019). *The frame problem: The AI “arms race” isn’t one*. Bulletin of the Atomic Scientists, 75(3), 95–98.
- Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall.
- Shilton, K. (2017). *Engaging Values Despite Neutrality: Challenges and Approaches to Values Reflection during the Design of Internet Infrastructure*. Science, Technology, & Human Values. 43. 016224391771486. 10.1177/0162243917714869.
- Sullins, J. P. (2012). “Robots, love, and sex: The ethics of building a love machine.” IEEE Transactions on Affective Computing, 3(4), 398-409.
- Taddeo, M. & Floridi, L. (2018). *How AI can be a force for good*. Science. 361. 751–752. 10.1126/science.aat5991.
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000380455>
- 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. 205630512090340. 10.1177/2056305120903408.
- Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.
- Wirtz, B., Weyerer, J. & Geyer, C. (2018). *Artificial Intelligence and the Public Sector—Applications and Challenges*. International Journal of Public Administration. 42. 10.1080/01900692.2018.1498103.
- Xia, Q., Zheng, H., Zou, H., Luo, D., Tang, H., Li, L., & Jiang, B. (2025). *A comprehensive review of deep learning for medical image segmentation*. Neurocomputing, 613, 128740. <https://doi.org/10.1016/j.neucom.2024.128740>
- Dwivedi, R., & Elluri, L. (2024). Exploring generative artificial intelligence research: A bibliometric analysis approach. *IEEE Access*, 12, 119884–119902. <https://doi.org/10.1109/ACCESS.2024.3450629>
- Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). *Sociotechnical safety evaluation of generative AI systems*. arXiv. <https://arxiv.org/abs/2310.11986>
- UNESCO. (2023). *Ethical impact assessment: A tool of the Recommendation on the Ethics of Artificial Intelligence*. UNESCO.
- Fosso Wamba, S., Bawack, R. E., Guthrie, C., Queiroz, M. M., & Carillo, K. D. A. (2021). Are we preparing for a good AI society? A bibliometric review and research agenda. *Technological Forecasting and Social Change*, 164, 120482. <https://doi.org/10.1016/j.techfore.2020.120482>
- Masle, N. (2024). *The AI Index 2024 annual report*. Stanford University, AI Index Steering Committee.
- Juric, M., Sandic, A., & Brcic, M. (2020). AI safety: State of the field through a quantitative lens. In *Proceedings of the 43rd International Convention on Information, Communication and Electronic Technology* (pp. 1254–1259).

- UN System Chief Executives Board for Coordination. (2024). *Summary of deliberations: Chief Executives Board for Coordination, 1st regular session of 2024 (Addendum)*. United Nations. <https://digitallibrary.un.org/record/4064466>
- Mäntymäki, M., Minkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics*, 2(4), 603–609. <https://doi.org/10.1007/s43681-022-00143-x>
- United Nations. (2024). *Governing AI for humanity*. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf
- UNESCO. (2022). *Recommendation on the Ethics of Artificial Intelligence*. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). <https://doi.org/10.6028/NIST.AI.100-1>
- Khan, A., Choudhury, N., Uddin, S., Hossain, L., & Baur, L. A. (2016). Longitudinal trends in global obesity research and collaboration: A bibliometric review. *Obesity Reviews*, 17(4), 377–385.
- Demiroz, F., & Haase, T. W. (2020). The concept of resilience: A bibliometric analysis of the emergency and disaster management literature. In *Local disaster management* (pp. 16–35). Routledge. <https://doi.org/10.4324/9781003036234-3>
- Laengle, S., Merigó, J. M., Miranda, J., Słowiński, R., Bomze, I., Borgonovo, E., Dyson, R. G., Oliveira, J. F., & Teunter, R. (2017). Forty years of the European Journal of Operational Research: A bibliometric overview. *European Journal of Operational Research*, 262(3), 803–816. <https://doi.org/10.1016/j.ejor.2017.04.027>
- Mingers, J., & Burrell, Q. L. (2006). Modeling citation behavior in management science journals. *Information Processing & Management*, 42(6), 1451–1464. <https://doi.org/10.1016/j.ipm.2006.03.012>
- Rosas, S. R., Kagan, J. M., Schouten, J. T., Slack, P. A., & Trochim, W. M. K. (2011). Evaluating research and impact: A bibliometric analysis of research by the NIH/NIAID HIV/AIDS clinical trials networks. *PLoS ONE*, 6(3), e17428.
- Comunale, M., & Manera, A. (2024). *The economic impacts and the regulation of AI: A review of the academic literature and policy actions*. International Monetary Fund. <https://www.imf.org/en/Publications/WP/Issues/2024/03/22/The-Economic-Impacts-and-the-Regulation-of-AI-A-Review-of-the-Academic-Literature-and-546645>
- Wasserman, S., & Galaskiewicz, J. (1997). *Advances in social network analysis: Research in the social and behavioral sciences*. Sage.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (Vol. 3, pp. 361–362).
