



ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

**Ανασκόπηση βιβλιογραφίας στα ζητήματα ηθικής στην
τεχνητή νοημοσύνη:
Συγκριτική μελέτη στα έτη 2019 και 2024**

ΔΙΠΛΩΜΑΤΙΚΗ

της

Τσίμα Βασιλικής

Επιβλέπουσα: Βασιλική Διαμαντοπούλου, Επ. Καθηγήτρια

Μέλη εξεταστικής επιτροπής: Μαρία Καρύδα, Καθηγήτρια, Λίλιαν Μήτρου, Καθηγήτρια

Σάμος, Φεβρουάριος 2025

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πρόλογος και ευχαριστίες

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια μου κα. Διαμαντοπούλου Βασιλική για την πολύτιμη βοήθεια της.

© [2025]

της

Τσίμα Βασιλικής

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1	Εισαγωγή	1
1.1	Η Τεχνητή Νοημοσύνη στη Σύγχρονη Εποχή: Προκλήσεις και Προβληματισμοί.....	1
1.2	Αντικείμενο της διπλωματικής εργασίας.....	2
1.3	Μεθοδολογική Προσέγγιση	4
1.3.1	Πηγές δεδομένων	5
1.3.2	Κριτήρια επιλογής/αποκλεισμού	5
1.4	Δομή της διπλωματικής	6
2	Θεωρητικό Υπόβαθρο	7
2.1	Βασικές Έννοιες και Ορισμοί.....	7
2.1.1	Τεχνητή Νοημοσύνη	7
2.1.2	Ηθική της Τεχνητής Νοημοσύνης	8
2.2	Νομικό Πλαίσιο και Κανονισμοί.....	10
2.2.1	GDPR και Προστασία Δεδομένων	10
2.2.2	Ευρωπαϊκό Πλαίσιο για την ΤΝ	11
2.2.3	Πρότυπα και Πιστοποιήσεις	12
2.3	Εφαρμογές και Επιπτώσεις της ΤΝ	13
2.3.1	Τομείς Εφαρμογής.....	13
2.3.2	Κοινωνικές Επιπτώσεις.....	14
2.3.3	Προκλήσεις και Κίνδυνοι	15
3	Μεθοδολογία Έρευνας.....	17
3.1	Ερευνητικά ερωτήματα.....	17
3.2	Μέθοδος Συλλογής Δεδομένων	18
▪	Φάση 1 (2015-2019)	18
▪	Φάση 2 (2024)	21
3.3	Κριτήρια Επιλογής και Αποκλεισμού.....	23
3.4	Περιορισμοί και Δυσκολίες	24
3.4.1	Ζητήματα Αναζήτησης και Σχετικότητας	24
3.4.2	Τεχνικοί Περιορισμοί στη Διαχείριση Αναφορών.....	25
3.4.3	Ζητήματα Πρόσβασης σε Έγγραφα.....	25
3.4.4	Ζητήματα Πληρότητας Εγγράφων	26
3.4.5	Διαχείριση Δεδομένων	26
4	Ανάλυση και Αποτελέσματα	28

4.1	Ευρήματα Περιόδου 2015-2019	28
4.1.1	Διασφάλιση ηθικής χρήσης TN.....	28
4.1.1.1	Αλγοριθμική δικαιοσύνη και λογοδοσία	29
4.1.1.2	Τεχνικές προσεγγίσεις για ηθική TN.....	30
4.1.1.3	Ρυθμιστικό και κανονιστικό πλαίσιο	30
4.1.1.4	Προκλήσεις και κίνδυνοι.....	31
4.1.1.5	Ανθρωποκεντρική προσέγγιση	32
4.1.2	Προϋποθέσεις ασφαλούς χρήσης TN.....	33
4.1.2.1	Θεμελιώδεις αρχές υπεύθυνης TN	33
4.1.2.2	Μηχανισμοί ασφάλειας και προστασίας	33
4.1.2.3	Διαφάνεια και επεξηγησιμότητα συστημάτων	34
4.1.2.4	Προετοιμασία για το μέλλον της TN.....	35
4.1.3	Υπευθυνότητα και ηθικός σχεδιασμός	35
4.1.3.1	Ορισμός και πλαίσιο υπευθυνότητας TN	35
4.1.3.2	Δεοντολογικές παράμετροι και πληροφοριακή ηθική.....	36
4.1.3.3	Αυτοματοποιημένη λήψη αποφάσεων και εποπτεία	37
4.1.3.4	Αυτονομία και όρια υπευθυνότητας.....	38
4.1.3.5	Κοινωνικές και ηθικές επιπτώσεις.....	38
4.1.4	Ζητήματα ιδιωτικότητας.....	39
4.1.4.1	Προστασία δεδομένων και ιδιωτικότητας στην εποχή της TN	39
4.1.4.2	Ρυθμιστικό πλαίσιο και μηχανισμοί προστασίας	40
4.1.4.3	Συγκρότηση πολιτικών και τεχνικών λύσεων	41
4.1.4.4	Προκλήσεις και μελλοντικές κατευθύνσεις.....	41
4.1.5	Ανάγκη ηθικής καθοδήγησης.....	42
4.1.5.1	Εκπαίδευση και επαγγελματική κατάρτιση.....	42
4.1.5.2	Θεσμικά πλαίσια και πρότυπα.....	43
4.1.5.3	Πρακτικές προκλήσεις εφαρμογής.....	44
4.1.5.4	Συνεργασία και συντονισμός φορέων	45
4.2	Ευρήματα Περιόδου 2024	45
4.2.1	Εξελιζέται στην ηθική χρήση της TN.....	45
4.2.1.1	Αλγοριθμική δικαιοσύνη και λογοδοσία στα σύγχρονα συστήματα TN	46
4.2.1.2	Τεχνικές προσεγγίσεις για ηθική TN.....	47
4.2.1.3	Ρυθμιστικό και κανονιστικό πλαίσιο	47
4.2.1.4	Προκλήσεις και κίνδυνοι.....	48
4.2.1.5	Ανθρωποκεντρική προσέγγιση	48
4.2.2	Νέες προκλήσεις ασφάλειας	49

4.2.2.1	Θεμελιώδεις αρχές υπεύθυνης ΤΝ	49
4.2.2.2	Μηχανισμοί ασφάλειας και προστασίας	49
4.2.2.3	Διαφάνεια και επεξηγησιμότητα συστημάτων	50
4.2.2.4	Προετοιμασία για το μέλλον της ΤΝ.....	50
4.2.3	<i>Σύγχρονες προσεγγίσεις υπευθυνότητας</i>	51
4.2.3.1	Ορισμός και πλαίσιο υπευθυνότητας ΤΝ	51
4.2.3.2	Δεοντολογικές παράμετροι και πληροφοριακή ηθική.....	51
4.2.3.3	Αυτοματοποιημένη λήψη αποφάσεων και εποπτεία	52
4.2.3.4	Αυτονομία και όρια υπευθυνότητας.....	52
4.2.3.5	Κοινωνικές και ηθικές επιπτώσεις.....	53
4.2.4	<i>Προστασία προσωπικών δεδομένων</i>	53
4.2.4.1	Προστασία δεδομένων και ιδιωτικότητας στην εποχή των γλωσσικών μοντέλων και του Generative AI	53
4.2.4.2	Ρυθμιστικό πλαίσιο και μηχανισμοί προστασίας στην εποχή του Generative AI.....	54
4.2.4.3	Συγκρότηση πολιτικών και τεχνικών λύσεων στην εποχή του Generative AI.....	54
4.2.4.4	Προκλήσεις και μελλοντικές κατευθύνσεις στην εποχή των Γλωσσικών Μοντέλων.....	55
4.2.5	<i>Σύγχρονα πλαίσια ηθικής καθοδήγησης</i>	55
4.2.5.1	Εκπαίδευση και ηθική κατάρτιση στην εποχή του Generative AI	55
4.2.5.2	Θεσμικά πλαίσια και πρότυπα στην εποχή της Generative AI.....	56
4.2.5.3	Πρακτικές προκλήσεις στην εφαρμογή ηθικών αρχών στη Generative AI	56
4.2.5.4	Συνεργασία και συντονισμός φορέων στην μετα-LLM εποχή.....	56
4.3	<i>Συγκριτική Ανάλυση Ευρημάτων</i>	57
4.3.1	<i>Διαχρονική εξέλιξη των ηθικών ζητημάτων</i>	57
4.3.2	<i>Αλλαγές στο ρυθμιστικό πλαίσιο</i>	57
4.3.3	<i>Νέες τάσεις και προκλήσεις</i>	57
4.3.4	<i>Μελλοντικές κατευθύνσεις</i>	58
5	Συμπεράσματα	59
5.1	<i>Συμπεράσματα Περιόδου 2015-2019</i>	59
5.1.1	<i>Σύνοψη ευρημάτων</i>	59
5.1.2	<i>Απαντήσεις στα ερευνητικά ερωτήματα</i>	60
5.1.3	<i>Κύρια συμπεράσματα και διαπιστώσεις</i>	61
5.2	<i>Συμπεράσματα Περιόδου 2024</i>	62
5.2.1	<i>Σύνοψη νέων ευρημάτων</i>	62
5.2.2	<i>Απαντήσεις στα ερευνητικά ερωτήματα υπό το πρίσμα των νέων δεδομένων</i>	62
5.3	<i>Συγκριτική Θεώρηση και Εξέλιξη του Πεδίου</i>	64
5.3.1	<i>Αλλαγές στην προσέγγιση της ηθικής διακυβέρνησης</i>	64
5.3.2	<i>Επίδραση του Generative AI στο πεδίο</i>	65

5.3.3	Ωρίμανση της ακαδημαϊκής και επιστημονικής κοινότητας.....	65
5.4	Προτάσεις για Μελλοντική Έρευνα	66
5.4.1	Αναδυόμενα ερευνητικά ερωτήματα.....	66
5.4.2	Προτεινόμενες κατευθύνσεις.....	67
5.4.3	Μεθοδολογικές προτάσεις.....	67
6	Βιβλιογραφία	69
6.1	Έρευνα 2015-2019.....	69
6.2	Έρευνα 2024.....	80

Λίστα Εικόνων

Εικόνα 1. Αναζήτηση όρων χωρίς αποστροφους.....	24
Εικόνα 2. Αναζήτηση όρων με αποστροφους για πιο συγκεκριμένα και λογικά λιγότερα σχετικά έγγραφα στα αποτελέσματα. Εμφάνιση εμφανώς περισσότερων μη-σχετικών αποτελεσμάτων.....	24
Εικόνα 3. Αποτελέσματα εννεηκοστής όγδοης σελίδας.	25
Εικόνα 4. Το συγκεκριμένο σφάλμα εμφανιζόταν με την επιλογή της εννεηκοστής ένατης σελίδας. Το σφάλμα εμφανιζόταν επανειλημμένα στην επιλογή της εννεηκοστής ένατης ή εκατοστής σελίδας (δοκιμάστηκε σε διαφορετικά χρονικά διαστήματα η αναζήτηση κατ' επανάληψη για να διαπιστωθεί ότι το σφάλμα ήταν μόνιμο), με αποτέλεσμα να σταματήσει σε αυτό το σημείο η αναζήτηση.	26
Εικόνα 5. Στοιχεία αναζήτησης μετά την ολοκλήρωση της φάσης.	27

Λίστα Πινάκων

Πίνακας 1. Περίληψη της έρευνας 2015-2019 σχετικά με τις στρατηγικές αναζήτησης ηθικής τεχνητής νοημοσύνης σε βάσεις δεδομένων. 21

Πίνακας 2. Περίληψη της έρευνας 2024 σχετικά με τις στρατηγικές αναζήτησης ηθικής τεχνητής νοημοσύνης σε βάσεις δεδομένων. 23

Ακρωνύμια

TN	Τεχνητή Νοημοσύνη
GDPR	General Data Protection Regulation
IVS	Intelligent Video Surveillance
ΟΟΣΑ	Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης
UNESCO	United Nations Educational, Scientific and Cultural Organization
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
PICOC	Population Intervention Comparison Outcomes Context
DPO	Data Protection Officer
ICO	Information Commissioner's Office
XAI	Explainable Artificial Intelligence
ML	Machine Learning
IEEE	Institute of Electrical and Electronics Engineers
CDIA	Consumer Data Industry Association
AI	Artificial Intelligence
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
LLM	Large Language Model
NFG	Normalized Fairness Gain
HF	Harmonic Fairness
NLP	Natural Language Processing
SDLC	Software Development Life Cycle
FIA	Feature Impact Analysis
I3	Integrated Invariance Interpretation
PCA/PIA	Principal Component Analysis
AILD	Artificial Intelligence Liability Directive
PII	Personally Identifiable Information
ZKP	Zero-Knowledge Proof
AIA	Artificial Intelligence Act
TLoD	Three Lines of Defense
RAISE	Responsible AI for Service Excellence
AIECT	AI-Enhanced Collaborative Theory

Πίνακας Ορολογίας

Ξενόγλωσσος όρος	Ελληνικός Όρος
General Data Protection Regulation	Γενικός Κανονισμός για την Προστασία των Δεδομένων
Envelopment	Περιβάλλουσα
Intelligent Video Surveillance	Ευφυής Βιντεο-επιτήρηση
Soft Law	Μη Δεσμευτικό Δίκαιο
International Organization for Standardization	Διεθνής Οργανισμός Τυποποίησης
International Electrotechnical Commission	Διεθνής Ηλεκτροτεχνική Επιτροπή
Population Intervention Comparison Outcomes Context	Πληθυσμός Παρέμβαση Σύγκριση Αποτελέσματα Συμπέρασμα
Explainable Artificial Intelligence	Επεξηγήσιμη Τεχνητή Νοημοσύνη
Machine Learning	Μηχανική Μάθηση
Generative AI	Παραγωγική ΤΝ
Generative Adversarial Network	Γενετικό Ανταγωνιστικό Δίκτυο
Variational Autoencoder	Μεταβλητός Αυτοκωδικοποιητής
Trustworthy AI	Αξιόπιστη ΤΝ
Large Language Model	Μεγάλο Γλωσσικό Μοντέλο
Normalized Fairness Gain	Κανονικοποιημένο Κέρδος Δικαιοσύνης
Harmonic Fairness	Αρμονική Δικαιοσύνη
Self-supervised	Αυτό-επιβλεπούμενη
Circuit Breakers	Διακόπτες Κυκλώματος
Hallucination	Ψευδαίσθηση
Natural Language Processing	Επεξεργασία Φυσικής Γλώσσας
Human-in-the-loop	Ενσωμάτωση του ανθρώπινου παράγοντα στον έλεγχο
Human-Centered Automation	Αυτοματοποίηση με Επίκεντρο τον Άνθρωπο
Software Development Life Cycle	Κύκλος Ζωής Ανάπτυξης Λογισμικού
Defense-in-Depth	Προσέγγιση Άμυνας σε Βάθος
Adversarial Training	Ανταγωνιστική Εκπαίδευση
Feature Impact Analysis	Ανάλυση Επίδρασης Χαρακτηριστικών
Principal Component Analysis	Ανάλυση Κύριων Συνιστωσών
Artificial Intelligence Liability Directive	Οδηγία περί Ευθύνης ΤΝ

Data Poisoning	Δηλητηρίαση Δεδομένων
Personally Identifiable Information	Προσωπικά Αναγνωρίσιμων Πληροφοριών
Jailbreaking	Επιθέσεις Παραβίασης
Differential Privacy	Διαφορική Ιδιωτικότητα
Replacement Encoding	Τεχνική κωδικοποίησης αντικατάστασης

Περίληψη

Η παρούσα διπλωματική εργασία διερευνά τα ηθικά ζητήματα που προκύπτουν από την ανάπτυξη και εφαρμογή της Τεχνητής Νοημοσύνης (TN), με ιδιαίτερη έμφαση στην εξέλιξή τους πριν και μετά την έλευση της Παραγωγικής TN. Μέσω μιας διττής συστηματικής βιβλιογραφικής ανασκόπησης που καλύπτει τις περιόδους έως και το 2019 και 2024, η εργασία εξετάζει πέντε βασικά ερευνητικά ερωτήματα σχετικά με τη διασφάλιση της ηθικής χρήσης της TN, τις προϋποθέσεις ασφαλούς εφαρμογής της, την υπευθυνότητα και τον ηθικό σχεδιασμό, την προστασία της ιδιωτικότητας, και την ανάγκη ηθικής καθοδήγησης.

Η έρευνα, που βασίστηκε στην ανάλυση περισσότερων από 9.000 επιστημονικών δημοσιεύσεων από δύο χρονικές περιόδους, αναδεικνύει τόσο τις διαχρονικές προκλήσεις όσο και τις νέες διαστάσεις που έχει προσθέσει η Παραγωγική TN στο πεδίο. Ενώ ζητήματα όπως η αλγοριθμική δικαιοσύνη και η λογοδοσία παραμένουν κρίσιμα, νέες προκλήσεις έχουν αναδυθεί, όπως η αυθεντικότητα του περιεχομένου που παράγεται από TN, η προστασία των πνευματικών δικαιωμάτων στην εποχή της μαζικής παραγωγής περιεχομένου, και η ανάγκη για διαφάνεια στη χρήση συστημάτων Παραγωγικής TN.

Η συγκριτική ανάλυση των δύο περιόδων αναδεικνύει τη σημαντική ωρίμανση του πεδίου της ηθικής της TN, με την ακαδημαϊκή κοινότητα να προτείνει πλέον πιο ολοκληρωμένα πλαίσια για την αντιμετώπιση των προκλήσεων. Ιδιαίτερη έμφαση δίνεται στην προστασία των προσωπικών δεδομένων και στην ανάγκη για ισχυρότερη ρύθμιση της χρήσης Παραγωγικής TN, ειδικά στο πλαίσιο της δημοκρατικοποίησης της τεχνολογίας.

Τα αποτελέσματα της έρευνας υπογραμμίζουν την ανάγκη για μια ολοκληρωμένη προσέγγιση στην ανάπτυξη και εφαρμογή της TN, που συνδυάζει τεχνικές λύσεις, κανονιστικά πλαίσια και ηθικές κατευθυντήριες γραμμές. Η εργασία προτείνει συγκεκριμένες κατευθύνσεις για μελλοντική έρευνα και τονίζει τη σημασία της διεθνούς συνεργασίας για την αντιμετώπιση των νέων προκλήσεων που θέτει η Παραγωγική TN στη σύγχρονη εποχή.

Λέξεις Κλειδιά: *τεχνητή νοημοσύνη, ηθική, GDPR, προστασία δεδομένων, αλγοριθμική δικαιοσύνη, ιδιωτικότητα, γενετική τεχνητή νοημοσύνη, μεγάλα γλωσσικά μοντέλα*

Abstract

This thesis investigates the ethical issues arising from the development and implementation of Artificial Intelligence (AI), with a particular focus on their evolution before and after the advent of Generative AI. Through a dual systematic literature review covering the periods up to 2019 and 2024, the study examines five key research questions concerning the assurance of ethical AI use, prerequisites for safe implementation, responsibility and ethical design, privacy protection, and the need for ethical guidance.

The research, based on the analysis of more than 9,000 scientific publications from two time periods, reveals both the enduring challenges and the new dimensions that Generative AI has added to the field. While issues such as algorithmic fairness and accountability remain critical, new challenges have emerged, such as the authenticity of AI-generated content, the protection of intellectual property rights in an era of mass content generation, and the need for transparency in the use of Generative AI systems.

The comparative analysis of the two periods highlights the significant maturation of the field of AI ethics, with the academic community now proposing more comprehensive frameworks to address these challenges. Particular emphasis is placed on personal data protection and the need for stronger regulation of Generative AI use, especially in the context of the democratization of technology.

The research findings underscore the need for a comprehensive approach to AI development and implementation that combines technical solutions, regulatory frameworks, and ethical guidelines. The thesis proposes specific directions for future research and emphasizes the importance of international collaboration in addressing the new challenges posed by Generative AI in the contemporary era.

Keywords: *artificial intelligence, ethics, GDPR, data protection, algorithmic fairness, privacy, generative AI, large language models*

1

Εισαγωγή

1.1 Η Τεχνητή Νοημοσύνη στη Σύγχρονη Εποχή: Προκλήσεις και Προβληματισμοί

Η ραγδαία ανάπτυξη της τεχνητής νοημοσύνης έχει φέρει επαναστατικές αλλαγές στον τρόπο που ζούμε, εργαζόμαστε και αλληλεπιδρούμε. Από την υγειονομική περίθαλψη και την εκπαίδευση μέχρι την ασφάλεια και τη διαχείριση κρίσεων, η ΤΝ προσφέρει πρωτοφανείς δυνατότητες για βελτίωση των υπηρεσιών και της ποιότητας ζωής [1]. Ως "ικανότητα ενός συστήματος να ερμηνεύει σωστά τα εξωτερικά δεδομένα, να μαθαίνει από αυτά και να χρησιμοποιεί τη γνώση για την επίτευξη συγκεκριμένων στόχων" [2], η ΤΝ έχει εξελιχθεί σε έναν από τους σημαντικότερους καταλύτες τεχνολογικής προόδου.

Ωστόσο, η ενσωμάτωση της ΤΝ στις κοινωνικοθεσμικές διαδικασίες φέρνει μαζί της σημαντικές προκλήσεις και κινδύνους [3]. Η αδιαφάνεια των αλγορίθμων, οι προκαταλήψεις στη λήψη αποφάσεων, και οι ανησυχίες για την προστασία των προσωπικών δεδομένων αποτελούν μερικά μόνο από τα ζητήματα που χρήζουν άμεσης αντιμετώπισης. Για παράδειγμα, τα συστήματα ΤΝ συχνά λειτουργούν ως "μαύρα κουτιά", καθιστώντας δύσκολη την κατανόηση και εξήγηση των αποφάσεών τους [4].

Η έλευση του Γενικού Κανονισμού για τη Προστασία των Δεδομένων (General Data Protection Regulation – GDPR) της Ευρωπαϊκής Ένωσης έχει θέσει νέες απαιτήσεις για τη διαφάνεια και τη λογοδοσία των συστημάτων ΤΝ [5]. Παράλληλα, η αυξανόμενη χρήση της ΤΝ σε κρίσιμους τομείς

όπως η υγειονομική περίθαλψη και η ποινική δικαιοσύνη εγείρει σοβαρά ερωτήματα σχετικά με την ηθική διάσταση των αποφάσεων που λαμβάνονται από τέτοια συστήματα [6].

Σε αυτό το πλαίσιο, καθίσταται επιτακτική η ανάγκη για ένα ολοκληρωμένο πλαίσιο που θα διασφαλίζει την ηθική και υπεύθυνη χρήση της ΤΝ. Η πρόκληση δεν περιορίζεται μόνο στην τεχνική διάσταση, αλλά επεκτείνεται στη διαμόρφωση κατάλληλων πολιτικών, κανονισμών και προτύπων δεοντολογίας που θα καθοδηγούν την ανάπτυξη και εφαρμογή των συστημάτων ΤΝ [7].

Από το 2019 έως σήμερα, το τοπίο της ΤΝ έχει μεταμορφωθεί ριζικά με την έλευση των Μεγάλων Γλωσσικών Μοντέλων (Large Language Models - LLMs) και της Παραγωγικής ΤΝ (Generative AI). Αυτή η εξέλιξη έχει οδηγήσει σε μια άνευ προηγουμένου δημοκρατικοποίηση της τεχνολογίας, καθιστώντας την ΤΝ προσβάσιμη σε ένα ευρύτερο κοινό μέσω φιλικών προς το χρήστη διεπαφών, όπως το ChatGPT.

Ωστόσο, αυτή η ευρεία διαθεσιμότητα έχει δημιουργήσει νέες και πιο σύνθετες προκλήσεις. Η ικανότητα της Παραγωγικής ΤΝ να δημιουργεί εξαιρετικά πειστικό συνθετικό περιεχόμενο έχει εντείνει τις ανησυχίες σχετικά με την παραπληροφόρηση και τα deepfakes. Η δυνατότητα μαζικής παραγωγής ψευδών πληροφοριών δεν απειλεί μόνο την εμπιστοσύνη και τις δημοκρατικές αξίες, αλλά εγείρει επίσης σημαντικά ζητήματα πνευματικής ιδιοκτησίας, καθώς τα μοντέλα ΤΝ εκπαιδεύονται σε τεράστιες ποσότητες δεδομένων, συχνά χωρίς ρητή άδεια.

Η ακαδημαϊκή κοινότητα, έχοντας ωριμάσει στην κατανόηση αυτών των προκλήσεων, προτείνει πλέον πιο ολοκληρωμένα πλαίσια για την αντιμετώπισή τους. Η ανάγκη για διεπιστημονικό διάλογο και δυναμική διακυβέρνηση έχει καταστεί επιτακτική, καθώς η τεχνολογία συνεχίζει να εξελίσσεται με πρωτοφανείς ρυθμούς. Αυτή η εξέλιξη απαιτεί μια ολιστική προσέγγιση που συνδυάζει τεχνικές λύσεις, ηθικές κατευθυντήριες γραμμές και ισχυρά κανονιστικά πλαίσια για τη διασφάλιση της υπεύθυνης ανάπτυξης και χρήσης της ΤΝ στην εποχή της Παραγωγικής ΤΝ.

1.2 Αντικείμενο της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία στοχεύει στη διερεύνηση των ηθικών και δεοντολογικών ζητημάτων που προκύπτουν από την αυξανόμενη χρήση της Τεχνητής Νοημοσύνης, με ιδιαίτερη έμφαση στο πλαίσιο του Γενικού Κανονισμού για τη Προστασία των Δεδομένων. Βασικός σκοπός είναι η ανάπτυξη ενός ολοκληρωμένου πλαισίου κατανόησης για την ηθική και ασφαλή χρήση της ΤΝ, καθώς και η διερεύνηση των προϋποθέσεων που απαιτούνται για την υπεύθυνη ανάπτυξη και εφαρμογή συστημάτων ΤΝ. Μέσω συστηματικής βιβλιογραφικής ανασκόπησης, η εργασία επιδιώκει να αναλύσει κρίσιμα ζητήματα όπως η προστασία της ιδιωτικότητας, η διαφάνεια των αλγορίθμων, και η ηθική λήψη αποφάσεων στα συστήματα ΤΝ. Παράλληλα, στοχεύει στην αξιολόγηση της αναγκαιότητας θέσπισης ενός πλαισίου ηθικής και δεοντολογικής καθοδήγησης για τους επαγγελματίες του χώρου, ώστε να διασφαλιστεί η υπεύθυνη ανάπτυξη και χρήση των τεχνολογιών ΤΝ προς όφελος της κοινωνίας.

Το πρώτο και θεμελιώδες ερώτημα που εξετάζει η εργασία αφορά τα απαραίτητα βήματα για τη διασφάλιση της ηθικής χρήσης της Τεχνητής Νοημοσύνης. Η μελέτη αναλύει το ζήτημα της αλγοριθμικής δικαιοσύνης και λογοδοσίας, εξετάζοντας πώς η αδιαφάνεια των αλγορίθμων μπορεί να αντιμετωπιστεί μέσω συγκεκριμένων περιορισμών και πλαισίων ελέγχου [8], [9]. Ιδιαίτερη έμφαση δίνεται στην έννοια της περιβάλλουσας (envelopment) ως εναλλακτική προσέγγιση στο ζήτημα της διαφάνειας [10], [11], καθώς και στη σημασία των συστημάτων ευφυούς βιντεοεπιτήρησης (Intelligent Video Surveillance – IVS) [12]. Επιπλέον, διερευνάται ο ρόλος του μη δεσμευτικού δικαίου (soft law) και των αρχών δεοντολογίας στη διαμόρφωση ενός αξιόπιστου πλαισίου για την ηθική χρήση της ΤΝ [13], [14], με ιδιαίτερη αναφορά στις πρωτοβουλίες της Ευρωπαϊκής Ένωσης [15], [16] και τις προκλήσεις που προκύπτουν από την έλλειψη κατάλληλων αρχών [2], [17].

Το δεύτερο ερευνητικό ερώτημα εστιάζει στις απαραίτητες προϋποθέσεις και τα χαρακτηριστικά που πρέπει να διαθέτει η Τεχνητή Νοημοσύνη για να θεωρείται ασφαλής. Κεντρικό ρόλο σε αυτή την ανάλυση διαδραματίζουν οι αρχές της υπεύθυνης ΤΝ, οι οποίες στηρίζονται στους πυλώνες της υπευθυνότητας, της διαφάνειας και της ενσωμάτωσης δικαιοσύνης και ηθικής [18]. Ιδιαίτερη έμφαση δίνεται στην ανάγκη για ασφάλεια και προστασία των συστημάτων ΤΝ, με την διαφάνεια να αποτελεί καθοριστικό παράγοντα στην κατανόηση των αλγοριθμικών αποφάσεων [19]. Παράλληλα, διερευνώνται ζητήματα λογοδοσίας και ευθύνης στην ΤΝ, με έμφαση στην ανάγκη επαλήθευσης και επικύρωσης των συστημάτων [20], ενώ εξετάζεται και η σημασία της ενσωμάτωσης ανθρώπινων αξιών για την οικοδόμηση εμπιστοσύνης [21]. Τέλος, αναδεικνύεται η αναγκαιότητα προετοιμασίας του εργατικού δυναμικού για τις προκλήσεις που φέρνει η ΤΝ, καθώς και η σημασία της διεθνούς συνεργασίας για τη θέσπιση κοινών προτύπων ασφαλείας [22].

Το τρίτο ερευνητικό ερώτημα εξετάζει το κρίσιμο ζήτημα της υπευθυνότητας της Τεχνητής Νοημοσύνης. Η μελέτη αναλύει το εξελισσόμενο πεδίο της ΤΝ και τις συνεχείς τροποποιήσεις στον ορισμό της, με ιδιαίτερη έμφαση στην αυξανόμενη ζήτηση για απόδοση ευθυνών [8], [23]. Εξετάζεται η σημασία της αλγοριθμικής υπευθυνότητας και της προέλευσης των αποφάσεων, καθώς τα συστήματα ΤΝ πρέπει να διασφαλίζουν τη δυνατότητα ανίχνευσης κάθε απόφασης [24]. Ιδιαίτερη προσοχή δίνεται στην εκπαίδευση των συστημάτων ΤΝ και την ικανότητά τους να αναγνωρίζουν και να ανταποκρίνονται σε ηθικά απαιτητικές καταστάσεις [25], ενώ παράλληλα αναλύονται οι προκλήσεις που προκύπτουν από την αυτοματοποιημένη λήψη αποφάσεων [20] και την αυτονομία των συστημάτων ΤΝ [26], [27]. Τέλος, εξετάζονται οι ευρύτερες επιπτώσεις της ΤΝ στην κοινωνία και η ανάγκη για διαφάνεια και λογοδοσία στην ανάπτυξη και χρήση της [28], [29].

Το τέταρτο ερευνητικό ερώτημα εξετάζει πώς η Τεχνητή Νοημοσύνη αντιμετωπίζει τα ζητήματα ιδιωτικότητας, με έμφαση στις προκλήσεις που προκύπτουν από τη διαχείριση και προστασία των προσωπικών δεδομένων. Η μελέτη αναλύει το ρόλο των μεγάλων τεχνολογικών εταιρειών στη διαμόρφωση πολιτικών προστασίας δεδομένων [30], τις επιπτώσεις της συγκέντρωσης δεδομένων στην κοινωνία [31], και την ανάγκη για διαφάνεια στα συστήματα ΤΝ

[32]. Ιδιαίτερη έμφαση δίνεται στην εξέταση των μηχανισμών προστασίας της ιδιωτικότητας και στην αξιολόγηση της αποτελεσματικότητας υφιστάμενων ρυθμιστικών πλαισίων [33]. Επιπλέον, διερευνώνται οι προκλήσεις που αντιμετωπίζουν οι αναπτυσσόμενες χώρες στην προστασία των δεδομένων των πολιτών τους [34], καθώς και ο ρόλος των διεθνών οργανισμών στη διαμόρφωση προτύπων προστασίας της ιδιωτικότητας [35].

Το πέμπτο ερευνητικό ερώτημα εξετάζει την αναγκαιότητα της ηθικής και δεοντολογικής καθοδήγησης στην ανάπτυξη και χρήση της Τεχνητής Νοημοσύνης. Η μελέτη αναλύει τη σημασία της ενσωμάτωσης ηθικών αρχών στη διδασκαλία της μηχανικής μάθησης [36], την ανάγκη για δεοντολογικά πρότυπα στην οικοδόμηση αξιόπιστων συστημάτων [37], και τις προκλήσεις στην εξισορρόπηση τεχνολογικής προόδου και κοινωνικής ευθύνης [22]. Ιδιαίτερη έμφαση δίνεται στο ρόλο των εκπαιδευτικών ιδρυμάτων και των επιχειρήσεων στην προώθηση της ηθικής συνείδησης [38], καθώς και στην ανάγκη για συνεργασία μεταξύ κυβέρνησης και ιδιωτικού τομέα στη διαμόρφωση κατάλληλων ρυθμιστικών πλαισίων [39]. Επιπλέον, εξετάζεται η σημασία της διεπιστημονικής προσέγγισης στην αντιμετώπιση των ηθικών προκλήσεων, καθώς και η ανάγκη για συνεχή αξιολόγηση και προσαρμογή των ηθικών κατευθυντήριων γραμμών στις εξελισσόμενες τεχνολογικές δυνατότητες [40].

Η συνεισφορά της παρούσας διπλωματικής εργασίας είναι πολυδιάστατη και σημαντική για την κατανόηση των ηθικών παραμέτρων της Τεχνητής Νοημοσύνης. Πρώτον, παρέχει μια συστηματική ανασκόπηση και σύνθεση της τρέχουσας βιβλιογραφίας σχετικά με τις ηθικές και δεοντολογικές προκλήσεις της ΤΝ, συμβάλλοντας στην καλύτερη κατανόηση του πεδίου. Δεύτερον, προσφέρει ένα ολοκληρωμένο πλαίσιο ανάλυσης των προϋποθέσεων για την ασφαλή και ηθική χρήση της ΤΝ, το οποίο μπορεί να αξιοποιηθεί τόσο από ερευνητές όσο και από επαγγελματίες του χώρου. Τρίτον, εξετάζει κριτικά τη σχέση μεταξύ τεχνολογικής καινοτομίας και προστασίας των ανθρωπίνων δικαιωμάτων, συμβάλλοντας στον διάλογο για τη διαμόρφωση κατάλληλων ρυθμιστικών πλαισίων. Τέταρτον, αναδεικνύει τη σημασία της διεπιστημονικής προσέγγισης στην αντιμετώπιση των προκλήσεων της ΤΝ, υπογραμμίζοντας την ανάγκη συνεργασίας μεταξύ τεχνικών εμπειρογνομόνων, ηθικών φιλοσόφων και υπευθύνων χάραξης πολιτικής. Τέλος, η εργασία προτείνει συγκεκριμένες κατευθύνσεις για μελλοντική έρευνα, συμβάλλοντας στην περαιτέρω ανάπτυξη του πεδίου της ηθικής της ΤΝ.

1.3 Μεθοδολογική Προσέγγιση

Η παρούσα διπλωματική εργασία ακολουθεί τη μεθοδολογία της συστηματικής βιβλιογραφικής ανασκόπησης σε δύο χρονικές περιόδους, εξετάζοντας την εξέλιξη των ηθικών ζητημάτων στην Τεχνητή Νοημοσύνη πριν και μετά την έλευση της Παραγωγικής ΤΝ. Η πρώτη φάση της έρευνας (2015-2019) εστιάζει στην αρχική ανάπτυξη του πεδίου, ενώ η δεύτερη φάση (2024) εξετάζει τις πρόσφατες εξελίξεις στην εποχή της Generative AI, επιτρέποντας μια συγκριτική ανάλυση των ευρημάτων.

1.3.1 Πηγές δεδομένων

Η αναζήτηση της βιβλιογραφίας πραγματοποιήθηκε σε δύο φάσεις:

1. **Φάση 1 (2015-2019):** Η αρχική έρευνα διεξήχθη κατά το διάστημα Ιούλιος 2019 - Νοέμβριος 2019, με αναζήτηση σε οκτώ κύριες βάσεις δεδομένων:

- IEEE Xplore Digital Library
- ACM Digital Library
- Google Scholar
- CiteSeerX
- Inspec
- ScienceDirect
- Scopus
- RefereX Engineering

Για κάθε βάση δεδομένων αναπτύχθηκε εξειδικευμένη στρατηγική αναζήτησης, χρησιμοποιώντας συνδυασμούς λέξεων-κλειδιών όπως "artificial intelligence", "ethics", "machine learning", "privacy", "GDPR", "responsible AI" κ.ά. Συνολικά συγκεντρώθηκαν 4,904 δημοσιεύσεις, οι οποίες μετά την αφαίρεση διπλοτύπων και την εφαρμογή των κριτηρίων επιλογής/αποκλεισμού περιορίστηκαν σε 4,036 μοναδικές δημοσιεύσεις.

2. **Φάση 2 (2024):** Η συμπληρωματική έρευνα πραγματοποιήθηκε στις ίδιες βάσεις δεδομένων, με επικαιροποιημένες λέξεις-κλειδιά που συμπεριλαμβάνουν νέους όρους όπως "generative AI", "large language models", "AI governance", και "AI democratization". Η νέα αναζήτηση επικεντρώθηκε στο διάστημα 1/1/2024 - 31/12/2024, με στόχο την καταγραφή των πρόσφατων εξελίξεων στο πεδίο.

1.3.2 Κριτήρια επιλογής/αποκλεισμού

Τα βασικά κριτήρια επιλογής παραμένουν σταθερά και για τις δύο φάσεις:

- Δημοσιεύσεις που εξετάζουν ζητήματα σχετικά με τον Γενικό Κανονισμό Προστασίας Δεδομένων (GDPR) και την Τεχνητή Νοημοσύνη
- Μελέτες που διερευνούν ηθικά ζητήματα στη χρήση της Τεχνητής Νοημοσύνης και της μηχανικής μάθησης
- Έρευνες από τον κλάδο της πληροφορικής και συναφών επιστημονικών πεδίων

Για τη Φάση 2, προστέθηκαν επιπλέον κριτήρια επιλογής:

- Μελέτες που εξετάζουν τις επιπτώσεις της Παραγωγικής ΤΝ
- Έρευνες σχετικές με τη δημοκρατικοποίηση της ΤΝ
- Δημοσιεύσεις που αναλύουν ζητήματα πνευματικής ιδιοκτησίας στην εποχή της Παραγωγικής ΤΝ

Τα κριτήρια αποκλεισμού παραμένουν τα ίδια:

- Δημοσιεύσεις σε γλώσσα διαφορετική της αγγλικής
- Μελέτες που δεν σχετίζονται με τον κλάδο της πληροφορικής
- Δημοσιεύσεις χωρίς επιστημονική αξιολόγηση (peer review).

Η συγκριτική ανάλυση των ευρημάτων από τις δύο φάσεις επιτρέπει την κατανόηση της εξέλιξης του πεδίου και την ανάδειξη νέων προκλήσεων και προσεγγίσεων στην ηθική διακυβέρνηση της ΤΝ. Όλες οι πηγές προέρχονται από αξιόπιστες επιστημονικές βάσεις δεδομένων με πρόσβαση μέσω του πανεπιστημιακού δικτύου, διασφαλίζοντας την εγκυρότητα και την αξιοπιστία της έρευνας.

1.4 Δομή της διπλωματικής

Η παρούσα διπλωματική εργασία είναι δομημένη σε πέντε κύρια κεφάλαια. Το Κεφάλαιο 1 αποτελεί την εισαγωγή, όπου παρουσιάζεται το γενικό πλαίσιο της έρευνας, ο σκοπός και οι στόχοι της εργασίας, καθώς και η μεθοδολογική προσέγγιση που ακολουθήθηκε.

Το Κεφάλαιο 2 παρέχει το θεωρητικό υπόβαθρο της έρευνας, εισάγοντας τις βασικές έννοιες και τους ορισμούς που είναι απαραίτητοι για την κατανόηση του θέματος, συμπεριλαμβανομένων των ειδικών χαρακτηριστικών του GenAI και των διαφορών του από άλλες μορφές ΤΝ. Αναλύεται το νομικό πλαίσιο με έμφαση στον Γενικό Κανονισμό Προστασίας Δεδομένων (GDPR) και το Ευρωπαϊκό Πλαίσιο για την ΤΝ, λαμβάνοντας υπόψη τις πρόσφατες εξελίξεις και προσαρμογές που σχετίζονται με το Generative AI.

Στο Κεφάλαιο 3 παρουσιάζεται αναλυτικά η μεθοδολογία της έρευνας. Διατυπώνονται τα ερευνητικά ερωτήματα, περιγράφεται η μέθοδος συλλογής δεδομένων και αναλύονται τα κριτήρια επιλογής και αποκλεισμού της βιβλιογραφίας. Επιπλέον, συζητούνται οι περιορισμοί και οι δυσκολίες που αντιμετωπίστηκαν κατά τη διεξαγωγή της έρευνας.

Το Κεφάλαιο 4 αποτελεί τον πυρήνα της εργασίας, παρουσιάζοντας την ανάλυση και τα αποτελέσματα της έρευνας. Το κεφάλαιο διαρθρώνεται σε πέντε βασικές ενότητες που αντιστοιχούν στα ερευνητικά ερωτήματα: διασφάλιση της ηθικής χρήσης της ΤΝ, προϋποθέσεις ασφαλούς χρήσης, ζητήματα υπευθυνότητας και ηθικού σχεδιασμού, προστασία της ιδιωτικότητας, και ανάγκη για ηθική καθοδήγηση, λαμβάνοντας υπόψη τις νέες προκλήσεις που εισάγει το GenAI. Εξετάζεται η διαχρονική εξέλιξη των ζητημάτων αυτών, συγκρίνοντας τα ευρήματα μεταξύ της περιόδου 2015-2019 και της περιόδου 2024.

Τέλος, το Κεφάλαιο 5 παρουσιάζει τα συμπεράσματα της έρευνας, συνοψίζοντας τα κύρια ευρήματα και παρέχοντας απαντήσεις στα ερευνητικά ερωτήματα. Εξετάζονται οι επιπτώσεις της εξέλιξης του Generative AI για μελλοντική έρευνα στο πεδίο της ηθικής της Τεχνητής Νοημοσύνης. Η εργασία ολοκληρώνεται με την παράθεση της βιβλιογραφίας και των παραρτημάτων που υποστηρίζουν την έρευνα.

2

Θεωρητικό Υπόβαθρο

Το παρόν κεφάλαιο επιχειρεί μια παρουσίαση του θεωρητικού υποβάθρου που είναι απαραίτητο για την κατανόηση των πολύπλευρων διαστάσεων της ΤΝ και των προκλήσεων που συνδέονται με την ηθική της χρήση. Ξεκινώντας από τους βασικούς ορισμούς και έννοιες, προχωρά στην ανάλυση του νομικού και κανονιστικού πλαισίου, με ιδιαίτερη έμφαση στον GDPR, και καταλήγει στη διερεύνηση των πρακτικών εφαρμογών και των κοινωνικών επιπτώσεων της ΤΝ [41]. Η κατανόηση αυτού του πλαισίου είναι θεμελιώδης για την ανάπτυξη συστημάτων ΤΝ που όχι μόνο είναι τεχνικά άρτια, αλλά και συμμορφώνονται με τις ηθικές αρχές και τις νομικές απαιτήσεις, διασφαλίζοντας παράλληλα τον σεβασμό στα ανθρώπινα δικαιώματα και την προστασία της ιδιωτικότητας [42], [43].

2.1 Βασικές Έννοιες και Ορισμοί

Η εξέλιξη των τεχνολογιών ΤΝ, από τα πρώιμα συστήματα κανόνων μέχρι τα σύγχρονα μοντέλα βαθιάς μάθησης, έχει διαμορφώσει ένα πολύπλοκο τοπίο όπου η ακριβής οριοθέτηση των εννοιών και η κατανόηση των διασυνδέσεων τους καθίσταται ολοένα και πιο σημαντική για ερευνητές, μηχανικούς και υπεύθυνους χάραξης πολιτικής [44], [45].

2.1.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη αναφέρεται στην ικανότητα των υπολογιστικών συστημάτων να εκτελούν εργασίες που παραδοσιακά απαιτούσαν ανθρώπινη νοημοσύνη [46]. Πιο συγκεκριμένα, ορίζεται ως η επιστήμη και η μηχανική της δημιουργίας ευφών μηχανών, ιδιαίτερα ευφών υπολογιστικών προγραμμάτων, που μπορούν να αντιλαμβάνονται το περιβάλλον τους, να μαθαίνουν από την εμπειρία, και να προσαρμόζουν τη συμπεριφορά τους για την επίτευξη συγκεκριμένων στόχων [47]. Τα βασικά χαρακτηριστικά της ΤΝ περιλαμβάνουν την ικανότητα επεξεργασίας φυσικής γλώσσας, τη μηχανική μάθηση, την αυτοματοποιημένη συλλογιστική και την αναγνώριση προτύπων [48].

Η ιστορία της ΤΝ ξεκινά από τη δεκαετία του 1950, με το θεμελιώδες έργο του Alan Turing και τη δημοσίευση του περίφημου "Turing Test" [49]. Η εξέλιξή της μπορεί να χωριστεί σε διακριτές περιόδους:

1. Πρώιμη περίοδος (1950-1969): Ανάπτυξη βασικών εννοιών και πρώτων προγραμμάτων επίλυσης προβλημάτων
2. Κλασική περίοδος (1970-1989): Ανάπτυξη έμπειρων συστημάτων και πρώτων εφαρμογών σε πραγματικά προβλήματα
3. Σύγχρονη περίοδος (1990-2010): Εμφάνιση νευρωνικών δικτύων και μηχανικής μάθησης
4. Εποχή βαθιάς μάθησης (2011-σήμερα): Ραγδαία ανάπτυξη μοντέλων βαθιάς μάθησης και εφαρμογών σε ευρύ φάσμα τομέων [50]

Η ΤΝ μπορεί να κατηγοριοποιηθεί με διάφορους τρόπους, με την πιο διαδεδομένη διάκριση να είναι μεταξύ:

1. **Στενή ή Ασθενής ΤΝ (Narrow/Weak AI):** Συστήματα σχεδιασμένα για την εκτέλεση συγκεκριμένων εργασιών, όπως αναγνώριση εικόνας ή επεξεργασία φυσικής γλώσσας.
2. **Γενική ή Ισχυρή ΤΝ (General/Strong AI):** Υποθετικά συστήματα που θα μπορούσαν να εκτελέσουν οποιαδήποτε διανοητική εργασία όπως ένας άνθρωπος [51].
3. **Υπερ-νοημοσύνη (Superintelligence):** Θεωρητικά συστήματα που θα ξεπερνούσαν τις ανθρώπινες γνωστικές ικανότητες σε όλους τους τομείς [52].

Επιπλέον, με βάση τη μεθοδολογία και την αρχιτεκτονική τους, τα συστήματα ΤΝ μπορούν να διακριθούν σε:

- Συστήματα βασισμένα σε κανόνες
- Συστήματα μηχανικής μάθησης
- Υβριδικά συστήματα
- Συστήματα βαθιάς μάθησης [53]

Στην πιο πρόσφατη εξέλιξη του πεδίου, η Παραγωγική ΤΝ (Generative AI) έχει αναδειχθεί ως μια νέα κατηγορία συστημάτων με μοναδικές δυνατότητες. Σε αντίθεση με τα παραδοσιακά συστήματα ΤΝ που επικεντρώνονται στην αναγνώριση προτύπων και τη λήψη αποφάσεων, η Παραγωγική ΤΝ μπορεί να δημιουργεί νέο περιεχόμενο μαθαίνοντας από υπάρχοντα σύνολα δεδομένων [54]. Χρησιμοποιεί προηγμένα μοντέλα όπως τα Γενετικά Ανταγωνιστικά Δίκτυα (GANs), τους Μεταβλητούς Αυτοκωδικοποιητές (VAEs) και τα αυτοπαλινδρομικά μοντέλα για την παραγωγή περιεχομένου υψηλής ποιότητας [55].