- Oldham, S., Fulcher, B., Parkes, L., Arnatkevičiūtė, A., Suo, C., & Fornito, A. (2019). Consistency and differences between centrality measures across distinct classes of networks. *PLoS ONE*, 14(7), e0220061.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- Willis, A., Fisher, A., & Lvov, I. M. (2015). Mapping networks of influence: Tracking Twitter conversations through time and space. *Participations: Journal of Audience and Reception Studies*, 494–530.
- Afzaal, M., Shanshan, X., Yan, D., & Younas, M. (2024). Mapping artificial intelligence integration in education: A decade of innovation and impact (2013–2023)—A bibliometric analysis. *IEEE Access*, 12, 113275–113299. <https://doi.org/10.1109/ACCESS.2024.3443313>
- McStay, A., & Rosner, G. (2021). Emotional artificial intelligence in children’s toys and devices: Ethics, governance and practical remedies. *Big Data & Society*, 8(1). <https://doi.org/10.1177/2053951721994877>
- Yew, G. C. K. (2021). Trust in and ethical design of carebots: The case for ethics of care. *International Journal of Social Robotics*, 13(4), 629–645. <https://doi.org/10.1007/s12369-020-00653-w>

- Jordan, A., & Natarajan, N. (2024). From TV to social media to ambient AI: Insights from 30 years of children's media policy in the United States. *Journal of Children and Media*, 18(3), 378–385. <https://doi.org/10.1080/17482798.2024.2345530>
- Lee, L., Salami, R. K., Martin, H., Shantharam, L., Thomas, K., Ashworth, E., Allan, E., Yung, K.-W., Pauling, C., Leyden, D., Arthurs, O. J., & Sheldermine, S. C. (2024). “How I would like AI used for my imaging”: Children and young persons’ perspectives. *European Radiology*, 34(12), 7751–7764. <https://doi.org/10.1007/s00330-024-10839-9>
- Mitsea, E., Drigas, A., & Skianis, C. (2023). Digitally assisted mindfulness in training self-regulation skills for sustainable mental health: A systematic review. *Behavioral Sciences*, 13(12), 1008. <https://doi.org/10.3390/bs13121008>
- Su, J. (2024). Kindergarten teachers’ perceptions of AI literacy education for young children. *International Journal of Technology and Design Education*, 34(5), 1665–1685. <https://doi.org/10.1007/s10798-024-09876-8>
- Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence: The International Journal of Research into New Media Technologies*, 27(1), 36–51. <https://doi.org/10.1177/1354856520923963>
- Portacolone, E., Halpern, J., Luxenberg, J., Harrison, K. L., & Covinsky, K. E. (2020). Ethical issues raised by the introduction of artificial companions to older adults with cognitive impairment. *Journal of Alzheimer's Disease*, 76(2), 445–455. <https://doi.org/10.3233/JAD-190952>
- Bjørlo, L., Moen, Ø., & Pasquine, M. (2021). The role of consumer autonomy in developing sustainable AI: A conceptual framework. *Sustainability*, 13(4), 2332. <https://doi.org/10.3390/su13042332>
- Rosenberg, L. (2023). The metaverse and conversational AI as a threat vector for targeted influence. In *Proceedings of the IEEE 13th Annual Computing and Communication Workshop and Conference* (pp. 504–510).
- Conca, S. D. (2023). The present looks nothing like the Jetsons: Deceptive design in virtual assistants and the protection of the rights of users. *Computer Law & Security Review*, 51, 105866. <https://doi.org/10.1016/j.clsr.2023.105866>
- Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). Characterizing manipulation from AI systems. Presented at *Equity, Access, Algorithms, Mechanisms, and Optimization*.
- Lu, S. (2022). Data privacy, human rights, and algorithmic opacity. *California Law Review*, 110, 2087–2147. <https://doi.org/10.15779/Z38804XM07>
- Yen, C.-P., & Hung, T.-W. (2021). Achieving equity with predictive policing algorithms: A social safety net perspective. *Science and Engineering Ethics*, 27(3), 36. <https://doi.org/10.1007/s11948-021-00312-x>
- Arnett, C. (2022). Black lives monitored. *UCLA Law Review*, 69, 1384–1435.
- Odilla, F. (2024). Unfairness in AI anti-corruption tools: Main drivers and consequences. *Minds and Machines*, 34(3), Article 28. <https://doi.org/10.1007/s11023-024-09688-8>
- Kurtpınar, E. Ö. (2024). Privacy's sky-high battle: The use of unmanned aircraft systems for law enforcement in the European Union. *Journal of Intelligent & Robotic Systems*, 110(3), Article 99. <https://doi.org/10.1007/s10846-024-02071-w>
- European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828. Official Journal of the European Union, L 1689.