Η Παραγωγική ΤΝ περιλαμβάνει διάφορες κατηγορίες με διακριτές εφαρμογές:

- Παραγωγή Κειμένου: Μοντέλα όπως το GPT μπορούν να δημιουργούν πρωτότυπο περιεχόμενο και να προτείνουν αφηγηματικά πλαίσια [56]
- Παραγωγή Εικόνας: Συστήματα που δημιουργούν πρωτότυπα έργα τέχνης συνδυάζοντας διαφορετικές αισθητικές [57]
- Παραγωγή Ήχου: Εργαλεία που επιτρέπουν τη σύνθεση μουσικής και τη διερεύνηση νέων μουσικών ειδών [58]

2.1.2 Ηθική της Τεχνητής Νοημοσύνης

Η ηθική της Τεχνητής Νοημοσύνης αναφέρεται στο σύνολο των αρχών και κατευθυντήριων γραμμών που διέπουν την ανάπτυξη και χρήση συστημάτων ΤΝ με τρόπο που σέβεται τις

ανθρώπινες αξίες και προάγει το κοινό καλό [59]. Οι θεμελιώδεις αρχές αποτελούν το πλαίσιο για την ηθική ανάπτυξη και εφαρμογή της ΤΝ.

Η αρχή της δικαιοσύνης και μη-διάκρισης επιτάσσει την ίση μεταχείριση όλων των χρηστών από τα συστήματα ΤΝ. Αυτό σημαίνει ότι οι αλγόριθμοι πρέπει να σχεδιάζονται ώστε να αποφεύγουν συστημικές προκαταλήψεις που θα μπορούσαν να οδηγήσουν σε διακρίσεις με βάση το φύλο, τη φυλή, την ηλικία ή άλλα χαρακτηριστικά. Επιπλέον, τα συστήματα ΤΝ θα πρέπει να είναι προσβάσιμα από όλες τις κοινωνικές ομάδες, συμπεριλαμβανομένων των ατόμων με αναπηρίες [60].

Η διαφάνεια και λογοδοσία αποτελούν κρίσιμες αρχές για την οικοδόμηση εμπιστοσύνης στα συστήματα ΤΝ. Η επεξηγησιμότητα των αποφάσεων απαιτεί τα συστήματα να μπορούν να παρέχουν κατανοητές εξηγήσεις για τις αποφάσεις τους. Η δυνατότητα ελέγχου επιτρέπει την αξιολόγηση της λειτουργίας των συστημάτων, ενώ ο καθορισμός ευθυνών διασφαλίζει ότι υπάρχει σαφής απόδοση ευθύνης σε περίπτωση προβλημάτων [61].

Η αυτονομία και ιδιωτικότητα αφορούν το σεβασμό της ανθρώπινης αυτοδιάθεσης και της προστασίας των προσωπικών δεδομένων. Τα συστήματα ΤΝ πρέπει να σχεδιάζονται ώστε να ενισχύουν και όχι να υπονομεύουν την ανθρώπινη αυτονομία, παρέχοντας στους χρήστες ουσιαστικό έλεγχο επί των δεδομένων τους και των αποφάσεων που τους επηρεάζουν [62].

Οι τεχνικές προκλήσεις στην εφαρμογή των ηθικών αρχών είναι πολυδιάστατες. Η επίτευξη πλήρους διαφάνειας συχνά περιορίζεται από την εγγενή πολυπλοκότητα των αλγορίθμων, ιδιαίτερα στα συστήματα βαθιάς μάθησης. Υπάρχει συχνά μια αντιστάθμιση μεταξύ της ακρίβειας των μοντέλων και της ερμηνευσιμότητάς τους [63].

Τα κοινωνικά διλήμματα αντανakλούν την πολυπλοκότητα της εφαρμογής της ΤΝ σε διαφορετικά κοινωνικά πλαίσια. Η εξισορρόπηση ατομικών και συλλογικών συμφερόντων απαιτεί προσεκτική στάθμιση, ενώ οι πολιτισμικές διαφορές μεταξύ κοινωνιών μπορεί να οδηγήσουν σε διαφορετικές προσεγγίσεις στην ηθική της ΤΝ. Οι επιπτώσεις στην απασχόληση αποτελούν επίσης σημαντική πρόκληση που απαιτεί προσεκτική διαχείριση [64].

Τα ηθικά διλήμματα θέτουν ιδιαίτερα δύσκολα ερωτήματα. Η λήψη αποφάσεων σε καταστάσεις ζωής και θανάτου, όπως στην περίπτωση αυτόνομων οχημάτων, απαιτεί σαφείς ηθικές κατευθυντήριες γραμμές. Η κατανομή περιορισμένων πόρων και η απόδοση ευθύνης για λάθη των συστημάτων ΤΝ είναι επίσης κρίσιμα ζητήματα [65].

Η κοινωνική αποδοχή της ΤΝ εξαρτάται σε μεγάλο βαθμό από την ενσωμάτωση ηθικών αρχών στην ανάπτυξή της. Η οικοδόμηση εμπιστοσύνης απαιτεί συνεχή διάλογο με την κοινωνία και διαφάνεια στη λήψη αποφάσεων. Η διασφάλιση δημόσιας υποστήριξης είναι απαραίτητη για την επιτυχή υιοθέτηση της τεχνολογίας [42].

Η βιώσιμη ανάπτυξη της ΤΝ προϋποθέτει την ισορροπία μεταξύ τεχνολογικής προόδου και κοινωνικής ευθύνης. Η πρόληψη αρνητικών επιπτώσεων και η συμβολή στην κοινωνική πρόοδο απαιτούν προσεκτικό σχεδιασμό και συνεχή αξιολόγηση των συστημάτων ΤΝ [66].

Η κανονιστική συμμόρφωση αποτελεί βασική πτυχή της ηθικής ανάπτυξης ΤΝ. Η εναρμόνιση με τα νομοθετικά πλαίσια και η προληπτική αντιμετώπιση των ρυθμιστικών απαιτήσεων διασφαλίζουν τη νομιμότητα και αποδοχή των συστημάτων ΤΝ. Η διεθνής συνεργασία είναι απαραίτητη για την ανάπτυξη κοινών προτύπων και πρακτικών [22].

2.2 Νομικό Πλαίσιο και Κανονισμοί

Η ταχεία ανάπτυξη και ευρεία εφαρμογή της Τεχνητής Νοημοσύνης έχει καταστήσει επιτακτική την ανάγκη για θέσπιση ενός ολοκληρωμένου νομικού και κανονιστικού πλαισίου που θα διασφαλίζει την υπεύθυνη και ηθική χρήση της τεχνολογίας [67]. Η Ευρωπαϊκή Ένωση, πρωτοπόρος στη ρύθμιση ψηφιακών τεχνολογιών, έχει αναλάβει ηγετικό ρόλο στη διαμόρφωση κανόνων και προτύπων, με τον Γενικό Κανονισμό Προστασίας Δεδομένων να αποτελεί θεμέλιο λίθο αυτής της προσπάθειας [68]. Το πλαίσιο αυτό στοχεύει όχι μόνο στην προστασία των θεμελιωδών δικαιωμάτων των πολιτών, αλλά και στη δημιουργία ενός περιβάλλοντος που προάγει την καινοτομία και την ανάπτυξη αξιόπιστων συστημάτων ΤΝ [34].

2.2.1 GDPR και Προστασία Δεδομένων

Ο Γενικός Κανονισμός για την Προστασία των Δεδομένων (GDPR) θεσπίστηκε το 2016 και τέθηκε σε εφαρμογή το 2018, αποτελώντας το πιο ολοκληρωμένο πλαίσιο προστασίας προσωπικών δεδομένων παγκοσμίως [69]. Η σημασία του είναι καίρια για τη ρύθμιση της χρήσης προσωπικών δεδομένων στα συστήματα Τεχνητής Νοημοσύνης, καθώς θέτει συγκεκριμένες απαιτήσεις και περιορισμούς στον τρόπο με τον οποίο τα δεδομένα μπορούν να συλλεχθούν και να χρησιμοποιηθούν [70].

Οι θεμελιώδεις αρχές του GDPR ορίζουν το πλαίσιο μέσα στο οποίο πρέπει να λειτουργούν τα συστήματα ΤΝ. Η αρχή της νομιμότητας, διαφάνειας και δικαιοσύνης απαιτεί την επεξεργασία των δεδομένων με νόμιμο, διαφανή και δίκαιο τρόπο. Οι οργανισμοί πρέπει να είναι ξεκάθαροι σχετικά με τον τρόπο που συλλέγουν και χρησιμοποιούν τα προσωπικά δεδομένα, και να διασφαλίζουν ότι η επεξεργασία γίνεται σύμφωνα με το νόμο [71].

Η αρχή του περιορισμού του σκοπού είναι επίσης εξαιρετικά σημαντική, η οποία ορίζει ότι τα δεδομένα πρέπει να συλλέγονται για συγκεκριμένους, ρητούς και νόμιμους σκοπούς και να μην υποβάλλονται σε περαιτέρω επεξεργασία κατά τρόπο ασύμβατο με τους σκοπούς αυτούς. Αυτό είναι ιδιαίτερα κρίσιμο για τα συστήματα ΤΝ, όπου τα δεδομένα μπορούν εύκολα να επαναχρησιμοποιηθούν για διαφορετικούς σκοπούς [72].

Η ελαχιστοποίηση των δεδομένων και η ακρίβεια λειτουργούν συμπληρωματικά στις παραπάνω αρχές, απαιτώντας τα δεδομένα να είναι επαρκή, σχετικά και να περιορίζονται στο αναγκαίο για τους σκοπούς της επεξεργασίας, ενώ παράλληλα πρέπει να διατηρούνται ακριβή και ενημερωμένα [73].

Στο πλαίσιο της ενίσχυσης της προστασίας των προσωπικών δεδομένων, ο GDPR θεσπίζει και ενδυναμώνει σημαντικά δικαιώματα για τα υποκείμενα των δεδομένων. Το δικαίωμα ενημέρωσης και πρόσβασης αποτελεί θεμελιώδες δικαίωμα, επιτρέποντας στα άτομα να γνωρίζουν ποια δεδομένα τους συλλέγονται και πώς χρησιμοποιούνται, καθώς και να αποκτούν αντίγραφο αυτών των δεδομένων [74].

Συμπληρωματικά, το δικαίωμα διόρθωσης και διαγραφής ("δικαίωμα στη λήθη") παρέχει στα άτομα τη δυνατότητα να ζητούν τη διόρθωση ανακριβών δεδομένων ή τη διαγραφή τους υπό συγκεκριμένες συνθήκες. Η εφαρμογή αυτών των δικαιωμάτων δημιουργεί ιδιαίτερες προκλήσεις για τα συστήματα ΤΝ, καθώς τα δεδομένα μπορεί να έχουν ήδη χρησιμοποιηθεί για την εκπαίδευση μοντέλων [75].

Οι οργανισμοί που αναπτύσσουν και χρησιμοποιούν συστήματα ΤΝ έχουν εκτεταμένες υποχρεώσεις υπό το GDPR. Η λογοδοσία και η συμμόρφωση απαιτούν την απόδειξη της τήρησης των αρχών του κανονισμού, συμπεριλαμβανομένης της τήρησης αρχείων επεξεργασίας, της διενέργειας εκτιμήσεων αντικτύπου και του ορισμού Υπευθύνου Προστασίας Δεδομένων όπου απαιτείται [76].

Η ασφάλεια των δεδομένων αποτελεί επίσης κρίσιμη υποχρέωση. Οι οργανισμοί πρέπει να εφαρμόζουν κατάλληλα τεχνικά και οργανωτικά μέτρα για την προστασία των δεδομένων από μη εξουσιοδοτημένη πρόσβαση, απώλεια ή καταστροφή. Σε αυτό το πλαίσιο, η υποχρέωση γνωστοποίησης παραβιάσεων εντός 72 ωρών στις αρχές και τα επηρεαζόμενα άτομα αποτελεί σημαντικό μηχανισμό διαφάνειας και λογοδοσίας [77], [78].

2.2.2 Ευρωπαϊκό Πλαίσιο για την ΤΝ

Η Ευρωπαϊκή Ένωση έχει αναλάβει ηγετικό ρόλο στη διαμόρφωση ενός ολοκληρωμένου ρυθμιστικού πλαισίου για την Τεχνητή Νοημοσύνη, με στόχο την προώθηση της καινοτομίας παράλληλα με τη διασφάλιση των θεμελιωδών δικαιωμάτων και της ασφάλειας των πολιτών [79].

Η Ευρωπαϊκή Επιτροπή έχει διαμορφώσει ένα πλαίσιο κατευθυντήριων γραμμών για την ανάπτυξη αξιόπιστης ΤΝ. Οι βασικές αρχές αυτού του πλαισίου περιλαμβάνουν τη διαφάνεια, τη λογοδοσία και την ανθρωποκεντρική προσέγγιση. Ιδιαίτερη έμφαση δίνεται στην ανάγκη για "αξιόπιστη ΤΝ" (Trustworthy AI), η οποία πρέπει να σέβεται τις θεμελιώδεις αξίες και τα δικαιώματα των πολιτών της ΕΕ, ενώ παράλληλα παραμένει τεχνικά εύρωστη και ασφαλής [80]. Η προσέγγιση της ΕΕ βασίζεται σε τρεις βασικούς πυλώνες:

1. Προώθηση της αριστείας στην ΤΝ
2. Διασφάλιση της εμπιστοσύνης στην τεχνολογία
3. Προστασία των θεμελιωδών δικαιωμάτων [81]

Ο Κανονισμός για την Τεχνητή Νοημοσύνη (AI Act) αποτελεί την πρώτη ολοκληρωμένη νομοθετική προσπάθεια παγκοσμίως για τη ρύθμιση της ΤΝ. Εισάγει μια προσέγγιση βασισμένη στον κίνδυνο, κατηγοριοποιώντας τις εφαρμογές ΤΝ σε τέσσερα επίπεδα κινδύνου:

- Μη αποδεκτός κίνδυνος (απαγορευμένες πρακτικές)
- Υψηλός κίνδυνος (αυστηρές απαιτήσεις)
- Περιορισμένος κίνδυνος (απαιτήσεις διαφάνειας)
- Ελάχιστος κίνδυνος (ελάχιστοι περιορισμοί) [82]

Παράλληλα, η ΕΕ έχει θεσπίσει συγκεκριμένες απαιτήσεις για συστήματα ΤΝ υψηλού κινδύνου, συμπεριλαμβανομένων:

- Υποχρεωτικής αξιολόγησης συμμόρφωσης πριν τη διάθεση στην αγορά
- Συστημάτων διαχείρισης κινδύνου
- Απαιτήσεων για υψηλής ποιότητας σύνολα δεδομένων
- Λεπτομερούς τεκμηρίωσης
- Διαφάνειας και παροχής πληροφοριών στους χρήστες [83]

Η ΕΕ προωθεί ενεργά τη διεθνή συνεργασία στον τομέα της ρύθμισης της ΤΝ, αναγνωρίζοντας τον παγκόσμιο χαρακτήρα της τεχνολογίας. Συγκεκριμένα:

- Συνεργάζεται με διεθνείς οργανισμούς όπως ο ΟΟΣΑ (Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης) και η UNESCO (United Nations Educational, Scientific and Cultural Organization) για την ανάπτυξη κοινών προτύπων
- Προωθεί τον διάλογο με στρατηγικούς εταίρους για την εναρμόνιση ρυθμιστικών προσεγγίσεων
- Υποστηρίζει την ανάπτυξη διεθνών προτύπων για την ΤΝ [22], [84]

Η προσέγγιση της ΕΕ στη ρύθμιση της ΤΝ αντανακλά μια προσπάθεια εξισορρόπησης μεταξύ της προώθησης της καινοτομίας και της προστασίας των θεμελιωδών δικαιωμάτων. Η επιτυχία αυτής της προσέγγισης θα εξαρτηθεί από την αποτελεσματική εφαρμογή των κανονισμών και την προσαρμοστικότητά τους στις ραγδαίες τεχνολογικές εξελίξεις [42], [85].

2.2.3 Πρότυπα και Πιστοποιήσεις

Τα πρότυπα και οι πιστοποιήσεις για συστήματα Τεχνητής Νοημοσύνης αποτελούν βασικά εργαλεία για τη διασφάλιση της ποιότητας, της αξιοπιστίας και της ασφάλειας των εφαρμογών ΤΝ. Η ανάπτυξή τους είναι ιδιαίτερα σημαντική καθώς παρέχει ένα κοινό πλαίσιο αναφοράς για την αξιολόγηση και πιστοποίηση των συστημάτων ΤΝ [86].

Οι διεθνείς οργανισμοί τυποποίησης έχουν αναπτύξει και συνεχίζουν να αναπτύσσουν πρότυπα για την ΤΝ. Ο Διεθνής Οργανισμός Τυποποίησης (International Organization for Standardization – ISO) και η Διεθνής Ηλεκτροτεχνική Επιτροπή (International Electrotechnical Commission – IEC) έχουν δημιουργήσει την κοινή τεχνική επιτροπή ISO/IEC JTC 1/SC 42, η οποία επικεντρώνεται αποκλειστικά στην τυποποίηση της ΤΝ [87]. Τα βασικά πρότυπα περιλαμβάνουν:

- ISO/IEC 22989: Παρέχει τις βασικές έννοιες και ορολογία για την ΤΝ [88]
- ISO/IEC 23053: Καθορίζει το πλαίσιο για συστήματα μηχανικής μάθησης [89]
- ISO/IEC 42001: Προσδιορίζει απαιτήσεις για συστήματα διαχείρισης ΤΝ [90]

Παράλληλα, έχουν αναπτυχθεί εξειδικευμένα πρότυπα για συγκεκριμένες πτυχές της ΤΝ:

- Ασφάλεια και διαχείριση κινδύνων
- Ποιότητα δεδομένων και διαχείριση προκαταλήψεων
- Διαφάνεια και δυνατότητα επεξήγησης αλγορίθμων [91]

Οι μηχανισμοί συμμόρφωσης για συστήματα ΤΝ έχουν εξελιχθεί σε ένα πολυεπίπεδο σύστημα αξιολόγησης και πιστοποίησης. Σε πρώτο επίπεδο, η αυτο-αξιολόγηση αποτελεί θεμελιώδη διαδικασία όπου οι οργανισμοί διενεργούν εσωτερική αξιολόγηση της συμμόρφωσής τους με τα καθιερωμένα πρότυπα. Αυτή η διαδικασία περιλαμβάνει λεπτομερή τεκμηρίωση των διαδικασιών και των αποτελεσμάτων, καθώς και την εφαρμογή μηχανισμών συνεχούς παρακολούθησης και βελτίωσης [92].

Συμπληρωματικά με την αυτο-αξιολόγηση, η εξωτερική πιστοποίηση παρέχει ένα επιπλέον επίπεδο διασφάλισης. Διαπιστευμένοι φορείς αναλαμβάνουν τον έλεγχο συμμόρφωσης με συγκεκριμένα πρότυπα, εφαρμόζοντας αυστηρά κριτήρια αξιολόγησης. Η διαδικασία αυτή δεν

είναι στατική, αλλά απαιτεί περιοδική επαναξιολόγηση για τη διασφάλιση της συνεχούς συμμόρφωσης και τη βελτίωση των συστημάτων [20].

Η αξιολόγηση συστημάτων TN απαιτεί μια ολιστική προσέγγιση που καλύπτει πολλαπλές διαστάσεις. Η τεχνική αξιολόγηση εστιάζει στις βασικές λειτουργικές παραμέτρους των συστημάτων, όπως η ακρίβεια και η απόδοση, η ευρωστία και η αξιοπιστία, καθώς και η ασφάλεια και η προστασία. Αυτές οι παράμετροι αξιολογούνται μέσω συγκεκριμένων μετρικών και δοκιμών που επιτρέπουν την αντικειμενική μέτρηση της απόδοσης του συστήματος [93].

Παράλληλα με την τεχνική διάσταση, η ηθική αξιολόγηση εξετάζει τη συμμόρφωση των συστημάτων με θεμελιώδεις αρχές όπως η δικαιοσύνη και η μη-διάκριση. Ιδιαίτερη έμφαση δίνεται στη διαφάνεια και τη λογοδοσία των συστημάτων, καθώς και στο σεβασμό της ιδιωτικότητας των χρηστών. Η αξιολόγηση αυτών των παραμέτρων απαιτεί τόσο ποσοτικές όσο και ποιοτικές μεθόδους ανάλυσης [94].

Η κοινωνική διάσταση της αξιολόγησης εξετάζει τις ευρύτερες επιπτώσεις των συστημάτων TN στην κοινωνία. Αυτό περιλαμβάνει την ανάλυση των κοινωνικών επιπτώσεων, την εκτίμηση του περιβαλλοντικού αντικτύπου και την αξιολόγηση των οικονομικών επιπτώσεων. Η πολυπλοκότητα αυτής της διάστασης απαιτεί τη συνεργασία διαφόρων ειδικών και την εφαρμογή διεπιστημονικών μεθόδων αξιολόγησης [95].

Η εφαρμογή προτύπων και πιστοποιήσεων είναι καίρια για την οικοδόμηση εμπιστοσύνης στα συστήματα TN. Ωστόσο, η ταχεία εξέλιξη της τεχνολογίας απαιτεί συνεχή επικαιροποίηση και προσαρμογή των προτύπων. Η συνεργασία μεταξύ ρυθμιστικών αρχών, βιομηχανίας και ακαδημαϊκής κοινότητας είναι απαραίτητη για την ανάπτυξη αποτελεσματικών και επίκαιρων προτύπων [96].

2.3 Εφαρμογές και Επιπτώσεις της TN

Έχοντας εξετάσει το θεωρητικό υπόβαθρο της Τεχνητής Νοημοσύνης και το νομικό-κανονιστικό πλαίσιο που τη διέπει, είναι σημαντικό να διερευνήσουμε τις πρακτικές εφαρμογές της τεχνολογίας και τις πολυδιάστατες επιπτώσεις της στην κοινωνία. Η ευρεία διείσδυση της TN σε διάφορους τομείς της ανθρώπινης δραστηριότητας, από την υγειονομική περίθαλψη μέχρι τη δημόσια διοίκηση, δημιουργεί τόσο ευκαιρίες για καινοτομία και πρόοδο όσο και προκλήσεις που απαιτούν προσεκτική διαχείριση και αντιμετώπιση.

2.3.1 Τομείς Εφαρμογής

Η Τεχνητή Νοημοσύνη βρίσκει εφαρμογή σε ένα ευρύ φάσμα τομέων, με κάθε περίπτωση να παρουσιάζει μοναδικές προκλήσεις και ευκαιρίες για την αξιοποίηση της τεχνολογίας προς όφελος της κοινωνίας [48].

Στον τομέα της υγείας, η TN έχει αναδειχθεί σε ισχυρό εργαλείο για τη βελτίωση της διάγνωσης, της θεραπείας και της διαχείρισης ασθενών. Τα συστήματα TN υποστηρίζουν την ανάλυση ιατρικών εικόνων, επιτρέποντας την έγκαιρη ανίχνευση παθήσεων όπως ο καρκίνος, με ακρίβεια που συχνά ξεπερνά αυτή των ανθρώπων ειδικών [97]. Παράλληλα, αλγόριθμοι μηχανικής μάθησης αναλύουν μεγάλους όγκους ιατρικών δεδομένων για την πρόβλεψη κλινικών αποτελεσμάτων και την εξατομίκευση θεραπευτικών προσεγγίσεων [98].

Ο χρηματοπιστωτικός τομέας έχει υιοθετήσει εκτενώς την ΤΝ για την ενίσχυση της ασφάλειας και της αποτελεσματικότητας των συναλλαγών. Τα συστήματα ανίχνευσης απάτης χρησιμοποιούν προηγμένους αλγόριθμους για τον εντοπισμό ύποπτων συναλλαγών σε πραγματικό χρόνο, ενώ συστήματα αυτοματοποιημένης διαπραγμάτευσης αναλύουν τις αγορές και εκτελούν συναλλαγές με βάση προκαθορισμένα κριτήρια [99]. Επιπλέον, η ΤΝ χρησιμοποιείται στην αξιολόγηση πιστωτικού κινδύνου και την παροχή εξατομικευμένων χρηματοοικονομικών συμβουλών [100].

Η εφαρμογή της ΤΝ στη δημόσια διοίκηση στοχεύει στη βελτίωση της αποτελεσματικότητας των υπηρεσιών και της εξυπηρέτησης των πολιτών. Έξυπνα συστήματα υποστηρίζουν τη λήψη αποφάσεων σε τομείς όπως ο πολεοδομικός σχεδιασμός, η διαχείριση πόρων και η παροχή δημόσιων υπηρεσιών. Chatbots και συστήματα αυτοματοποιημένης εξυπηρέτησης διευκολύνουν την επικοινωνία με τους πολίτες, ενώ αλγόριθμοι ανάλυσης δεδομένων βοηθούν στην πρόβλεψη και αντιμετώπιση κοινωνικών προβλημάτων [101].

Στον τομέα της ασφάλειας, η ΤΝ χρησιμοποιείται για την ενίσχυση της προστασίας και την πρόληψη εγκληματικών ενεργειών. Συστήματα αναγνώρισης προσώπου και ανάλυσης συμπεριφοράς υποστηρίζουν την επιτήρηση δημόσιων χώρων, ενώ προηγμένα συστήματα κυβερνοασφάλειας προστατεύουν κρίσιμες υποδομές από ψηφιακές απειλές [102]. Ωστόσο, η εκτεταμένη χρήση τέτοιων τεχνολογιών εγείρει σημαντικά ζητήματα ιδιωτικότητας και πολιτικών ελευθεριών που απαιτούν προσεκτική στάθμιση και ρύθμιση [103].

Η ΤΝ διαδραματίζει καίριο ρόλο στην αντιμετώπιση περιβαλλοντικών προκλήσεων. Συστήματα πρόβλεψης και παρακολούθησης υποστηρίζουν την κλιματική έρευνα, την προστασία της βιοποικιλότητας και τη διαχείριση φυσικών πόρων. Αλγόριθμοι βελτιστοποίησης συμβάλλουν στη μείωση της ενεργειακής κατανάλωσης και των εκπομπών άνθρακα, ενώ συστήματα παρακολούθησης της ρύπανσης επιτρέπουν την έγκαιρη λήψη μέτρων προστασίας [104], [105].

2.3.2 Κοινωνικές Επιπτώσεις

Η διείσδυση της Τεχνητής Νοημοσύνης στον κοινωνικό ιστό επιφέρει βαθιές και πολύπλευρες αλλαγές που επηρεάζουν τόσο τη λειτουργία των θεσμών όσο και την καθημερινότητα των πολιτών [106].

Η επίδραση της ΤΝ στην αγορά εργασίας είναι πολυδιάστατη και προκαλεί σημαντικούς μετασχηματισμούς. Η αυτοματοποίηση ορισμένων εργασιακών καθηκόντων οδηγεί στην εξάλειψη παραδοσιακών θέσεων εργασίας, ενώ παράλληλα δημιουργεί νέες επαγγελματικές ευκαιρίες και ειδικότητες [107]. Η αλλαγή αυτή απαιτεί:

- Επανεκπαίδευση και αναβάθμιση δεξιοτήτων του εργατικού δυναμικού
- Προσαρμογή των εκπαιδευτικών συστημάτων στις νέες απαιτήσεις
- Ανάπτυξη πολιτικών για την αντιμετώπιση της εργασιακής ανασφάλειας [108]

Η εκτεταμένη χρήση συστημάτων ΤΝ εγείρει σοβαρά ζητήματα προστασίας της ιδιωτικότητας και των θεμελιωδών δικαιωμάτων. Η συλλογή και επεξεργασία μεγάλου όγκου προσωπικών δεδομένων δημιουργεί κινδύνους για:

- Την παραβίαση της προσωπικής ζωής
- Τη μη εξουσιοδοτημένη χρήση προσωπικών πληροφοριών

- Την ψηφιακή επιτήρηση και τον κοινωνικό έλεγχο [109]

Οι προκλήσεις αυτές απαιτούν την ανάπτυξη ισχυρών μηχανισμών προστασίας και την καλλιέργεια ψηφιακής παιδείας στους πολίτες [110].

Η ΤΝ επηρεάζει σημαντικά τη λειτουργία των δημοκρατικών θεσμών και τη διαδικασία λήψης αποφάσεων. Η χρήση αλγοριθμικών συστημάτων στη δημόσια διοίκηση και την πολιτική επικοινωνία έχει σημαντικές επιπτώσεις:

- Στη διαμόρφωση της κοινής γνώμης μέσω εξατομικευμένης πληροφόρησης
- Στη διαφάνεια και λογοδοσία των δημόσιων οργανισμών
- Στην ποιότητα του δημοκρατικού διαλόγου [111]

Παράλληλα, η ψηφιακή διακυβέρνηση προσφέρει ευκαιρίες για:

- Αυξημένη συμμετοχή των πολιτών στα κοινά
- Βελτιωμένη αποτελεσματικότητα των δημόσιων υπηρεσιών
- Καλύτερη διαχείριση των κοινωνικών πόρων [112]

Η διάχυση της ΤΝ μπορεί να επηρεάσει σημαντικά την κοινωνική συνοχή και τις υπάρχουσες ανισότητες. Παρατηρούνται φαινόμενα όπως:

- Το ψηφιακό χάσμα μεταξύ διαφορετικών κοινωνικών ομάδων
- Η ενίσχυση υφιστάμενων κοινωνικοοικονομικών ανισοτήτων
- Η δημιουργία νέων μορφών κοινωνικού αποκλεισμού [113]

Η αντιμετώπιση αυτών των προκλήσεων απαιτεί:

- Στοχευμένες πολιτικές κοινωνικής ένταξης
- Προγράμματα ψηφιακού αλφαριθμητισμού
- Μέτρα για την ισότιμη πρόσβαση στις νέες τεχνολογίες [114], [115]

2.3.3 Προκλήσεις και Κίνδυνοι

Παρά τις σημαντικές δυνατότητες που προσφέρει η Τεχνητή Νοημοσύνη, η ανάπτυξη και εφαρμογή της συνοδεύεται από σημαντικές προκλήσεις και κινδύνους που χρήζουν προσεκτικής αντιμετώπισης [116]. Οι προκλήσεις αυτές εκτείνονται από τεχνικά ζητήματα μέχρι κοινωνικές και ηθικές επιπτώσεις, απαιτώντας μια ολοκληρωμένη προσέγγιση για την αντιμετώπισή τους.

Η αλγοριθμική προκατάληψη αποτελεί μία από τις πιο σημαντικές προκλήσεις στην ανάπτυξη συστημάτων ΤΝ. Οι προκαταλήψεις μπορούν να προκύψουν από διάφορους παράγοντες, συμπεριλαμβανομένων των μεροληπτικών δεδομένων εκπαίδευσης που αντανακλούν υπάρχουσες κοινωνικές προκαταλήψεις, της ελλιπούς αντιπροσώπευσης συγκεκριμένων κοινωνικών ομάδων στα δεδομένα, και των προκατειλημμένων υποθέσεων στο σχεδιασμό των αλγορίθμων [117]. Οι επιπτώσεις αυτών των προκαταλήψεων μπορεί να είναι ιδιαίτερα σοβαρές, οδηγώντας σε διακρίσεις σε κρίσιμους τομείς όπως η πρόσβαση σε χρηματοπιστωτικές υπηρεσίες, οι ευκαιρίες απασχόλησης, και η ποινική δικαιοσύνη [60].

Η διαφάνεια των συστημάτων ΤΝ αποτελεί μια ακόμη κρίσιμη πρόκληση. Το πρόβλημα του "μαύρου κουτιού" και η περιορισμένη δυνατότητα επεξήγησης των αποφάσεων που

λαμβάνονται από συστήματα ΤΝ δυσχεραίνουν την κατανόηση και αξιολόγηση των αποτελεσμάτων, τον εντοπισμό και τη διόρθωση σφαλμάτων, καθώς και τη λογοδοσία σε περιπτώσεις εσφαλμένων αποφάσεων [63]. Για την αντιμετώπιση αυτών των προκλήσεων, αναπτύσσονται συνεχώς νέες μέθοδοι και τεχνικές για επεξηγήσιμη ΤΝ, δυνατότητα ανίχνευσης των αποφάσεων και διαφανή τεκμηρίωση των αλγοριθμικών διαδικασιών [2].

Η ασφάλεια των συστημάτων ΤΝ περιλαμβάνει πολλαπλές διαστάσεις που απαιτούν συστηματική προσέγγιση. Σε τεχνικό επίπεδο, η προστασία από κυβερνοεπιθέσεις, η αντιμετώπιση κακόβουλης χρήσης και η διασφάλιση της αξιοπιστίας των συστημάτων αποτελούν βασικές προτεραιότητες [118]. Παράλληλα, η λειτουργική ασφάλεια επικεντρώνεται στην πρόληψη ατυχημάτων και δυσλειτουργιών, τη διασφάλιση σταθερής απόδοσης και την προστασία από απρόβλεπτες συμπεριφορές [119]. Η αντιμετώπιση των προκλήσεων ασφαλείας απαιτεί την ανάπτυξη ισχυρών πρωτοκόλλων, τακτικούς ελέγχους και συνεχή επικαιροποίηση των μέτρων προστασίας [120].

Πέρα από τις άμεσες προκλήσεις, οι μακροπρόθεσμοι κίνδυνοι της ΤΝ απαιτούν ιδιαίτερη προσοχή και προνοητικό σχεδιασμό. Οι πιθανές επιπτώσεις στην ανθρώπινη αυτονομία και λήψη αποφάσεων, η αυξανόμενη εξάρτηση από συστήματα ΤΝ σε κρίσιμους τομείς, και τα ζητήματα ελέγχου και διακυβέρνησης προηγμένων συστημάτων ΤΝ αποτελούν σημαντικές προκλήσεις που απαιτούν προσεκτική εξέταση και αντιμετώπιση [121], [122].

3

Μεθοδολογία Έρευνας

Στο παρόν κεφάλαιο παρουσιάζεται αναλυτικά η μεθοδολογία που ακολουθήθηκε για τη διεξαγωγή της διττής συστηματικής βιβλιογραφικής ανασκόπησης αναφορικά με τα ηθικά ζητήματα της Τεχνητής Νοημοσύνης, καλύπτοντας δύο διακριτές χρονικές περιόδους: 2015-2019 και 2024. Με βάση το θεωρητικό πλαίσιο που αναπτύχθηκε στο προηγούμενο κεφάλαιο, διατυπώνονται τα κύρια ερευνητικά ερωτήματα που καθοδήγησαν την έρευνα και την επέκτασή της στην εποχή της Παραγωγικής ΤΝ. Στη συνέχεια, περιγράφεται λεπτομερώς η διαδικασία συλλογής των δεδομένων από τις επιλεγμένες βιβλιογραφικές πηγές για κάθε χρονική περίοδο, συμπεριλαμβανομένων των προσαρμοσμένων στρατηγικών αναζήτησης για τη μελέτη των νέων προκλήσεων που έχει φέρει η Παραγωγική ΤΝ. Παρουσιάζονται τα κριτήρια συμπερίληψης και αποκλεισμού των μελετών στην τελική ανάλυση, καθώς και η μεθοδολογική προσέγγιση για τη συγκριτική ανάλυση των ευρημάτων από τις δύο περιόδους. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση των περιορισμών και των δυσκολιών που προέκυψαν κατά τη διάρκεια της ερευνητικής διαδικασίας, τόσο στην αρχική έρευνα όσο και στην επέκτασή της, προσφέροντας έτσι μια ολοκληρωμένη εικόνα της μεθοδολογικής προσέγγισης που υιοθετήθηκε για την κατανόηση της εξέλιξης των ηθικών ζητημάτων στην ΤΝ διαχρονικά.

3.1 Ερευνητικά ερωτήματα

Η παρούσα συστηματική βιβλιογραφική ανασκόπηση καθοδηγείται από πέντε κύρια ερευνητικά ερωτήματα, τα οποία διαμορφώθηκαν με βάση τις ανάγκες διερεύνησης των ηθικών ζητημάτων που προκύπτουν από την ανάπτυξη και εφαρμογή της Τεχνητής Νοημοσύνης. Τα ερωτήματα αυτά είναι:

1. Τι πρέπει να γίνει για να διασφαλιστεί η ηθική χρήση της Τεχνητής Νοημοσύνης;
2. Ποιες οι προϋποθέσεις για ασφαλή χρήση της Τεχνητής Νοημοσύνης και ποια είναι τα χαρακτηριστικά ασφαλούς Τεχνητής Νοημοσύνης;
3. Μπορεί να θεωρηθεί η Τεχνητή Νοημοσύνη υπεύθυνη (ηθικός σχεδιασμός);
4. Πώς αντιμετωπίζει η Τεχνητή Νοημοσύνη ζητήματα ιδιωτικότητας;
5. Κρίνεται αναγκαία η ηθική και δεοντολογική καθοδήγηση σε ζητήματα Τεχνητής νοημοσύνης;

Σύμφωνα με τις κατευθυντήριες γραμμές για τη διεξαγωγή συστηματικών βιβλιογραφικών ανασκοπήσεων στη Μηχανική Λογισμικού, κάθε ερευνητικό ερώτημα εμπίπτει σε συγκεκριμένο τύπο ερώτησης. Το πρώτο ερώτημα εστιάζει στον "Προσδιορισμό των επιπτώσεων των τεχνολογιών στην αξιοπιστία, τις επιδόσεις και τα μοντέλα κόστους", καθώς διερευνά τις απαραίτητες ενέργειες για την εξασφάλιση ηθικής χρήσης της TN.

Το δεύτερο ερώτημα αφορά την "Αξιολόγηση της συχνότητας ή του ποσοστού ενός συντελεστή ανάπτυξης έργου", εξετάζοντας τις προϋποθέσεις και τα χαρακτηριστικά που καθιστούν την TN ασφαλή.

Το τρίτο ερώτημα σχετίζεται με τον "Προσδιορισμό του κόστους και των παραγόντων κινδύνου που σχετίζονται με μια τεχνολογία", διερευνώντας την υπευθυνότητα και τον ηθικό σχεδιασμό των συστημάτων TN.

Το τέταρτο ερώτημα εμπίπτει στην "Αξιολόγηση της επίδρασης μιας τεχνολογίας", εστιάζοντας στον τρόπο που η TN διαχειρίζεται ζητήματα ιδιωτικότητας.

Το πέμπτο ερώτημα αφορά την "Ανάλυση οφέλους από την αξιοποίηση συγκεκριμένων τεχνολογιών", εξετάζοντας την αναγκαιότητα ηθικής καθοδήγησης στην ανάπτυξη και χρήση της TN.

Τα ερευνητικά ερωτήματα δομήθηκαν επίσης με βάση τα κριτήρια PICOC (Population/Πληθυσμός, Intervention/Παρέμβαση, Comparison/Σύγκριση, Outcomes/Αποτελέσματα, Context/Συμπέρασμα):

- **Πληθυσμός:** Μηχανικοί λογισμικού, προγραμματιστές, DPOs (Data Protection Officers), testers, νομικοί και επιχειρήσεις/οργανισμοί που δραστηριοποιούνται στον τομέα της TN εντός της Ευρωπαϊκής Ένωσης.
- **Παρέμβαση:** Νομοθεσία (GDPR), διεθνή πρότυπα (27001, 02, 27701) και σχετικές τεχνολογίες.
- **Σύγκριση:** Μεμονωμένες νομοθεσίες κάθε χώρας και ξεχωριστές πολιτικές κάθε επιχείρησης.
- **Αποτελέσματα:** Δημιουργία ενός κοινού πλαισίου ηθικού σχεδιασμού και ανάπτυξης έργων Τεχνητής Νοημοσύνης.

3.2 Μέθοδος Συλλογής Δεδομένων

Η συλλογή των δεδομένων για την παρούσα συστηματική βιβλιογραφική ανασκόπηση πραγματοποιήθηκε σε δύο φάσεις, καλύπτοντας διαφορετικές χρονικές περιόδους.

▪ Φάση 1 (2015-2019)

Η αρχική συλλογή δεδομένων πραγματοποιήθηκε μέσω εκτενούς αναζήτησης σε επτά κύριες βάσεις δεδομένων και βιβλιοθήκες κατά το διάστημα Ιούλιος 2019 - Νοέμβριος 2019. Συγκεκριμένα, οι πηγές που χρησιμοποιήθηκαν περιλαμβάνουν:

Πηγή δεδομένων	Έγγραφα/Τεκμήρια
----------------	------------------

Ψηφιακή βιβλιοθήκη	Όνομα βάσης δεδομένων: IEEEExplore Στρατηγική αναζήτησης για την βάση: (machine OR robot) AND ethics, ai ethics, ai gdpr, ieeexplore autonomous AND (robots OR systems) AND ethics, cyberethics, roboethics Ημερομηνία αναζήτησης: 16/07/2019 - 21/08/2019 Διάστημα έρευνας: 36 ημέρες
Ψηφιακή βιβλιοθήκη	Όνομα βάσης δεδομένων: ACM Digital library Στρατηγική αναζήτησης για την βάση: (ai OR artificial intelligence OR robot) AND ethical AND issues, (ai OR artificial intelligence) AND ethical AND (design OR framework), (ethical OR moral) AND (ai OR artificial intelligence OR machine OR robot OR agent), ai AND gdpr, autonomous AND (robots OR systems) AND ethics, cyberethics, roboethics Ημερομηνία αναζήτησης: 16/08/2019 - 19/08/2019 Διάστημα έρευνας: 4 ημέρες
Βιβλιογραφική Βάση Δεδομένων	Όνομα βάσης δεδομένων: Google Scholar Στρατηγική αναζήτησης για την βάση: (law OR legal OR responsible OR moral OR ethical OR framework OR ethical design) AND ((safe OR safety) AND (policy OR use)) AND (robot OR machine OR ai OR artificial intelligence OR agent OR computer), (ethical OR framework OR ethical design OR legal OR moral) AND (artificial intelligence OR robot OR machine OR ai OR agent OR machine OR autonomous) AND (gdpr OR general data protection regulation) Ημερομηνία αναζήτησης: 05/11/2019 - 13/11/2019 Διάστημα έρευνας: 9 ημέρες
Βιβλιογραφική Βάση Δεδομένων	Όνομα βάσης δεδομένων: Citeseer library (CiteseerX) //τα πιο πρόσφατα είναι του 2015-2016 Στρατηγική αναζήτησης για την βάση: gdpr AND (artificial intelligence OR robot OR machine OR ai),

	(ethical OR framework OR ethical design OR legal OR moral) AND (artificial intelligence OR robot OR machine OR ai OR agent OR machine) AND (gdpr OR general data protection regulation) Ημερομηνία αναζήτησης: 04/10/2019 Διάστημα έρευνας: 1 ημέρα
Βιβλιογραφική και Τεχνική Βάση Δεδομένων	Όνομα βάσης δεδομένων: Inspec (https://digital-library.theiet.org/) //θεμα με τα πιο προσφατα Στρατηγική αναζήτησης για την βάση: (ethical OR moral) AND (ai OR artificial intelligence OR machine OR robot OR agent) Ημερομηνία αναζήτησης: 1/10/2019 Διάστημα έρευνας: 1 ημέρα
Πλατφόρμα επιστημονικής βιβλιογραφίας	Όνομα βάσης δεδομένων: ScienceDirect Στρατηγική αναζήτησης για την βάση: Ethical decision making, ai AND gdpr, cyberethics, roboethics, (law OR legal OR responsible) AND (robot OR machine OR ai OR artificial intelligence OR agent OR computer), (ethical OR framework OR ethical design OR legal OR moral) AND (artificial intelligence OR robot OR machine OR ai OR agent OR machine) AND (gdpr OR general data protection regulation) Ημερομηνία αναζήτησης: 25/10/2019 Διάστημα έρευνας: 1 ημέρα
Βιβλιογραφική Βάση Δεδομένων	Όνομα βάσης δεδομένων: SCOPUS Στρατηγική αναζήτησης για την βάση: (law OR legal OR responsible) AND (robot OR machine OR ai OR artificial intelligence OR agent OR computer), (ethical OR framework OR ethical design OR legal OR moral) AND (artificial intelligence OR robot OR machine OR ai OR agent OR machine) AND (gdpr OR general data protection regulation), roboethics, cyberethics Ημερομηνία αναζήτησης: 25/10/2019 - 26-10/2019 Διάστημα έρευνας: 2 ημέρες

Πίνακας 1. Περίληψη της έρευνας 2015-2019 σχετικά με τις στρατηγικές αναζήτησης ηθικής τεχνητής νοημοσύνης σε βάσεις δεδομένων.

Συνολικά, η διαδικασία συλλογής δεδομένων απέφερε 4.904 αρχεία, τα οποία μετά την απαλοιφή των διπλοτύπων και την αφαίρεση των κατεστραμμένων αρχείων περιορίστηκαν σε 4.080 μοναδικά

έγγραφα. Μετά την αφαίρεση των αρχείων αναφορών και των ξενόγλωσσων αρχείων (πλην αγγλικών), το τελικό σύνολο των προς ανάλυση εγγράφων διαμορφώθηκε σε 4.036 έγγραφα.

Η κατανομή των εγγράφων ανά πηγή έχει ως εξής:

- ACM Digital Library: 1.281 έγγραφα
- CiteSeerX: 98 έγγραφα
- Google Scholar: 1.605 έγγραφα
- IEEE Xplore: 563 έγγραφα
- Inspec: 686 έγγραφα
- Referex Engineering - ScienceDirect: 6 έγγραφα
- SCOPUS: 665 έγγραφα

Από το σύνολο των συλλεχθέντων εγγράφων, εντοπίστηκαν 179 έγγραφα που κρίθηκαν άμεσα σχετικά με τα ερευνητικά ερωτήματα της μελέτης, ενώ δύο μελέτες εμφανίστηκαν σε περισσότερες από μία πηγές δεδομένων.

▪ Φάση 2 (2024)

Η δεύτερη φάση της έρευνας πραγματοποιήθηκε στις ίδιες βάσεις δεδομένων κατά το διάστημα 1/1/2024 - 31/12/2024, με επικαιροποιημένες στρατηγικές αναζήτησης που συμπεριέλαβαν επιπλέον όρους σχετικούς με την Παραγωγική TN:

Πηγή δεδομένων	Έγγραφα/Τεκμήρια
Ψηφιακή βιβλιοθήκη	Όνομα βάσης δεδομένων: IEEEExplore Στρατηγική αναζήτησης για την βάση: "generative AI" AND ethics "large language models" AND ethics AI governance AND "generative AI" AI democratization "generative AI" AND bias deepfake AND ethics Ημερομηνία αναζήτησης: 30/01/2025 - 15/02/2025 Διάστημα έρευνας: 16 ημέρες
Ψηφιακή βιβλιοθήκη	Όνομα βάσης δεδομένων: ACM Digital library "generative AI" AND (ethics OR ethical OR moral)"LLM" AND (ethics OR ethical OR moral)"AI content generation" AND ethics"AI intellectual property" AND "generative AI" Ημερομηνία αναζήτησης: 30/01/2025 - 15/02/2025 Διάστημα έρευνας: 16 ημέρες

Βιβλιογραφική Βάση Δεδομένων	Όνομα βάσης δεδομένων: Google Scholar Στρατηγική αναζήτησης για την βάση: "generative AI" AND (ethics OR ethical OR moral)"large language models" AND ethics AND governance"AI content" AND (authenticity OR transparency)"generative AI" AND intellectual property Ημερομηνία αναζήτησης: 30/01/2025 - 15/02/2025 Διάστημα έρευνας: 16 ημέρες
Βιβλιογραφική Βάση Δεδομένων	Όνομα βάσης δεδομένων: Citeseer library (CiteseerX) Στρατηγική αναζήτησης για την βάση: "generative AI" AND ethics"AI governance" AND "content generation" Ημερομηνία αναζήτησης: 30/01/2025 - 15/02/2025 Διάστημα έρευνας: 16 ημέρες
Βιβλιογραφική και Τεχνική Βάση Δεδομένων	Όνομα βάσης δεδομένων: Inspec (https://digital-library.theiet.org/) //θέμα με τα πιο προσφατά Στρατηγική αναζήτησης για την βάση: "generative AI" AND (ethics OR governance)"AI content generation" AND ethics Ημερομηνία αναζήτησης: 30/01/2025 - 15/02/2025 Διάστημα έρευνας: 16 ημέρες
Πλατφόρμα επιστημονικής βιβλιογραφίας	Όνομα βάσης δεδομένων: ScienceDirect Στρατηγική αναζήτησης για την βάση: "generative AI" AND ethics "AI content" AND authenticity "large language models" AND governance "AI democratization" Ημερομηνία αναζήτησης: 30/01/2025 - 15/02/2025 Διάστημα έρευνας: 16 ημέρες
Βιβλιογραφική Βάση Δεδομένων	Όνομα βάσης δεδομένων: SCOPUS "generative AI" AND (ethics OR governance) "AI content generation" AND (authenticity OR transparency) "large language models" AND ethics "AI intellectual property" AND "generative AI" Ημερομηνία αναζήτησης: 30/01/2025 - 15/02/2025 Διάστημα έρευνας: 16 ημέρες

Πίνακας 2. Περίληψη της έρευνας 2024 σχετικά με τις στρατηγικές αναζήτησης ηθικής τεχνητής νοημοσύνης σε βάσεις δεδομένων.

Από το σύνολο των 4.128 εγγράφων που συλλέχθηκαν στη δεύτερη φάση, μετά την αφαίρεση διπλοτύπων και την εφαρμογή των κριτηρίων επιλογής/αποκλεισμού, 139 άρθρα κρίθηκαν σχετικά με τα ερευνητικά ερωτήματα της μελέτης.

3.3 Κριτήρια Επιλογής και Αποκλεισμού

Για τη διασφάλιση της ποιότητας και της καταλληλότητας των επιλεγμένων μελετών, καθορίστηκαν συγκεκριμένα κριτήρια συμπερίληψης και αποκλεισμού. Τα κριτήρια αυτά διαμορφώθηκαν με γνώμονα την εξασφάλιση της συνάφειας των μελετών με τα ερευνητικά ερωτήματα και την εγκυρότητα των ευρημάτων.

Κριτήρια Συμπερίληψης

1. Σχετικότητα με GDPR και Τεχνητή Νοημοσύνη
 - Μελέτες που λαμβάνουν υπόψη ή αναφέρουν τον GDPR
 - Έργα που πραγματεύονται ζητήματα τεχνητής νοημοσύνης
 - Μελέτες που στη βιβλιογραφία τους περιλαμβάνουν έγγραφα σχετικά με τον GDPR
2. Ηθική Διάσταση
 - Έγγραφα που εξετάζουν ηθικά ζητήματα σχετικά με τη χρήση της τεχνητής νοημοσύνης
 - Μελέτες που αναλύουν ηθικά θέματα που προκύπτουν από τη χρήση υποσυνόλων της ΤΝ (π.χ. μηχανική μάθηση)
3. Επιστημονικό Πεδίο
 - Μελέτες που εντάσσονται στον κλάδο της πληροφορικής ή συγγενικών επιστημονικών πεδίων

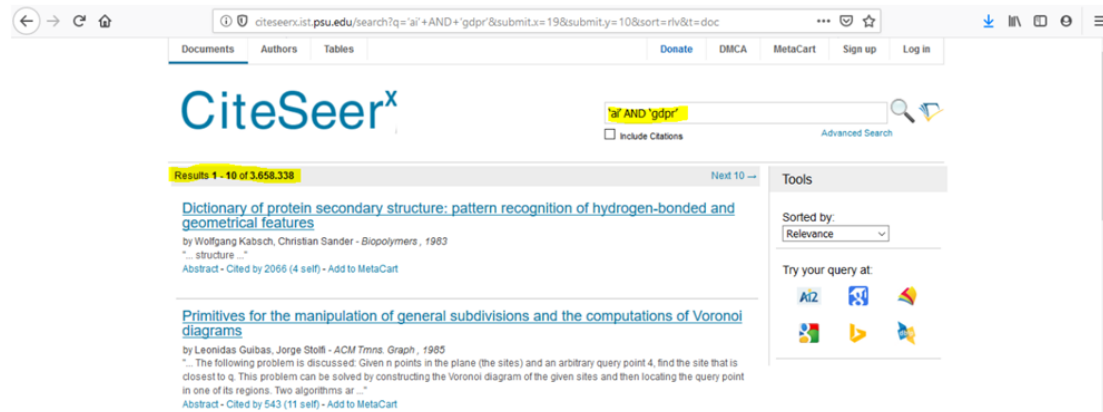
Κριτήρια Αποκλεισμού

1. Γλωσσικός Περιορισμός
 - Αποκλείστηκαν όλες οι μελέτες που δεν ήταν γραμμένες στην αγγλική γλώσσα
 - Η επιλογή της αγγλικής γλώσσας ως μοναδικής γλώσσας αποδοχής έγινε για να διασφαλιστεί η ομοιογένεια στην ανάλυση και την κατανόηση των κειμένων
2. Θεματικός Περιορισμός
 - Αποκλείστηκαν μελέτες που, παρότι μπορεί να αναφέρονταν στην τεχνητή νοημοσύνη, είχαν ως κύριο αντικείμενο άλλους επιστημονικούς κλάδους
 - Αυτός ο περιορισμός εξασφάλισε την εστίαση στις τεχνολογικές και ηθικές πτυχές της ΤΝ από την οπτική της πληροφορικής

Είναι σημαντικό να σημειωθεί ότι η αξιοπιστία των πηγών διασφαλίστηκε μέσω:

- Της χρήσης αναγνωρισμένων ακαδημαϊκών βάσεων δεδομένων
- Της πρόσβασης μέσω πανεπιστημιακών διαπιστευτηρίων
- Της επιλογής δημοσιεύσεων που έχουν υποστεί διαδικασία αξιολόγησης από ομότιμους κριτές

Τα παραπάνω κριτήρια εφαρμόστηκαν συστηματικά κατά τη διαδικασία επιλογής των μελετών, οδηγώντας στον εντοπισμό των σχετικών εγγράφων από το αρχικό σύνολο των 4.036 μελετών που πληρούσαν τις βασικές προϋποθέσεις γλώσσας και πρόσβασης.



Εικόνα 1. Αναζήτηση όρων χωρίς αποστροφές.

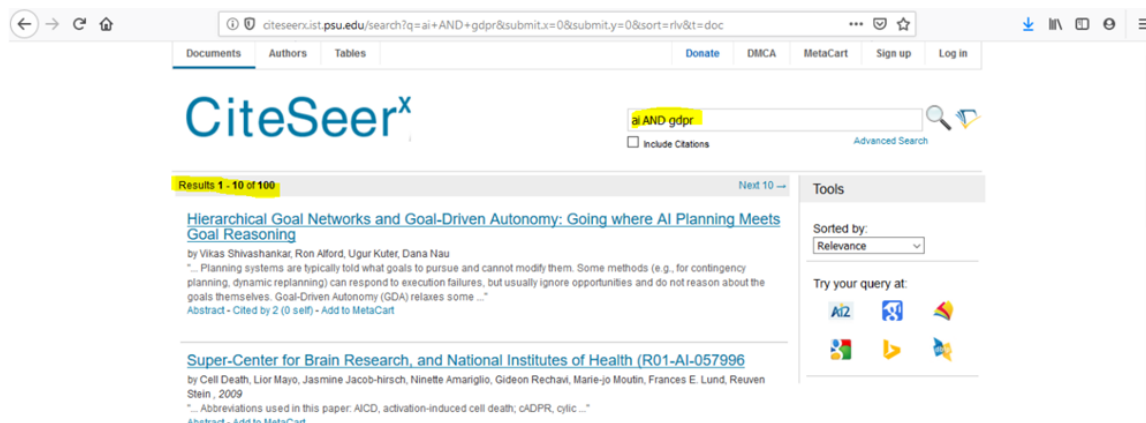
3.4 Περιορισμοί και Δυσκολίες

Κατά τη διεξαγωγή της συστηματικής βιβλιογραφικής ανασκόπησης, προέκυψαν διάφοροι περιορισμοί και δυσκολίες που επηρέασαν τη διαδικασία συλλογής και επεξεργασίας των δεδομένων.

3.4.1 Ζητήματα Αναζήτησης και Σχετικότητας

Μη-σχετικά Αποτελέσματα Αναζήτησης

- Στη βιβλιογραφική βάση δεδομένων CiteSeerX παρατηρήθηκαν σημαντικά προβλήματα με τους τελεστές αναζήτησης (AND, OR)
- Οι συνδυασμοί όρων δεν λειτουργούσαν αποτελεσματικά, ακόμη και με χρήση εισαγωγικών
- Η αναζήτηση βασιζόταν στα μεμονωμένα γράμματα αντί για ολόκληρες λέξεις, οδηγώντας



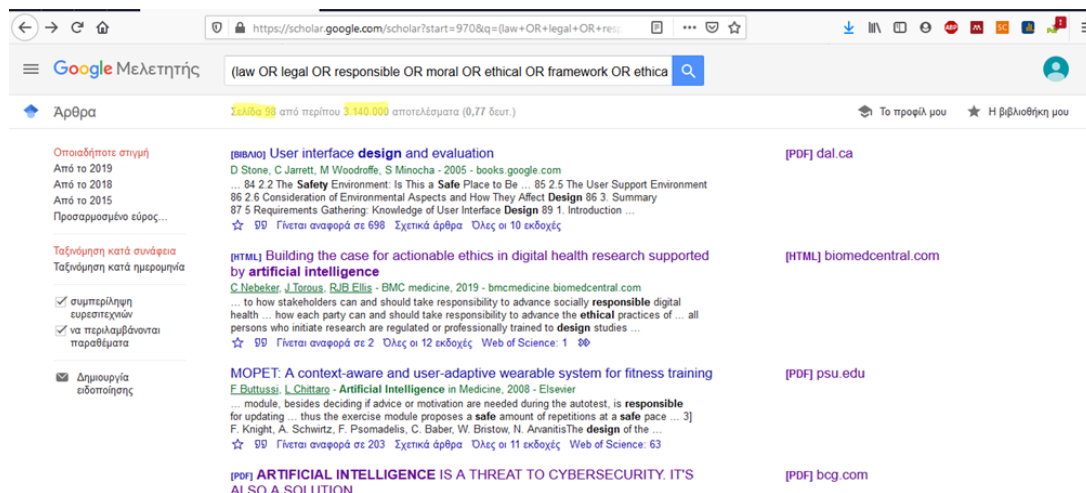
Εικόνα 2. Αναζήτηση όρων με αποστροφές για πιο συγκεκριμένα και λογικά λιγότερα σχετικά έγγραφα στα αποτελέσματα. Εμφάνιση εμφανώς περισσότερων μη-σχετικών αποτελεσμάτων.

σε άσχετα αποτελέσματα

- Παρά την εξειδίκευση της βάσης στην επιστήμη υπολογιστών, αποδείχθηκε ακατάλληλη για τους σκοπούς της συγκεκριμένης έρευνας

3.4.2 Τεχνικοί Περιορισμοί στη Διαχείριση Αναφορών

Google Scholar



Εικόνα 3. Αποτελέσματα ενενηκοστής όγδοης σελίδας.

- Απουσία δυνατότητας αυτοματοποιημένης εξαγωγής αποτελεσμάτων αναζήτησης σε μορφή συμβατή με προγράμματα διαχείρισης αναφορών
- Απαίτηση χειροκίνητης προσθήκης μεμονωμένων αναφορών, καθιστώντας τη διαδικασία εξαιρετικά χρονοβόρα για μεγάλο όγκο δεδομένων

CiteSeerX

- Πλήρης απουσία δυνατοτήτων εξαγωγής αρχείου αναφορών

Science Direct

- Περιορισμένη συμβατότητα των εξαγόμενων αρχείων με προγράμματα διαχείρισης βιβλιογραφίας

SCOPUS

- Περιορισμός στην εξαγωγή αποτελεσμάτων (μέχρι 500 έγγραφα ή επιλεγμένα <500)

3.4.3 Ζητήματα Πρόσβασης σε Έγγραφα

IEEEExplore

- Ασυνέπεια στις ενδείξεις προσβασιμότητας (λουκέτα)
- Περιορισμοί στον αριθμό λήψεων (παύση συστήματος μετά από 30 λήψεις)
- Απαίτηση αναμονής 1-1.5 ώρας λόγω "Excessive Downloading"

Science Direct

- Περιορισμένη πρόσβαση σε πλήρη κείμενα
- Διαθεσιμότητα μόνο βασικών πληροφοριών (περιεχόμενα, στοιχεία βιβλίου)

SCOPUS

- Απαίτηση επιβεβαίωσης πρόσβασης για έγγραφα IEEEExplore

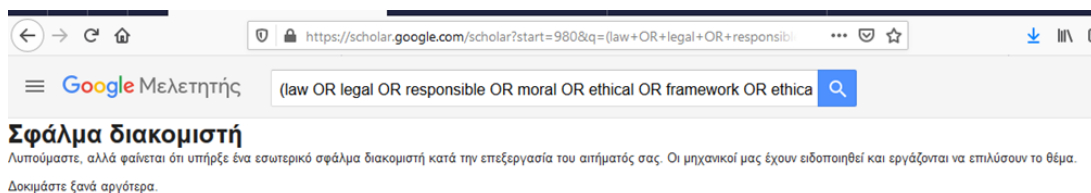
- Προβλήματα με ονοματοδοσία αρχείων κατά τη λήψη

Google Scholar

- Περιορισμοί στην πρόσβαση μετά από συγκεκριμένο αριθμό σελίδων (σφάλμα στην 99η σελίδα)
- Ζητήματα με την ονοματοδοσία των ληφθέντων αρχείων

3.4.4 Ζητήματα Πληρότητας Εγγράφων

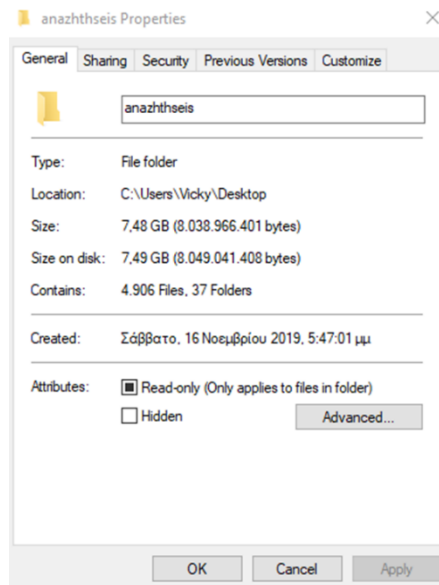
- Τα έγγραφα από τον ιστότοπο IET συχνά περιείχαν μόνο βασικές πληροφορίες (περιγραφή, ευρετήριο, περιεχόμενα)
- Έλλειψη πλήρους κειμένου σε πολλές περιπτώσεις
- Ανάγκη χρήσης εναλλακτικών μεθόδων για τη λήψη περιεχομένου (π.χ., χρήση πρόσθετων PDF)



Εικόνα 4. Το συγκεκριμένο σφάλμα εμφανιζόταν με την επιλογή της ενενηκοστής ένατης σελίδας. Το σφάλμα εμφανιζόταν επανειλημμένα στην επιλογή της ενενηκοστής ένατης ή εκατοστής σελίδας (δοκιμάστηκε σε διαφορετικά χρονικά διαστήματα η αναζήτηση κατ' επανάληψη για να διαπιστωθεί ότι το σφάλμα ήταν μόνιμο), με αποτέλεσμα να σταματήσει σε αυτό το σημείο η αναζήτηση.

3.4.5 Διαχείριση Δεδομένων

- Συνολικός όγκος αρχείων: 7,48 GB (4.901 έγγραφα αρχικά)
- Ανάγκη δημιουργίας αντιγράφων για διατήρηση της αρχικής αναζήτησης
- Απαίτηση πολλαπλών σταδίων επεξεργασίας για την απομάκρυνση διπλοτύπων και την εφαρμογή φίλτρων



Εικόνα 5. Στοιχεία αναζήτησης μετά την ολοκλήρωση της φάσης.

Οι παραπάνω περιορισμοί και δυσκολίες επηρέασαν σημαντικά τη χρονική διάρκεια και την αποτελεσματικότητα της ερευνητικής διαδικασίας. Ωστόσο, η συστηματική προσέγγιση και η χρήση εναλλακτικών μεθόδων όπου ήταν απαραίτητο, επέτρεψαν την επιτυχή ολοκλήρωση της συλλογής και επεξεργασίας των δεδομένων.

4

Ανάλυση και Αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζονται και αναλύονται τα ευρήματα της συστηματικής βιβλιογραφικής ανασκόπησης που διεξήχθη σε δύο χρονικές περιόδους (2015-2019 και 2024), σύμφωνα με τη μεθοδολογία που περιγράφηκε στο προηγούμενο κεφάλαιο. Η ανάλυση βασίζεται στη μελέτη σχετικών ερευνών που προέκυψαν από την αναζήτηση σε επτά βιβλιογραφικές βάσεις δεδομένων και παρέχει απαντήσεις στα πέντε βασικά ερευνητικά ερωτήματα που καθοδήγησαν την έρευνα. Η διττή χρονική εξέταση επιτρέπει τη συγκριτική ανάλυση της εξέλιξης του πεδίου, ιδιαίτερα μετά την έλευση της Γενετικής ΤΝ και των LLMs που έχουν μετασχηματίσει το τοπίο της ΤΝ. Συγκεκριμένα, εξετάζονται διαδοχικά τα ζητήματα της διασφάλισης ηθικής χρήσης της ΤΝ, των προϋποθέσεων ασφαλούς χρήσης, της υπευθυνότητας και του ηθικού σχεδιασμού, της προστασίας της ιδιωτικότητας, και της αναγκαιότητας ηθικής καθοδήγησης. Για κάθε ερευνητικό ερώτημα παρουσιάζονται τα κύρια ευρήματα ανά χρονική περίοδο, συνθέτονται οι διαφορετικές προσεγγίσεις που εντοπίστηκαν στη βιβλιογραφία, και αναδεικνύονται οι σημαντικότερες προκλήσεις και προτεινόμενες λύσεις. Η συγκριτική ανάλυση των ευρημάτων αναδεικνύει την εξέλιξη του πεδίου και τις νέες προκλήσεις που έχουν προκύψει με την ωρίμανση των τεχνολογιών ΤΝ.

4.1 Ευρήματα Περιόδου 2015-2019

Η περίοδος 2015-2019 αποτέλεσε μια κρίσιμη φάση για τη διαμόρφωση του πλαισίου ηθικής στην ΤΝ, καθώς η ραγδαία ανάπτυξη και υιοθέτηση συστημάτων τεχνητής νοημοσύνης σε διάφορους τομείς ανέδειξε την επιτακτική ανάγκη για ηθικές κατευθυντήριες γραμμές και ρυθμιστικά πλαίσια. Η ανάλυση των ευρημάτων αυτής της περιόδου αποκαλύπτει τις θεμελιώδεις προκλήσεις και τις αρχικές προσπάθειες της ερευνητικής κοινότητας να διαμορφώσει ένα συνεκτικό πλαίσιο για την ηθική ανάπτυξη και χρήση της ΤΝ.

4.1.1 Διασφάλιση ηθικής χρήσης ΤΝ

Η διασφάλιση της ηθικής χρήσης της Τεχνητής Νοημοσύνης αποτελεί θεμελιώδη πρόκληση που απαιτεί πολυεπίπεδη προσέγγιση. Με την αυξανόμενη διείσδυση της ΤΝ σε κρίσιμους τομείς όπως η υγειονομική περίθαλψη, η ποινική δικαιοσύνη και η εκπαίδευση, η ανάγκη για αποτελεσματικούς μηχανισμούς ελέγχου και ηθικής καθοδήγησης καθίσταται επιτακτική [3], [123]. Στην παρούσα

ενότητα εξετάζονται οι βασικοί πυλώνες που συμβάλλουν στην ηθική χρήση της ΤΝ, από την αλγοριθμική δικαιοσύνη και τις τεχνικές λύσεις μέχρι το ρυθμιστικό πλαίσιο και την ανθρωποκεντρική προσέγγιση. Ιδιαίτερη έμφαση δίνεται στην αντιμετώπιση των προκλήσεων που προκύπτουν από την αδιαφάνεια των αλγορίθμων, την προστασία των προσωπικών δεδομένων και την ανάγκη οικοδόμησης εμπιστοσύνης μεταξύ ανθρώπων και συστημάτων ΤΝ [8], [9].

4.1.1.1 Αλγοριθμική δικαιοσύνη και λογοδοσία

Η αλγοριθμική δικαιοσύνη και η λογοδοσία αποτελούν θεμελιώδεις προϋποθέσεις για την ηθική χρήση της Τεχνητής Νοημοσύνης. Η χρήση αδιαφανών αλγορίθμων μπορεί να είναι αποδεκτή μόνο όταν συνοδεύεται από ένα ολοκληρωμένο σύστημα περιορισμών που διασφαλίζουν την ηθική λειτουργία τους [8]. Ωστόσο, η δημιουργία αυτών των περιορισμών προϋποθέτει βαθιά κατανόηση των εισόδων, εξόδων, λειτουργίας και ορίων της ΤΝ, καθώς και των δεδομένων εκπαίδευσης που χρησιμοποιούνται - στοιχεία που συχνά δεν είναι εύκολα προσβάσιμα ή κατανοητά [8], [9].

Ιδιαίτερη έμφαση πρέπει να δοθεί στην ποιότητα των δεδομένων εκπαίδευσης, καθώς αυτά είναι καθοριστικής σημασίας για την απόδοση των αλγορίθμων μηχανικής μάθησης [123]. Η αυξημένη προτεραιοποίηση στην ηθική χρήση της ΤΝ μπορεί να επιτευχθεί μέσω της αμερόληπτης συμμετοχής πολλαπλών εταιρειών και της ευαισθητοποίησης για τις αλγοριθμικές διακρίσεις και τις προκαταλήψεις ταξινόμησης [3], [123]. Η δικαιοσύνη μπορεί επίσης να προωθηθεί μέσω της εκπαίδευσης των χρηστών και της τροποποίησης των διαδικασιών από συνειδητοποιημένες εταιρείες [9].

Η μείωση των ποσοστών σφαλμάτων σε υποομάδες απαιτεί έναν ουσιαστικό μετασχηματισμό στην ανάπτυξη, την εποπτεία και τη ρύθμιση της τεχνολογίας [123]. Οι αλγοριθμικοί έλεγχοι, αν και συχνά δεν οδηγούν σε άμεσες αλλαγές στα εμπορικά συστήματα, είναι απαραίτητοι για την αποκάλυψη προβληματικών προτύπων [3]. Η Information Commissioner's Office (ICO) διαδραματίζει σημαντικό ρόλο στην προώθηση ηθικών προσεγγίσεων και αλγοριθμικής διαφάνειας, συμβάλλοντας στην ανάπτυξη αξιόπιστων συστημάτων ΤΝ [123].

Ένα σημαντικό ζήτημα που χρήζει προσοχής είναι η λειτουργία των αλγορίθμων μηχανικής μάθησης ως "μαύρων κουτιών", καθιστώντας δύσκολη την κατανόηση και εξήγηση των αποφάσεών τους [3]. Το μέγεθος των τεχνικών εμποδίων που καθιστούν δύσκολη την εξήγηση των αυτόνομων αποφάσεων εξαρτάται από την πολυπλοκότητα του κάθε αλγορίθμου [9]. Πολλοί ερευνητές υποστηρίζουν ότι είναι σχεδόν αδύνατο να εξηγηθεί η λογική πίσω από μια απόφαση που λαμβάνεται από έναν αλγόριθμο λόγω της πολυπλοκότητας και της αδιαφάνειας που τους χαρακτηρίζει [3]. Μια προτεινόμενη λύση είναι η χρήση ενός δευτερεύοντος αλγορίθμου για την εξήγηση της λογικής του πρωτεύοντος, προσφέροντας μια πιο κατανοητή ερμηνεία [9], [123].

Η αύξηση των μεγάλων δεδομένων και της μηχανικής μάθησης έχει οδηγήσει σε σημαντική αύξηση των αλγοριθμικών αποφάσεων σε κρίσιμους τομείς όπως τα χρηματοπιστωτικά, η νομική, η ποινική δικαιοσύνη, η εκπαίδευση, η υγειονομική περίθαλψη και η πρόσληψη προσωπικού [3], [123]. Σε αυτό το πλαίσιο, το προτεινόμενο Κέντρο Δεοντολογίας και Καινοτομίας Δεδομένων αναμένεται να διαδραματίσει καθοριστικό ρόλο στην εποπτεία της ανάπτυξης των αλγορίθμων και των αποφάσεών τους [123].

4.1.1.2 Τεχνικές προσεγγίσεις για ηθική TN

Οι τεχνικές προσεγγίσεις για τη διασφάλιση της ηθικής χρήσης της TN περιλαμβάνουν διάφορες μεθοδολογίες, με την περιβάλλουσα να αναδεικνύεται ως μια σημαντική εναλλακτική στο πρόβλημα της αδιαφάνειας της TN. Η περιβάλλουσα προσφέρει ένα πλαίσιο για τη ρύθμιση και την ηθική πλοήγηση σε ένα περιβάλλον όπου η TN διαδραματίζει όλο και σημαντικότερο ρόλο [10]. Αναφέρεται στον καθορισμό σαφών ορίων στην εμβέλεια και λειτουργία των συστημάτων TN, επιτρέποντας στους αδιαφανείς αλγόριθμους να λειτουργούν αποτελεσματικά εντός προκαθορισμένων πλαισίων [11].

Για την επιτυχή εφαρμογή της περιβάλλουσας, απαιτείται λεπτομερής γνώση σχετικά με τα δεδομένα εκπαίδευσης, τις εισόδους, τις εξόδους, τη λειτουργία και τα όρια του συστήματος TN [10], [124]. Ωστόσο, η προσέγγιση αυτή εγείρει προβληματισμούς σχετικά με την πιθανή μείωση της ανθρώπινης αυτονομίας, καθώς οι ίδιες οι περιβάλλουσες μπορεί να ενσωματώνουν ανήθικες παραμέτρους [11].

Παράλληλα, τα συστήματα ευφυούς βιντεοεπιτήρησης αποτελούν χαρακτηριστικό παράδειγμα όπου η διαφάνεια είναι κρίσιμη για την ηθική χρήση της TN [12]. Η διαφάνεια σε αυτά τα συστήματα επιτρέπει στα ενδιαφερόμενα μέρη να κατανοήσουν τη λογική πίσω από συγκεκριμένες αποφάσεις ή ενέργειες [12], [125], [126]. Για την επίτευξη διαφάνειας στα συστήματα IVS, είναι απαραίτητη η υιοθέτηση συγκεκριμένων προτύπων σχεδιασμού και η χρήση προηγμένων τεχνικών μηχανικής μάθησης [125].

Η οικοδόμηση εμπιστοσύνης στα σύνθετα συστήματα TN μπορεί να επιτευχθεί μέσω της χρήσης κατανεμημένων συστημάτων πρόβλεψης, τα οποία παρέχουν προβλέψεις και εξηγήσεις με αποκεντρωμένο, ασφαλή και αξιόπιστο τρόπο [127]. Η χρήση τεχνολογιών όπως το blockchain και τα έξυπνα συμβόλαια μπορεί να ενισχύσει την ασφάλεια και την αξιοπιστία των συστημάτων TN, παρέχοντας ένα δημόσια προσβάσιμο ψηφιακό μητρώο που διευκολύνει τη συνεργατική λήψη αποφάσεων [28].

Ωστόσο, παραμένουν σημαντικές προκλήσεις, όπως η ελαχιστοποίηση της ανθρώπινης παρέμβασης στην επικύρωση των εξηγήσεων και η διασφάλιση της λειτουργίας σε πραγματικό χρόνο [127]. Επιπλέον, η ανάγκη για μετρικές ή δείκτες πιστότητας γίνεται ιδιαίτερα σημαντική σε περιπτώσεις όπου δεν υπάρχει ανθρώπινη εποπτεία, ειδικά όταν παρατηρούνται ασυνεπείς εξηγήσεις για την ίδια πρόβλεψη [28].

4.1.1.3 Ρυθμιστικό και κανονιστικό πλαίσιο

Το μη δεσμευτικό δίκαιο (soft law) αναδεικνύεται ως ένα σημαντικό εργαλείο για την ενίσχυση της εμπιστοσύνης στις νέες τεχνολογίες TN [13], [128]. Συνδυάζοντας τη σαφήνεια των διεθνών νομικών πράξεων με την ευελιξία των μη δεσμευτικών πράξεων, προσφέρει ένα "συναινετικό εργαλείο με διεθνές πεδίο εφαρμογής" [13]. Το soft law μπορεί να λάβει διάφορες μορφές, όπως κατευθυντήριες γραμμές, κώδικες δεοντολογίας, πρότυπα και μηχανισμούς πιστοποίησης, ενώ η εφαρμογή του μπορεί να επιτευχθεί είτε μέσω διοικητικών διαδικασιών είτε μέσω πολιτικής επιρροής [13].

Η Ευρωπαϊκή Ένωση έχει αναλάβει ηγετικό ρόλο στη διαμόρφωση του ρυθμιστικού πλαισίου για την TN, θεσπίζοντας νέους κανόνες για την ενίσχυση της οικονομίας των δεδομένων

και την ανάπτυξη διασυνοριακών αυτόνομων συστημάτων [15]. Η πολιτική της ΕΕ για την ΤΝ ξεχωρίζει τόσο για την ανάγκη καθορισμού του ρόλου της ΕΕ στη χάραξη πολιτικής όσο και για το διεθνές πλαίσιο ενός αγώνα δρόμου για την ΤΝ [16].

Τα συστήματα ΤΝ πρέπει να συμμορφώνονται με συγκεκριμένες ηθικές αρχές και κατευθυντήριες γραμμές [14]. Η Ευρωπαϊκή προσέγγιση δίνει έμφαση στην ανάγκη για έναν ηθικό σκοπό που βασίζεται σε θεμελιώδη δικαιώματα, κοινωνικές αξίες και ηθικές αρχές, συμπεριλαμβανομένων των αρχών της μη κακομεταχείρισης, της φιλικότητας προς το περιβάλλον, της διατήρησης της ανθρώπινης δράσης, της δικαιοσύνης και της αμεροληψίας [14], [129].

Οι κατευθυντήριες γραμμές δεοντολογίας της ΕΕ στοχεύουν στην παροχή οδηγιών για προγραμματιστές, πωλητές, διανομείς και οργανισμούς ΤΝ, διασφαλίζοντας τη συμμόρφωση με τις ηθικές αρχές και τα δικαιώματα των χρηστών [15], [16]. Η ομάδα εμπειρογνομόνων υψηλού επιπέδου (AI HLEG) υποστηρίζει μια ηθική προσέγγιση στην τεχνητή νοημοσύνη, βασισμένη στα θεμελιώδη δικαιώματα της ΕΕ [16], τονίζοντας τη σημασία του σεβασμού της ανθρώπινης αξιοπρέπειας και προωθώντας μια ανθρωποκεντρική προσέγγιση στην ΤΝ [16], [130].

Η ρύθμιση των νέων και συνεχώς εξελισσόμενων τεχνολογιών ΤΝ αποτελεί σημαντική πρόκληση για τις ρυθμιστικές αρχές παγκοσμίως [2]. Οι κανονισμοί πρέπει να εξισορροπούν την ανάγκη για ανάπτυξη με την προστασία των πολιτών, καθορίζοντας κριτήρια όπως η αναγκαιότητα, η διαφάνεια, η αποτελεσματικότητα και η νομιμότητα [2]. Το δίλημμα του Collingridge προσθέτει μια επιπλέον διάσταση δυσκολίας, καθώς οι τεχνολογίες είναι δύσκολο να ρυθμιστούν αρχικά λόγω περιορισμένης γνώσης, αλλά γίνονται ακόμη δυσκολότερες να ελεγχθούν μόλις εδραιωθούν [2].

4.1.1.4 Προκλήσεις και κίνδυνοι

Η έλλειψη διαφάνειας στη λήψη αποφάσεων από συστήματα ΤΝ αποτελεί μία από τις σημαντικότερες προκλήσεις, θέτοντας σε κίνδυνο τις θεμελιώδεις αρχές που διέπουν τη χρήση τους [17]. Η αυξανόμενη πίεση από το κοινό για την παροχή εξηγήσεων, ιδιαίτερα για αποφάσεις που επηρεάζουν αρνητικά τα συμφέροντα ιδιωτών και μικρών επιχειρήσεων, καθιστά επιτακτική την ανάγκη για μεγαλύτερη διαφάνεια [17].

Στο πλαίσιο της συμμόρφωσης με τον GDPR, αναδεικνύονται σημαντικές προκλήσεις στη διασφάλιση της δικαιοσύνης στη χρήση της ΤΝ [4], [5]. Παρότι η ΤΝ θα μπορούσε να είναι πιο αμερόληπτη από τους ανθρώπους στη λήψη αποφάσεων, παραμένουν σοβαρές ανησυχίες σχετικά με τις προκαταλήψεις και την υποκειμενικότητα [4]. Οι απαιτήσεις του GDPR για τη λήψη συγκατάθεσης των καταναλωτών πριν τη χρήση των πληροφοριών τους μπορεί να λειτουργήσουν ως εμπόδιο στην εξατομίκευση και τη συμπεριφορική στόχευση [5].

Η ρύθμιση των νέων τεχνολογιών παίρνει διάφορες μορφές, συμπεριλαμβανομένων εγγενών και φυσικών ελέγχων, αυτοελέγχου, διαφόρων επιπέδων "μαλακών" κοινοτικών ελέγχων, καθώς και ρύθμισης από την υποδομή [17]. Σε ορισμένες δικαιοδοσίες, η "αρχή της προφύλαξης" που εφαρμόζεται ήδη στην περιβαλλοντική νομοθεσία προτείνεται να επεκταθεί στις περιπτώσεις όπου η ΤΝ ενδέχεται να έχει σημαντικό αντίκτυπο [17].

Η συζήτηση σχετικά με το εάν ο GDPR περιλαμβάνει το δικαίωμα λήψης εξηγήσεων για συγκεκριμένες αυτοματοποιημένες διαδικασίες λήψης αποφάσεων παραμένει ενεργή [2]. Οι ρυθμιστικές αρχές καλούνται να επανεξετάσουν το άρθρο 22 του GDPR, το οποίο επιτρέπει στους

πολίτες να μην υπόκεινται σε αυτοματοποιημένη λήψη αποφάσεων, καθώς θα μπορούσε να επηρεάσει αρνητικά τόσο τους καταναλωτές όσο και τη βιομηχανία [131].

Η μη συμμόρφωση με τον GDPR μπορεί να οδηγήσει σε υψηλά πρόστιμα, γεγονός που αποτελεί ιδιαίτερη πρόκληση για τις μικρές και μεσαίες επιχειρήσεις [5], ενώ η συμμόρφωση είναι απαραίτητη για διεθνείς εμπόρους που επικοινωνούν με πολίτες της ΕΕ [131]. Ωστόσο, οι κανονισμοί του GDPR μπορούν να μετατραπούν σε ευκαιρία για τη βελτίωση της εμπιστοσύνης προς την ΤΝ μέσω διαφανών πολιτικών δεδομένων [4].

Υπάρχει επείγουσα ανάγκη για έρευνα στη διαμόρφωση υπεύθυνων στρατηγικών ΤΝ που θα επιτρέψουν στις επιχειρήσεις να αξιοποιήσουν την ΤΝ αποτελεσματικά και ηθικά [4]. Οι εργαζόμενοι θα πρέπει να εκπαιδεύονται και να ενημερώνονται για τις απειλές και τις διαδικασίες ασφαλείας, ενώ οι διαχειριστές και οι ιδιοκτήτες συστημάτων θα πρέπει να ευθυγραμμίζουν τις πολιτικές και τα πρότυπα με τις απαιτήσεις συμμόρφωσης και να διενεργούν τακτικούς ελέγχους [4], [131].

4.1.1.5 Ανθρωποκεντρική προσέγγιση

Η αλληλεπίδραση μεταξύ ανθρώπων και αλγορίθμων πρέπει να σχεδιάζεται με τρόπο που να επιτρέπει την κατανόηση και την αμφισβήτηση των αλγοριθμικών αποφάσεων, ακολουθώντας το πλαίσιο της δικαιοσύνης, της λογοδοσίας και της διαφάνειας [6]. Η αποτελεσματική αλγοριθμική διαφάνεια πρέπει να υπερβαίνει την απλή πρόσβαση στον πηγαίο κώδικα, παρέχοντας ουσιαστική δυνατότητα κατανόησης και αξιολόγησης των αποφάσεων [6].

Η εκπαίδευση της επόμενης γενιάς επιστημόνων μηχανικής μάθησης είναι ζωτικής σημασίας για την εφαρμογή βέλτιστων πρακτικών στην αξιολόγηση της ΤΝ [6], [7]. Αυτό περιλαμβάνει τη θέσπιση αυστηρότερων κριτηρίων αξιολόγησης, τη δημοσίευση του πηγαίου κώδικα και των δεδομένων, καθώς και τη δημιουργία προβλημάτων αναφοράς για συγκεκριμένους τομείς, ώστε να παρέχεται ένα πιο αμερόληπτο σύστημα σύγκρισης αλγορίθμων ΤΝ [6], [7].

Ιδιαίτερη προσοχή απαιτείται στον τομέα της υγειονομικής περίθαλψης, όπου η σύγχρονη μηχανική μάθηση βασίζεται κυρίως σε προσεγγίσεις "μαύρου κουτιού" [6], [7]. Η ερμηνευσιμότητα της μηχανικής μάθησης και της ΤΝ είναι κρίσιμη για την επιτυχία τους στην υγειονομική περίθαλψη, ενώ δεοντολογικά και ρυθμιστικά ζητήματα περιορίζουν τη διαθεσιμότητα ιατρικών δεδομένων [6], [7].

Η έλλειψη διαφάνειας στα μοντέλα μηχανικής μάθησης μπορεί να παρεμποδίσει τη λογοδοσία και να επιδεινώσει τις ασυμμετρίες ισχύος μεταξύ εταιρειών/κυβερνήσεων και πολιτών [6], [7]. Η διαφάνεια πρέπει να εξετάζεται σε τρία επίπεδα: του συνολικού μοντέλου, των μεμονωμένων στοιχείων και των συγκεκριμένων αλγορίθμων [6], [7]. Η αδιαφάνεια στη λήψη αλγοριθμικών αποφάσεων μπορεί να είναι σκόπιμη, να οφείλεται σε έλλειψη τεχνικής παιδείας ή να είναι εγγενής σε ορισμένες μεθόδους μηχανικής μάθησης [6], [7].

Τέλος, είναι απαραίτητο να διερευνηθεί ο αντίκτυπος της ΤΝ στην ανθρώπινη νοημοσύνη και στις ανθρώπινες σχέσεις, με ιδιαίτερη έμφαση στις επιπτώσεις στα παιδιά [6], [7]. Οι δεοντολογικές αρχές και η ποικιλομορφία πρέπει να ενσωματώνονται στην ανάπτυξη συστημάτων μηχανικής νοημοσύνης, ενώ η αλγοριθμική διαφάνεια πρέπει να εξετάζεται σε διάφορους τομείς εφαρμογών, με την ερμηνευσιμότητα να είναι ιδιαίτερα κρίσιμη στις εφαρμογές υγειονομικής περίθαλψης [7].

4.1.2 Προϋποθέσεις ασφαλούς χρήσης ΤΝ

Μετά τη διερεύνηση των τρόπων διασφάλισης της ηθικής χρήσης της ΤΝ, είναι σημαντικό να εξεταστούν οι συγκεκριμένες προϋποθέσεις και τα χαρακτηριστικά που καθιστούν ένα σύστημα ΤΝ ασφαλές. Η ασφαλής χρήση της Τεχνητής Νοημοσύνης στηρίζεται σε πολλαπλά θεμέλια, από τις αρχές της υπεύθυνης ΤΝ [18] μέχρι τους μηχανισμούς προστασίας και ασφάλειας [19], [132]. Η ενότητα αυτή εξετάζει τις βασικές προϋποθέσεις και τα χαρακτηριστικά που πρέπει να πληρούνται για την ασφαλή ανάπτυξη και εφαρμογή συστημάτων ΤΝ, δίνοντας έμφαση στη διαφάνεια των διαδικασιών, την προστασία των δεδομένων, και την αξιοπιστία των συστημάτων.

4.1.2.1 Θεμελιώδεις αρχές υπεύθυνης ΤΝ

Η υπεύθυνη τεχνητή νοημοσύνη βασίζεται σε τρεις θεμελιώδεις πυλώνες που διασφαλίζουν την ασφαλή της χρήση: την υπευθυνότητα, τη διαφάνεια και την ενσωμάτωση δικαιοσύνης και ηθικής [18]. Η υπευθυνότητα αναφέρεται τόσο στο ρόλο των ανθρώπων όσο και στις ικανότητες των συστημάτων ΤΝ, διασφαλίζοντας ότι οι αποφάσεις που λαμβάνονται είναι σύμφωνες με τις κοινωνικές αξίες και προσδοκίες [18].

Ο Χάρτης της ΤΝ καθορίζει πέντε θεμελιώδεις αρχές που πρέπει να διέπουν την ανάπτυξη και χρήση συστημάτων ΤΝ: το σεβασμό των θεμελιωδών δικαιωμάτων, την απαγόρευση των διακρίσεων, την ποιότητα και ασφάλεια των εργαλείων, τη διαφάνεια και αμεροληψία, καθώς και τη χρήση υπό έλεγχο [133]. Αυτές οι αρχές δεν αποτελούν απλώς θεωρητικές κατευθύνσεις, αλλά πρακτικές προϋποθέσεις για την ασφαλή εφαρμογή της ΤΝ.

Η δυνατότητα επαλήθευσης και επικύρωσης των συστημάτων ΤΝ αποτελεί επίσης κρίσιμο στοιχείο για τη διασφάλιση της ασφαλούς λειτουργίας τους, ιδιαίτερα όταν τα συστήματα εφαρμόζονται σε πραγματικές συνθήκες [20]. Για τα κρίσιμα συστήματα ασφαλείας, η συμμόρφωση με δημοσιευμένα πρότυπα μπορεί να αποτελεί νομική απαίτηση, ενώ οι ρυθμιστικοί φορείς, σε συνεργασία με το κοινό, οφείλουν να εξασφαλίζουν την αξιοπιστία των ρυθμιστικών διαδικασιών [20].

Η έμφαση στην ποιότητα των δεδομένων και στην αλγοριθμική δικαιοσύνη είναι εξίσου σημαντική για την ασφαλή χρήση της ΤΝ. Ένα ποικίλο σύνολο αλγορίθμων μπορεί να συμβάλει στη μείωση των κινδύνων που προκύπτουν από την παραβίαση ενός μόνο κύριου αλγορίθμου [134], ενώ η συνεχής παρακολούθηση και αξιολόγηση των συστημάτων είναι απαραίτητη για τη διασφάλιση της αποτελεσματικότητας και της ασφάλειάς τους.

4.1.2.2 Μηχανισμοί ασφάλειας και προστασίας

Η ασφάλεια και προστασία των συστημάτων ΤΝ απαιτεί ολοκληρωμένους μηχανισμούς που διασφαλίζουν τόσο την τεχνική αρτιότητα όσο και την προστασία των δεδομένων. Όταν πρόκειται για εφαρμογές κρίσιμες για την ασφάλεια, όπως η αυτόνομη οδήγηση και ο ιατρικός τομέας, οι απαιτήσεις για αξιοπιστία και ασφάλεια είναι ιδιαίτερα αυστηρές [24].

Ένας βασικός μηχανισμός ασφαλείας είναι η δυνατότητα παρέμβασης και ελέγχου από ανθρώπους. Για παράδειγμα, σε συστήματα που αλληλεπιδρούν με χρήστες, η ύπαρξη ενός "κουμπιού επαναφοράς" επιτρέπει την άμεση διακοπή ή αναπροσαρμογή της λειτουργίας όταν εντοπίζονται προβλήματα ή ανεπιθύμητες συμπεριφορές [135]. Παράλληλα, είναι απαραίτητη η

θέσπιση αυστηρών ορίων και περιορισμών, ιδιαίτερα για ευαίσθητα προϊόντα και υπηρεσίες ή για την προστασία ευάλωτων ομάδων [135].

Η ασφάλεια των δεδομένων αποτελεί επίσης κρίσιμο στοιχείο, με την εφαρμογή προηγμένων τεχνικών κρυπτογράφησης και στρατηγικών διαχείρισης [136]. Ο δημόσιος τομέας, ειδικότερα, οφείλει να επενδύει σε ισχυρές τεχνικές υποδομές και να εφαρμόζει συστηματικούς ελέγχους για την εξέταση κρίσιμων παραμέτρων ασφαλείας [136].

Η ανάπτυξη συστημάτων TN απαιτεί επίσης την εφαρμογή της αρχής "ασφάλεια μέσω σχεδιασμού" (security by design), όπου οι μηχανισμοί προστασίας ενσωματώνονται εξ αρχής στην αρχιτεκτονική του συστήματος [137]. Αυτή η προσέγγιση συνδυάζεται με την "ιδιωτικότητα μέσω σχεδιασμού" (privacy by design) και την "αυτονομία μέσω σχεδιασμού" (autonomy by design), δημιουργώντας ένα ολοκληρωμένο πλαίσιο προστασίας [137].

Για την αντιμετώπιση των κυβερνοαπειλών, είναι απαραίτητη η ανάπτυξη ανθεκτικών συστημάτων που μπορούν να προσαρμόζονται και να αντιδρούν σε νέες απειλές [138]. Η διασφάλιση της διαλειτουργικότητας, η τυποποίηση και η πιστοποίηση των συστημάτων, καθώς και η αποτελεσματική ανταλλαγή πληροφοριών για απειλές, συμβάλλουν στην ενίσχυση της συνολικής ασφάλειας [137].

4.1.2.3 Διαφάνεια και επεξηγησιμότητα συστημάτων

Η διαφάνεια και εξηγησιμότητα αποτελούν θεμελιώδεις προϋποθέσεις για την ασφαλή λειτουργία συστημάτων TN, ιδιαίτερα σε εφαρμογές που επηρεάζουν σημαντικά την ανθρώπινη ζωή. Ενώ η διαφάνεια δεν αποτελεί απαραίτητη προϋπόθεση για καταναλωτικά προϊόντα ή απλές διαδικτυακές υπηρεσίες, είναι κρίσιμη για εφαρμογές όπως η αυτόνομη οδήγηση και η ιατρική διάγνωση [21].

Οι εξηγήσεις που παρέχει ένα σύστημα TN πρέπει να είναι κατανοητές από τους ανθρώπους που τις χρειάζονται, είτε πρόκειται για τελικούς χρήστες, δικαστές ή ρυθμιστικές αρχές [25]. Ωστόσο, η επίτευξη αυτού του στόχου παρουσιάζει σημαντικές προκλήσεις, ιδιαίτερα όταν χρησιμοποιούνται σύνθετες τεχνολογίες όπως τα νευρωνικά δίκτυα, όπου η παροχή εξηγήσεων μπορεί να είναι δύσκολη ή αδύνατη [25].

Το ζήτημα της διαφάνειας διακρίνεται σε δύο τύπους:

- Η εκ των προτέρων διαφάνεια (prospective transparency), που αφορά την κατανόηση του τρόπου λειτουργίας του συστήματος πριν τη χρήση του
- Η εκ των υστέρων διαφάνεια (retrospective transparency), που επιτρέπει την ανάλυση και κατανόηση των αποφάσεων μετά τη λήψη τους [25]

Για την αντιμετώπιση του προβλήματος του "μαύρου κουτιού" στα συστήματα TN, έχει αναπτυχθεί ο τομέας της εξηγήσιμης τεχνητής νοημοσύνης (Explainable Artificial Intelligence – XAI). Παρά τη σημαντική πρόοδο, οι υπάρχουσες μέθοδοι επεξήγησης παραμένουν περιορισμένες, ιδιαίτερα στην ικανότητά τους να παρουσιάζουν τις σχέσεις μεταξύ των χαρακτηριστικών εισόδου σε υψηλό επίπεδο αφαίρεσης [21]. Για το λόγο αυτό, αναπτύσσονται μετα-εξηγήσεις που συγκεντρώνουν πληροφορίες από εργαλεία χαμηλού επιπέδου για να παρέχουν πιο ολοκληρωμένη κατανόηση [21].

Η διαφάνεια δεν περιορίζεται μόνο στον τεχνικό τομέα αλλά επεκτείνεται και στη διοικητική διαδικασία. Οι ρυθμιστικοί φορείς, σε συνεργασία με το κοινό, πρέπει να εξασφαλίζουν

τη διαφάνεια και την εμπιστοσύνη στην αξιοπιστία των ρυθμιστικών διαδικασιών [20], ενώ οι επεξηγήσεις που παρέχονται πρέπει να είναι επαληθεύσιμες και να επιτρέπουν τον εντοπισμό τυχόν προβληματικών συμπεριφορών [21].

4.1.2.4 Προετοιμασία για το μέλλον της TN

Οι ψηφιακές τεχνολογίες, και ειδικότερα η Τεχνητή Νοημοσύνη, αναπτύσσονται με εξαιρετικά γρήγορους ρυθμούς, επιφέροντας σημαντικές αλλαγές στην κοινωνία και την οικονομία [22]. Αυτός ο ταχύτατος ρυθμός ανάπτυξης απαιτεί ολοκληρωμένη προετοιμασία σε πολλαπλά επίπεδα, με έμφαση στην εκπαίδευση και κατάρτιση του ανθρώπινου δυναμικού για τις θέσεις εργασίας του μέλλοντος [22].

Η εκπαίδευση και κατάρτιση των πολιτών, ειδικά όσον αφορά τις νέες τεχνολογικές δεξιότητες, καθίσταται ζωτικής σημασίας για τη διατήρηση της ανταγωνιστικότητας και της οικονομικής ευημερίας [22]. Παράλληλα, η παροχή βοήθειας στους εργαζόμενους κατά τη μετάβαση σε νέα επαγγέλματα και η ενδυνάμωσή τους είναι ουσιώδεις προϋποθέσεις για να διασφαλιστεί μια ανάπτυξη που θα ωφελήσει ευρέως την κοινωνία [22].

Ιδιαίτερη σημασία έχει η προετοιμασία για την αντιμετώπιση των προκλήσεων που φέρνει η αυξανόμενη αυτοματοποίηση. Η υιοθέτηση της TN, ενώ μπορεί να αυξήσει σημαντικά την παραγωγικότητα, ενέχει τον κίνδυνο διεύρυνσης του ψηφιακού χάσματος και δημιουργίας έντονου ανταγωνισμού για την προσέλκυση εξειδικευμένου προσωπικού [22]. Για την αντιμετώπιση αυτών των προκλήσεων, απαιτείται η ανάπτυξη προγραμμάτων επανεκπαίδευσης και η δημιουργία νέων εκπαιδευτικών προγραμμάτων που θα προετοιμάζουν το εργατικό δυναμικό για τις μελλοντικές απαιτήσεις της αγοράς εργασίας [22].

Η διεθνής συνεργασία και η ανταλλαγή γνώσεων αποτελούν επίσης κρίσιμους παράγοντες για την επιτυχή προετοιμασία για το μέλλον της TN. Η στενή συνεργασία μεταξύ ερευνητών, βιομηχανίας και κυβερνητικών φορέων είναι απαραίτητη για την πλήρη αξιοποίηση των δυνατοτήτων των νέων τεχνολογιών και την αποτελεσματική αντιμετώπιση των προκλήσεων που αναδύονται [22].

4.1.3 Υπευθυνότητα και ηθικός σχεδιασμός

Μετά τη διερεύνηση των τρόπων διασφάλισης της ηθικής χρήσης της TN και των προϋποθέσεων για την ασφαλή εφαρμογή της, είναι σημαντικό να εξεταστεί το ζήτημα της υπευθυνότητας και του ηθικού σχεδιασμού των συστημάτων TN. Η έννοια της υπευθυνότητας στην TN είναι πολυδιάστατη, περιλαμβάνοντας τόσο τεχνικές όσο και ηθικές παραμέτρους, από την ικανότητα λογοδοσίας και την προέλευση των αποφάσεων μέχρι την εκπαίδευση των συστημάτων για την αναγνώριση και διαχείριση ηθικά απαιτητικών καταστάσεων. Η παρούσα ενότητα εξετάζει το πώς μπορούμε να διασφαλίσουμε ότι τα συστήματα TN δεν είναι απλώς τεχνικά άρτια, αλλά και ηθικά υπεύθυνα στο σχεδιασμό και τη λειτουργία τους.

4.1.3.1 Ορισμός και πλαίσιο υπευθυνότητας TN

Η τεχνητή νοημοσύνη βρίσκεται σε μια διαρκή πορεία εξέλιξης, με τον ορισμό της να τροποποιείται συνεχώς, κυρίως λόγω των προσπαθειών για την υλοποίηση υπερ-νοημοσύνης [139]. Αυτές οι διαφοροποιήσεις στον ορισμό καθίστανται ιδιαίτερα σημαντικές όταν εξετάζονται ζητήματα

ευθύνης, ειδικά σε περιπτώσεις όπου η χρήση της ΤΝ έχει προκαλέσει βλάβη [139]. Παράλληλα, παρατηρείται μια αυξανόμενη απαίτηση για απόδοση ευθυνών, με στόχο τον περιορισμό των ελαττωμάτων που συνδέονται με τη χρήση της ΤΝ [23].

Η καθολική εφαρμογή της ΤΝ καθιστά τη λογοδοσία θεμελιώδη για τη διατήρηση της αλληλεπίδρασης ανθρώπου-υπολογιστή εντός αποδεκτών και νόμιμων ορίων. Χαρακτηριστικά παραδείγματα που υπογραμμίζουν τη σημασία της λογοδοσίας αποτελούν περιπτώσεις όπως η αναγνώριση του Whanganui River στη Νέα Ζηλανδία και της Sophia στη Σαουδική Αραβία ως ρομπότ ΤΝ με καθεστώς ιθαγένειας [23], [139].

Καθώς οι τεχνητές οντότητες αποκτούν ανθρώπινη ή οιονεί ανθρώπινη υπόσταση, εγείρονται σύνθετα νομικά και ηθικά ζητήματα, ιδιαίτερα στον δυτικό και αραβικό κόσμο [140]. Η αλγοριθμική υπευθυνότητα και η προέλευση των αποφάσεων αναδεικνύονται ως καίρια ζητήματα, με τα συστήματα ΤΝ να πρέπει να διασφαλίζουν ότι κάθε απόφαση είναι ανιχνεύσιμη και συμβάλλει στη διάδοση υπεύθυνων αλγοριθμικών συστημάτων [23], [36].

Ο ρυθμιστικός έλεγχος της αυτοματοποιημένης λήψης αποφάσεων παρουσιάζει σημαντικές διαφορές μεταξύ των εθνικών νομοθεσιών, με ορισμένες χώρες να επιλέγουν μια πιο διευρυμένη και άλλες μια πιο στενή προσέγγιση [141]. Η ανάγκη για ανθρώπινη συμμετοχή παραμένει ουσιώδης, αλλά αυτή αποκτά νόημα μόνο όταν συνδυάζεται με σαφή ενημέρωση των εμπλεκόμενων ατόμων [141]. Η ιδιωτικότητα και η ανωνυμία αποτελούν επίσης κρίσιμα ζητήματα λογοδοσίας, με τους εμπλεκόμενους να πρέπει να διασφαλίζουν την προστασία των δικαιωμάτων των τελικών χρηστών και άλλων ενδιαφερόμενων φορέων.

Καθώς η ΤΝ συνεχίζει να εξελίσσεται, παρουσιάζει νέους κινδύνους που οι ισχύοντες νόμοι ενδέχεται να μην μπορούν να αντιμετωπίσουν αποτελεσματικά [142]. Αυτό καθιστά αναγκαία την επανεξέταση και προσαρμογή των νομικών πλαισίων, λαμβάνοντας υπόψη τόσο τις τεχνολογικές εξελίξεις όσο και τις κοινωνικές επιπτώσεις της χρήσης της ΤΝ.

4.1.3.2 Δεοντολογικές παράμετροι και πληροφοριακή ηθική

Η ηθική της πληροφορίας επικεντρώνεται στην πολύπλευρη σχέση μεταξύ της δημιουργίας, οργάνωσης, διάδοσης και χρήσης της πληροφορίας μέσω συστημάτων τεχνητής νοημοσύνης. Ιδιαίτερη έμφαση δίνεται στη συγκέντρωση βασικών πληροφοριών για άτομα και στην ανταλλαγή δεδομένων μεταξύ εταιρειών, με στόχο τη δημιουργία ακριβέστερων προφίλ [135]. Τα προφίλ αυτά χρησιμοποιούνται συχνά ως εργαλεία καθοδήγησης της συμπεριφοράς των χρηστών προς επιθυμητές κατευθύνσεις, κυρίως για εμπορικούς σκοπούς.

Αναφορικά με την ευθύνη, οι κατασκευαστές και παραγωγοί καλούνται να αναλάβουν αυστηρή υπευθυνότητα για ζημιές που αποδίδονται στα σχέδια των ρομπότ. Η έννοια της ηλεκτρονικής προσωπικότητας εισάγεται ως μηχανισμός που συμβάλλει στην αυτοασφάλιση του ρομπότ και την εγγραφή του κατά τρόπο ανάλογο με την καταχώριση εταιρειών [143]. Προτείνεται επίσης ένα σύστημα καταχώρισης των προηγμένων ρομπότ βάσει συγκεκριμένων κριτηρίων ταξινόμησης, όπου μόνο τα δίκαια, χωρίς διακρίσεις και ειλικρινή μοντέλα πρέπει να ενσωματώνονται στις καθημερινές λειτουργίες.

Η διατήρηση του σώματος κανόνων είναι κρίσιμη, καθώς παρωχημένοι κανόνες μπορούν να αυξήσουν τα ψευδώς θετικά αποτελέσματα [18], [138]. Η ηθική αξιολόγηση των αλγορίθμων εστιάζει συχνά στη δράση, με ορισμένες ενέργειες να κρίνονται προκατειλημμένες αποκλειστικά

από την επίδρασή τους σε προστατευόμενες ομάδες [43]. Η εξάπλωση της χρήσης αλγορίθμων θέτει ηθικές προκλήσεις που δεν μπορούν πάντα να αναχθούν σε σαφείς περιπτώσεις επιστημικών ή ηθικών αποτυχιών, δεδομένου ότι οι αλγόριθμοι μπορούν να επηρεάσουν την αντίληψη του κόσμου και να τροποποιήσουν την κοινωνική και πολιτική οργάνωση [43].

Η ενσυναίσθηση και ο σεβασμός στη σχεδίαση προϊόντων και υπηρεσιών είναι απαραίτητες για τη μείωση των γνωστικών προκαταλήψεων και τη δημιουργία ικανοποιητικών εμπειριών για όλους τους χρήστες, σύμφωνα με τις απαιτήσεις του GDPR. Η μηχανική μάθηση αποδεικνύεται αποτελεσματικό εργαλείο για την ευαισθητοποίηση της κοινωνίας σχετικά με την προκατάληψη και τις παραβιάσεις δεδομένων, προετοιμάζοντας τα εμπλεκόμενα μέρη για την προσαρμογή σε νέους ή τροποποιημένους αλγόριθμους.

Παρά τις προσπάθειες μετριασμού των αθέμιτων προκαταλήψεων και διατύπωσης ηθικών αρχών για τον σχεδιασμό συστημάτων ΤΝ, η έλλειψη υποστηρικτικών ηθικών κανόνων ή κατευθυντήριων γραμμών πολιτικής παραμένει ένα σημαντικό κενό στη ρύθμιση των προγραμματιστών [144]. Αυτό υπογραμμίζει την ανάγκη για συνεχή ανάπτυξη και επικαιροποίηση των δεοντολογικών πλαισίων που διέπουν την ανάπτυξη και χρήση συστημάτων ΤΝ.

4.1.3.3 Αυτοματοποιημένη λήψη αποφάσεων και εποπτεία

Η αυτοματοποιημένη λήψη αποφάσεων αποτελεί ένα πεδίο που χρήζει ιδιαίτερης προσοχής σε ό,τι αφορά την υπευθυνότητα των συστημάτων ΤΝ. Σύμφωνα με την A29WP (Article 29 Data Protection Working Party - συμβουλευτικό όργανο στην Ευρωπαϊκή Ένωση με επίκεντρο θέματα προστασίας δεδομένων και απορρήτου), όταν μια μηχανή υπερέχει έναντι του ανθρώπου και αντιμετωπίζεται ως τέτοια, η οποιαδήποτε ανθρώπινη εμπλοκή στη διαδικασία δεν μπορεί να είναι απλά συμβολική [34]. Για τη διασφάλιση των θεμελιωδών δικαιωμάτων των χρηστών, απαιτείται "ουσιαστική ανθρώπινη συμβολή" και όχι απλώς μια "συμβολική χειρονομία".

Στις Ηνωμένες Πολιτείες, τα δικαιώματα που απειλούνται περισσότερο από τις αποφάσεις ΤΝ αφορούν τη μη διάκριση λόγω φυλής, φύλου, σεξουαλικού προσανατολισμού και θρησκείας [140]. Σε περίπτωση που η λήψη αποφάσεων από ανθρώπους αντικατασταθεί πλήρως από την ΤΝ, η ευθύνη θα συγκεντρώνεται σε μια πολύ μικρή ομάδα παραγωγών τεχνολογίας, δημιουργώντας έτσι σημαντικά ηθικά και νομικά ζητήματα [43], [145].

Το ρυθμιστικό πλαίσιο για τους ρομποτικούς συμβούλους εισάγει την έννοια της αυτορρύθμισης, όπου ο ιδιωτικός τομέας αναλαμβάνει ρυθμιστικές δραστηριότητες που παλαιότερα ήταν αποκλειστικά ευθύνη της κυβέρνησης [40]. Η πρώτη πρόκληση αφορά την αναγνώριση της ΤΝ ως υπεύθυνης οντότητας με νομικό καθεστώς και ευθύνη, ενώ η δεύτερη σχετίζεται με τη νομική ευθύνη και την προστασία των χρηστών που χρησιμοποιούν τέτοιες υπηρεσίες [40].

Σύμφωνα με τον GDPR, μια αυτοματοποιημένη απόφαση πρέπει να βασίζεται αποκλειστικά στην αυτοματοποιημένη επεξεργασία [146], [147]. Ωστόσο, η ευρύτητα του όρου αυτοματοποιημένη λήψη αποφάσεων καλύπτει διάφορους τύπους αποφάσεων, γεγονός που περιπλέκει την κατανόηση και την εφαρμογή του [146]. Τα άρθρα τονίζουν την ανάγκη για διαφάνεια και λογοδοσία στη διαχείριση αυτοματοποιημένων συστημάτων, ιδίως όταν αυτά χρησιμοποιούνται για τη λήψη αποφάσεων με νομικές επιπτώσεις [43], [148].

Η γλώσσα του GDPR υποδηλώνει ότι τα υποκείμενα των δεδομένων πρέπει να ενημερώνονται για τη λειτουργία των συστημάτων αυτοματοποιημένης λήψης αποφάσεων και τις προβλεπόμενες επιπτώσεις πριν ληφθούν αυτοματοποιημένες αποφάσεις [70]. Αν και το δικαίωμα επεξήγησης δεν είναι νομικά υποχρεωτικό, συνιστάται να παρέχεται ενημέρωση σχετικά με την ύπαρξη αυτοματοποιημένης λήψης αποφάσεων και τη λειτουργικότητα του συστήματος [70]. Το δικαίωμα του ατόμου να μην υπόκειται σε αυτοματοποιημένη λήψη αποφάσεων, όπως περιγράφεται στο άρθρο 22 του GDPR, παραμένει ασαφές και στερείται σαφώς καθορισμένων δικαιωμάτων και διασφαλίσεων [70].

4.1.3.4 Αυτονομία και όρια ευθυνότητας

Η εξέλιξη των αυτόνομων συστημάτων δημιουργεί σημαντικές προκλήσεις σε θέματα ευθύνης και ηθικής, ιδιαίτερα στην περίπτωση αυτοκινούμενων οχημάτων και άλλων συστημάτων ΤΝ [26]. Οι παραγωγοί αυτών των συστημάτων αντιμετωπίζουν δυσκολίες στην απόδοση ευθύνης για τις ζημιές που προκαλούνται, καθώς τα συστήματα αναμένεται να εξελίσσονται αυτόνομα [26]. Το ζήτημα της ευθύνης αναδεικνύεται ως ένα από τα πιο πιεστικά θέματα, ιδίως με την εισαγωγή των αυτοκινούμενων οχημάτων και την συνεχή εξέλιξη της τεχνολογίας [26].

Η ποικιλομορφία του εργατικού δυναμικού στον τομέα της ΤΝ διαδραματίζει κρίσιμο ρόλο, καθώς η έλλειψή της μπορεί να οδηγήσει σε αλγοριθμικές προκαταλήψεις που επηρεάζουν αρνητικά την ποιότητα των αποτελεσμάτων της ΤΝ [27]. Οι προκαταλήψεις αυτές μπορεί είτε να ενσωματωθούν στα συστήματα είτε, αντίθετα, η ΤΝ να συμβάλει στην αποκάλυψη και εξάλειψή τους [27]. Η διασφάλιση της δίκαιης και αμερόληπτης λήψης αποφάσεων μέσω της ΤΝ απαιτεί τη συνεργασία κυβερνήσεων, βιομηχανίας και ακαδημαϊκού κόσμου [27].

Η απόδειξη της διαφάνειας στις αποφάσεις που βασίζονται σε ΤΝ παρουσιάζει ιδιαίτερες δυσκολίες, λόγω της πολυπλοκότητας και της αυτονομίας της ΤΝ, που συχνά συνδέεται με έννοιες όπως η συνείδηση και η αυτογνωσία [149]. Αυτό περιπλέκει τη συζήτηση γύρω από την ευθύνη και τη διαχείριση της ΤΝ, ιδίως όταν δεν γίνεται επαρκής διάκριση μεταξύ των διαδοχικών φάσεων της ανάπτυξής της [149].

Η ηθική και η διακυβέρνηση αναδεικνύονται ως ζωτικής σημασίας παράμετροι για την ανάπτυξη και εφαρμογή της ΤΝ, επιτρέποντας την εστίαση σε βασικά ζητήματα που αντιμετωπίζουν οι οργανισμοί κατά την εφαρμογή των ερευνητικών αποτελεσμάτων [149]. Ωστόσο, η υπάρχουσα νομοθεσία ενδέχεται να εμποδίζει την καινοτομία και την εφαρμογή νέων τεχνολογιών, καθώς οι νόμοι συχνά δεν είναι προσαρμοσμένοι στις μοναδικές προκλήσεις που παρουσιάζει η ΤΝ [149]. Αν και οι εφαρμογές της ΤΝ υπόκεινται γενικά στις διάφορες μορφές του εμπορικού δικαίου, είναι αναγκαίο να επαναπροσδιοριστεί το πεδίο εφαρμογής της ΤΝ ώστε να ανταποκρίνεται στις σύγχρονες ηθικές και νομικές απαιτήσεις [149].

4.1.3.5 Κοινωνικές και ηθικές επιπτώσεις

Η χρήση της τεχνητής νοημοσύνης αναμένεται να προσφέρει σημαντική αξία στην οικονομία της Ευρωπαϊκής Ένωσης, ωστόσο το ισχύον νομικό πλαίσιο για την προστασία και την ιδιοκτησία των δεδομένων ενδέχεται να μην επαρκεί για την αντιμετώπιση της αναμενόμενης μαζικής ροής δεδομένων [150]. Παράλληλα, παρατηρείται ασάφεια στην ευρωπαϊκή νομοθεσία σχετικά με τη

δυνατότητα πνευματικής ιδιοκτησίας για τα έργα που παράγονται από την ΤΝ, καθώς τα συστήματα ΤΝ δημιουργούν όλο και περισσότερα έργα χωρίς ανθρώπινη παρέμβαση [40], [150].

Η κύρια ανησυχία του κοινού επικεντρώνεται στη δυνατότητα των συστημάτων ΤΝ να επιδείξουν προκαταλήψεις και να επηρεάσουν την εγκυρότητα των αποφάσεων. Η διαφανής λήψη αποφάσεων αναδεικνύεται ως θεμελιώδης προϋπόθεση για τα συστήματα ΤΝ, ώστε να εξασφαλιστούν δίκαιες και αξιόπιστες λύσεις, αν και η διαφάνεια από μόνη της δεν επαρκεί για την αντιμετώπιση όλων των ζητημάτων που προκύπτουν [32].

Η συμμετοχή του κοινού και η διαφάνεια στην ανάπτυξη των συστημάτων ΤΝ είναι κρίσιμες παράμετροι. Τα άτομα που χρησιμοποιούν συστήματα και αλγόριθμους ΤΝ πρέπει να έχουν τη δυνατότητα να ελέγχουν τα ιδρύματα που αναπτύσσουν την ΤΝ, ενώ τόσο οι προγραμματιστές όσο και οι χρήστες οφείλουν να συμμετέχουν ενεργά στις διαδικασίες ανάπτυξης και αυτοματοποίησης της ΤΝ [32], [151]. Ιδιαίτερη έμφαση δίνεται στη διατήρηση της ιδιωτικότητας και της ανωνυμίας των χρηστών, καθώς και στη διασφάλιση της εγκυρότητας των χρησιμοποιούμενων δεδομένων [36].

Η ανάπτυξη και η μαζική χρήση των συστημάτων ΤΝ έχουν εγείρει νέα ερωτήματα σχετικά με το καθεστώς ευθύνης των παρόχων υπηρεσιών διαδικτύου, γεγονός που μπορεί να οδηγήσει σε μια νέα εποχή ευθύνης γι' αυτούς [152]. Η επίτευξη πλήρους λογοδοσίας των προγραμματιστών της ΤΝ αποτελεί ένα περίπλοκο ζήτημα, ιδίως όταν εμπλέκονται μεγάλες εταιρείες όπως η Google και το Facebook, οι οποίες διαδραματίζουν κεντρικό ρόλο στη δημιουργία και χρηματοδότηση σημαντικών συστημάτων ΤΝ [147], [153].

Οι υπεύθυνοι χάραξης πολιτικής και οι διοικητικές αρχές καλούνται να διαδραματίσουν ενεργό ρόλο στην κατανόηση του τρόπου με τον οποίο οι οργανισμοί χρησιμοποιούν τα δεδομένα για την υπεράσπιση της χρήσης αλγορίθμων ΤΝ [154]. Οι προκλήσεις αυτές υπογραμμίζουν την ανάγκη για συνεχή επανεξέταση και προσαρμογή των ρυθμιστικών πλαισίων ώστε να εξασφαλιστεί ότι η χρήση της ΤΝ θα ωφεληθεί το ευρύ κοινό με δίκαιο και υπεύθυνο τρόπο.

4.1.4 Ζητήματα ιδιωτικότητας

Μετά τη διερεύνηση των τρόπων διασφάλισης της ηθικής χρήσης της ΤΝ, των προϋποθέσεων για την ασφαλή εφαρμογή της και των ζητημάτων υπευθυνότητας και ηθικού σχεδιασμού, είναι σημαντικό να εξεταστεί ένα από τα πιο κρίσιμα θέματα στη σύγχρονη εποχή - η προστασία της ιδιωτικότητας στο πλαίσιο της Τεχνητής Νοημοσύνης. Το ζήτημα αυτό αποκτά ιδιαίτερη σημασία στο σύγχρονο ψηφιακό περιβάλλον, όπου η συλλογή και επεξεργασία προσωπικών δεδομένων αποτελεί αναπόσπαστο μέρος των συστημάτων ΤΝ. Η παρούσα ενότητα εξετάζει τις προκλήσεις και τις λύσεις που σχετίζονται με την προστασία της ιδιωτικότητας στο πλαίσιο της ΤΝ, επικεντρώνοντας στη σχέση μεταξύ τεχνολογικής καινοτομίας και προστασίας προσωπικών δεδομένων.

4.1.4.1 Προστασία δεδομένων και ιδιωτικότητας στην εποχή της ΤΝ

Η προστασία των προσωπικών δεδομένων και της ιδιωτικότητας αναδεικνύεται ως μία από τις σημαντικότερες προκλήσεις στην εποχή της Τεχνητής Νοημοσύνης. Κεντρικό ρόλο σε αυτό το ζήτημα διαδραματίζουν οι μεγάλες τεχνολογικές εταιρείες, όπως η Microsoft, η Google, η IBM, η Baidu, η Tencent, η Intel, η Sony, η Workday, η SAP και η Sage, οι οποίες όχι μόνο κατασκευάζουν

μερικές από τις πιο σημαντικές συσκευές διαδικτύου, λογισμικού και υλικού στον κόσμο, αλλά διαχειρίζονται επίσης τεράστιες ποσότητες δεδομένων και ηγούνται των εξελίξεων της ΤΝ σε διάφορους τομείς [30].

Η συγκέντρωση δεδομένων σε λίγους μεγάλους παρόχους δημιουργεί σημαντικές ανησυχίες για την προστασία της ιδιωτικότητας. Μια περαιτέρω συγκέντρωση ισχύος είναι πιθανό να συμβεί σε λίγες μεγάλες επιχειρήσεις λόγω της κυριαρχίας τους στα δεδομένα, των αποτελεσμάτων πλατφόρμας και της πολλαπλασιαστικής δύναμης των δικτύων [35]. Σύμφωνα με το σχέδιο του Ηνωμένου Βασιλείου για την ΤΝ, οι μικρότερες επιχειρήσεις μπορεί να έχουν λιγότερη επιρροή στην πολιτική και τη συζήτηση για την ΤΝ, λόγω της έλλειψης πρόσβασης σε κρίσιμες πηγές και πληροφορίες [35].

Η διαφάνεια αναδεικνύεται ως θεμελιώδης προϋπόθεση για τα συστήματα ΤΝ, με στόχο την εξασφάλιση δίκαιων και αξιόπιστων λύσεων [155]. Ωστόσο, η διαφάνεια από μόνη της δεν επαρκεί για την αντιμετώπιση όλων των ζητημάτων που προκύπτουν από τη διαχείριση προσωπικών δεδομένων. Οι μεγάλες τεχνολογικές εταιρείες έχουν συχνά συναντήσεις για να συζητήσουν τη δεοντολογία των προγραμματιστών, τις τακτικές εσωτερικής διακυβέρνησης όπως οι συμβουλευτικές επιτροπές, καθώς και τη λειτουργία της εξωτερικής διακυβέρνησης [156].

Ενώ αυτές οι εταιρείες μπορεί να υποστηρίζουν την αυτοδιαχείριση ή τη συλλογική ρύθμιση του κλάδου, προειδοποιούν επίσης κατά της πολύ γρήγορης κυβερνητικής ρύθμισης [157]. Όπως και ο δημόσιος τομέας, αυτές οι πολυεθνικές εταιρείες μπορούν να δημιουργήσουν ανισορροπίες ισχύος που επηρεάζουν αρνητικά τις μικρές και μεσαίες επιχειρήσεις, τους πελάτες, ακόμη και τις κυβερνήσεις [143]. Τέλος, η επέκταση των αντιμονοπωλιακών μέτρων ή η αλλαγή του καθεστώτος διακυβέρνησης της ασφάλειας δεδομένων και της ιδιωτικής ζωής θα μπορούσαν να επηρεαστούν ή και να περιοριστούν από τις πιέσεις αυτών των τεχνολογικών γιγάντων [158].

4.1.4.2 Ρυθμιστικό πλαίσιο και μηχανισμοί προστασίας

Η ανάγκη για ένα ολοκληρωμένο ρυθμιστικό πλαίσιο που θα διασφαλίζει την προστασία της ιδιωτικότητας στο περιβάλλον της ΤΝ καθίσταται ολοένα και πιο επιτακτική. Η ανησυχία ότι η τηλεϊατρική και τα ευφυή συστήματα διδασκαλίας μπορεί να αυτοματοποιήσουν θέσεις εργασίας στους τομείς της υγείας και της εκπαίδευσης είναι ιδιαίτερα έντονη, καθώς αυτό θα μπορούσε να οδηγήσει σε μείωση των επενδύσεων στα συστήματα υγείας και εκπαίδευσης σε αναπτυσσόμενες χώρες, εμποδίζοντας μακροπρόθεσμες αναπτυξιακές πορείες υπέρ γρήγορων λύσεων ή βραχυπρόθεσμων κερδών [159].

Ιδιαίτερη πρόκληση αποτελεί η προστασία των δεδομένων στις αναπτυσσόμενες χώρες. Υπάρχει ο κίνδυνος οι πιο ευάλωτες περιοχές και πληθυσμοί να περιθωριοποιηθούν αν ο παγκόσμιος διάλογος για την ΤΝ και την πολιτική της διαμορφώνεται κυρίως από μια μικρή ομάδα ισχυρών χωρών ή περιφερειακών οργανισμών, όπως η Ευρωπαϊκή Ένωση [33]. Παρά τις προσπάθειες για την αντιμετώπιση των ηθικών και κοινωνικών ζητημάτων που αναδεικνύουν οι κορυφαίοι παράγοντες της ΤΝ, υπάρχει ο κίνδυνος να επιδεινωθούν οι παγκόσμιες ανισότητες και ανισορροπίες ισχύος, με τις αναπτυσσόμενες χώρες να βρίσκονται σε μειονεκτική θέση [34].

Η πολιτική ΤΝ του Μεξικού και η εθνική στρατηγική της Ινδίας για την ΤΝ, με το σύνθημα #AIFORALL, εκφράζουν ανησυχίες σχετικά με την πιθανή μετατόπιση θέσεων εργασίας στον τομέα της εξυπηρέτησης πελατών και της τεχνικής υποστήριξης λόγω της αυτοματοποίησης και της

ανάθεσης εργασιών σε εξωτερικούς συνεργάτες σε χώρες με μεγαλύτερο διαθέσιμο εισόδημα [160]. Αυτή η εξέλιξη μπορεί να δημιουργήσει νέες προκλήσεις για τις αναπτυσσόμενες χώρες, περιορίζοντας τα οφέλη της ΤΝ σε επίπεδο τοπικής ανάπτυξης και απασχόλησης.

Οι διεθνείς οργανισμοί διαδραματίζουν καίριο ρόλο στη διαμόρφωση των μηχανισμών προστασίας. Η διεθνής συνεργασία, ιδιαίτερα μέσω οργανισμών όπως ο ΟΟΣΑ, η G7 και η G20, καθώς και πολλές οντότητες της ΕΕ, έχει αποδειχθεί ζωτικής σημασίας για τη διαμόρφωση της πολιτικής και της ηθικής της ΤΝ σε διεθνές επίπεδο [31]. Η αποτελεσματικότητα των ρυθμιστικών πλαισίων εξαρτάται σε μεγάλο βαθμό από τη συνεργασία μεταξύ των διαφόρων ενδιαφερόμενων μερών και την ικανότητά τους να προσαρμόζονται στις ταχέως μεταβαλλόμενες τεχνολογικές εξελίξεις.

4.1.4.3 Συγκρότηση πολιτικών και τεχνικών λύσεων

Η συγκρότηση αποτελεσματικών πολιτικών και τεχνικών λύσεων για την προστασία της ιδιωτικότητας στην εποχή της ΤΝ απαιτεί μια πολυδιάστατη προσέγγιση. Η σοβαρότητα ενός ηθικού εγγράφου ή μιας πρωτοβουλίας μπορεί να εκτιμηθεί από το επίπεδο της επιβολής και της επιτήρησης που περιγράφεται [161]. Όταν μια εταιρεία είναι πρόθυμη να υποβληθεί σε δαπανηρές επιθεωρήσεις, ιδίως όταν αυτές περιλαμβάνουν ανεξάρτητη εξωτερική επιτήρηση, αυτό δείχνει μια σοβαρή δέσμευση πέρα από την απλή σηματοδότηση καλών προθέσεων [162].

Οι στρατηγικές των εταιρειών για την προστασία δεδομένων πρέπει να περιλαμβάνουν συγκεκριμένα μέτρα και μηχανισμούς ασφαλείας. Τα σχέδια για επανάληψη και συνεχή παρακολούθηση είναι κρίσιμα, διότι δείχνουν ότι μια επιχείρηση δεν αρκείται σε μια εφάπαξ δήλωση, αλλά επιδιώκει την εφαρμογή και την εξέλιξη των δεσμεύσεων της με βάση τις εξελισσόμενες ανάγκες και προκλήσεις [163]. Ωστόσο, εξακολουθούν να υπάρχουν ανησυχίες σχετικά με τις πιο αποτελεσματικές μεθόδους ηγεσίας και το επίπεδο εμπιστοσύνης που μπορεί να επιτευχθεί μέσω των πρωτοβουλιών που προτείνονται [160].

Η εξισορρόπηση μεταξύ καινοτομίας και προστασίας της ιδιωτικότητας αποτελεί μια συνεχή πρόκληση. Δεδομένης της συνεχώς μεταβαλλόμενης φύσης της τεχνητής νοημοσύνης και των συνακόλουθων κοινωνικών, ηθικών και πολιτικών ζητημάτων, οι εταιρείες και οι οργανισμοί μπορεί να χρειαστεί να υιοθετήσουν μέτρα όπως η διεξαγωγή πιλοτικών δοκιμών στρατηγικών και η διαβούλευση με επαγγελματίες και δημόσιους φορείς για συνεχή καθοδήγηση και προσαρμογή των πρακτικών τους [164].

Η συνεργασία μεταξύ των διαφόρων φορέων είναι ζωτικής σημασίας για την επιτυχία των πολιτικών προστασίας της ιδιωτικότητας. Είναι χαρακτηριστικό ότι από τα 100 δεοντολογικά άρθρα που αναλύθηκαν, τα 20 προέρχονταν από τον ιδιωτικό τομέα, γεγονός που δείχνει ότι τόσο ο δημόσιος όσο και ο εμπορικός τομέας παίζουν καθοριστικό ρόλο στην ηγεσία της ΤΝ [165]. Η αποτελεσματική προστασία της ιδιωτικότητας απαιτεί τη συνεργασία μεταξύ κυβερνήσεων, επιχειρήσεων, ακαδημαϊκών ιδρυμάτων και της κοινωνίας των πολιτών για την ανάπτυξη και εφαρμογή ολοκληρωμένων λύσεων.

4.1.4.4 Προκλήσεις και μελλοντικές κατευθύνσεις

Οι προκλήσεις στην προστασία της ιδιωτικότητας στο πλαίσιο της ΤΝ είναι πολυεπίπεδες και απαιτούν συστηματική αντιμετώπιση. Μια σημαντική πρόκληση είναι οι αυξανόμενες

ανισορροπίες ισχύος μεταξύ των παρόχων υπηρεσιών. Η τάση της μαζικής μετανάστευσης εξειδικευμένων επαγγελματιών από τις αναπτυσσόμενες χώρες είναι πιθανό να συνεχιστεί, καθώς υπάρχει έντονος ανταγωνισμός για την πρόσληψη εμπειρογνομόνων υψηλής εξειδίκευσης στην ΤΝ [166]. Επιπλέον, τα αναπτυσσόμενα έθνη μπορεί να μην επωφεληθούν πλήρως από την τεχνολογία αυτοματοποίησης, η οποία επιβάλλεται από τα εύπορα κράτη ως μέσο εξοικονόμησης κεφαλαίου και εργατικού δυναμικού [167].

Η προστασία των ευάλωτων ομάδων αποτελεί μια ακόμη σημαντική πρόκληση. Σύμφωνα με το σχέδιο του Ηνωμένου Βασιλείου για την ΤΝ, οι μικρότερες επιχειρήσεις και οι αναπτυσσόμενες χώρες μπορεί να έχουν λιγότερη επιρροή στην πολιτική και τη συζήτηση για την ΤΝ [35]. Αυτό μπορεί να οδηγήσει σε περαιτέρω περιθωριοποίηση και να εντείνει τις υπάρχουσες ανισότητες στην πρόσβαση και τον έλεγχο των προσωπικών δεδομένων.

Η διεθνής συνεργασία αναδεικνύεται ως βασικός παράγοντας για την αντιμετώπιση των προκλήσεων. Οι φορείς του δημόσιου τομέα συχνά ενσωματώνουν ουσιαστικά ηθικά επιχειρήματα σε ευρύτερα σχέδια πολιτικής και όχι μόνο σε κείμενα που αφορούν τη δεοντολογία [155]. Αν και η πλειονότητα των συγγραφέων των εγγράφων αυτών προέρχεται από τον Παγκόσμιο Βορρά, υπάρχει σημαντική γεωγραφική διαφοροποίηση [168]. Η συμμετοχή διεθνών οργανισμών και η συνεργασία μεταξύ διαφορετικών χωρών είναι απαραίτητη για την ανάπτυξη αποτελεσματικών λύσεων.

Όσον αφορά τις μελλοντικές κατευθύνσεις, η δημόσια πολιτική αναδεικνύεται ως κύριος τομέας εστίασης. Οι οργανισμοί μπορεί να αρχίσουν με γενικές παραδοχές, αλλά στη συνέχεια να προσθέσουν λεπτομέρειες καθώς το πλαίσιο προστασίας της ιδιωτικότητας εξελίσσεται [169]. Η εμπέδωση και η έκθεση των πολιτικών, δηλαδή πόσο συχνά αναφέρονται ή διατίθενται δημόσια, μπορεί να επηρεάσει την αντίληψη της αξίας και της ποιότητάς τους [170], ενώ η ανάπτυξη ισχυρών μηχανισμών προστασίας της ιδιωτικότητας θα απαιτήσει συνεχή προσαρμογή στις εξελισσόμενες τεχνολογικές δυνατότητες και κοινωνικές ανάγκες [171].

4.1.5 Ανάγκη ηθικής καθοδήγησης

Καθώς η Τεχνητή Νοημοσύνη εξελίσσεται από εργαλείο σε συνεργάτη στη λήψη κρίσιμων αποφάσεων, το ερώτημα της ηθικής καθοδήγησης μετατοπίζεται από το πεδίο της θεωρητικής συζήτησης στο επίκεντρο της πρακτικής ανάπτυξης συστημάτων ΤΝ. Οι προγραμματιστές και μηχανικοί που χτίζουν αυτά τα συστήματα δεν είναι απλώς τεχνικοί δημιουργοί αλλά και αρχιτέκτονες μιας νέας μορφής λήψης αποφάσεων που επηρεάζει εκατομμύρια ανθρώπους. Ωστόσο, ενώ η τεχνική εκπαίδευση και καθοδήγηση είναι εκτενής και συστηματική, η αντίστοιχη ηθική προετοιμασία παραμένει συχνά αποσπασματική και διαισθητική. Η παρούσα ενότητα εξετάζει αυτό το κενό μεταξύ τεχνικής αριστείας και ηθικής επάρκειας, διερευνώντας τους τρόπους με τους οποίους μπορούμε να μεταβούμε από την απλή συζήτηση περί δεοντολογίας στην ουσιαστική ενσωμάτωση ηθικών αρχών στην καθημερινή πρακτική ανάπτυξης συστημάτων ΤΝ.

4.1.5.1 Εκπαίδευση και επαγγελματική κατάρτιση

Η ταχεία ενσωμάτωση της Τεχνητής Νοημοσύνης σε όλους τους τομείς της κοινωνίας καθιστά επιτακτική την ανάγκη για συστηματική ηθική καθοδήγηση, ξεκινώντας από την εκπαίδευση των μελλοντικών επαγγελματιών του κλάδου. Είναι κρίσιμο οι μηχανικοί μηχανικής μάθησης (Machine

Learning – ML) να αναπτύξουν βαθιά επίγνωση των κοινωνικών και ηθικών επιπτώσεων της integrating εργασίας τους [36]. Αυτό απαιτεί την ενσωμάτωση της ηθικής ως θεμελιώδους συστατικού της εκπαίδευσης ML, όχι απλώς ως πρόσθετου μαθήματος αλλά ως αναπόσπαστου μέρους του προγράμματος σπουδών [36].

Τα εκπαιδευτικά ιδρύματα και οι εταιρείες καλούνται να προσφέρουν μαθήματα που εξετάζουν τη διασύνδεση της TN και της ανθρωπότητας, καλύπτοντας ζωτικής σημασίας ζητήματα όπως η ισότητα, η ηθική και η ιδιωτικότητα [36], [137], [139]. Η εκπαίδευση πρέπει να υπερβαίνει την απλή εξοικείωση με τις τυπικές ηθικές προκλήσεις και να στοχεύει στην ενσωμάτωση ηθικών προβληματισμών μέσα στις βασικές έννοιες της μηχανικής μάθησης [36].

Απαιτείται μια νέα εκπαιδευτική στρατηγική που θα συνδυάζει τα ηθικά πρότυπα και τις προτάσεις για την TN σε δημόσιες πολιτικές, επηρεάζοντας ουσιαστικά τον τρόπο διδασκαλίας της εκπαίδευσης των μηχανικών [38]. Ιδιαίτερη έμφαση πρέπει να δοθεί στην προώθηση διεπιστημονικών αλληλεπιδράσεων μεταξύ μηχανικών, σχεδιαστών, ηθικολόγων και κοινωνικών επιστημόνων [38]. Αυτή η διεπιστημονική προσέγγιση είναι απαραίτητη για την κατανόηση του τρόπου με τον οποίο η αυτονομία των μηχανών πρέπει να εξισορροπείται με κοινωνικούς και οικονομικούς παράγοντες, λαμβάνοντας υπόψη ηθικές, θρησκευτικές και νομικές παραμέτρους [38].

Η εκπαίδευση πρέπει επίσης να αντιμετωπίσει το ζήτημα του τεχνολογικού αναλφαριθμητισμού που παρατηρείται μεταξύ των ρυθμιστικών αρχών, καθώς αυτό μπορεί να οδηγήσει σε αναποτελεσματικότητα και δυνητική δημόσια βλάβη [134]. Η έλλειψη ευαισθητοποίησης και κατανόησης σχετικά με την τεχνητή νοημοσύνη μεταξύ διαφόρων ενδιαφερομένων, συμπεριλαμβανομένων των καταναλωτών, των φορέων χάραξης πολιτικής, των δικηγόρων και των ακαδημαϊκών, υπογραμμίζει την ανάγκη για ολοκληρωμένη εκπαίδευση και κατάρτιση σε όλα τα επίπεδα [172].

4.1.5.2 Θεσμικά πλαίσια και πρότυπα

Η ανάγκη για ισχυρά θεσμικά πλαίσια και πρότυπα στον τομέα της Τεχνητής Νοημοσύνης γίνεται όλο και πιο επιτακτική. Οι επιτροπές δεοντολογίας διαδραματίζουν καίριο ρόλο σε αυτή την προσπάθεια, ωστόσο η τρέχουσα λειτουργία τους σε απομόνωση και η αδιαφάνεια των διαδικασιών λήψης αποφάσεων υπογραμμίζουν την ανάγκη για αυξημένη επικοινωνία και ανταλλαγή γνώσεων μεταξύ όλων των ενδιαφερομένων, συμπεριλαμβανομένων των δημιουργών TN, των υπευθύνων χάραξης πολιτικής και των ηθικολόγων [173].

Η αδιάλειπτη προσπάθεια του IEEE (Institute of Electrical and Electronics Engineers) για τον καθορισμό προτύπων στοχεύει στη δημιουργία πιστοποιήσεων που επικεντρώνονται στη λογοδοσία και τη διαφάνεια των αλγοριθμικών λειτουργιών. Ωστόσο, υπάρχουν ακόμη κενά στον καθορισμό ενός σαφούς πλαισίου συμμόρφωσης και βέλτιστων πρακτικών για την ανάπτυξη συστημάτων που βασίζονται σε δεδομένα [174]. Η δημιουργία ενός ενιαίου αρχείου αποφάσεων των επιτροπών δεοντολογίας θα μπορούσε να ενισχύσει σημαντικά την ανάπτυξη σαφών κανόνων και να υποστηρίξει την υπεύθυνη καινοτομία στην TN [173].

Τα δεοντολογικά πρότυπα είναι ζωτικής σημασίας για την οικοδόμηση αξιόπιστων συστημάτων, με τις επιχειρήσεις να καλούνται να ακολουθούν συγκεκριμένες αξίες και κατευθυντήριες αρχές για τη νόμιμη, δίκαιη και αμερόληπτη επεξεργασία των δεδομένων [37].

Προς αυτή την κατεύθυνση, προτείνονται ολοκληρωμένες εκτιμήσεις επιπτώσεων δεδομένων (Consumer Data Industry Association – CDIA) για την εξέταση των οφελών, των κινδύνων, των δικαιωμάτων και των συμφερόντων όλων των εμπλεκόμενων φορέων [37].

Η σύσταση ενός Συμβουλίου για την ΤΝ και τη δεοντολογία των δεδομένων θα μπορούσε να προσφέρει πολύτιμη καθοδήγηση στο ταχέως αναπτυσσόμενο πεδίο της ΤΝ [22]. Ένα τέτοιο συμβούλιο θα μπορούσε να λειτουργήσει ως κεντρικός φορέας για τη διαμόρφωση πολιτικών και την παροχή ηθικής καθοδήγησης, διασφαλίζοντας ότι η ανάπτυξη της ΤΝ παραμένει προσανατολισμένη στο κοινωνικό όφελος [22]. Τα ρυθμιστικά πλαίσια πρέπει να εξελίσσονται μέσω συνεχών πειραματισμών, αντιμετωπίζοντας ευαίσθητα ζητήματα όπως οι οικονομικές επιπτώσεις στις θέσεις εργασίας και η ανάγκη συνδυασμού σκληρών και ήπιων ρυθμίσεων για τη διαχείριση των κινδύνων [22].

4.1.5.3 Πρακτικές προκλήσεις εφαρμογής

Η εφαρμογή ηθικής καθοδήγησης στην Τεχνητή Νοημοσύνη αντιμετωπίζει σημαντικές πρακτικές προκλήσεις. Μια κρίσιμη πρόκληση είναι η εξισορρόπηση μεταξύ της προώθησης της καινοτομίας και της αντιμετώπισης πιθανών κινδύνων και ηθικών προβλημάτων [134]. Η διαμάχη μεταξύ των αρχών της επιφυλακτικότητας και της καινοτομίας στην πολιτική για την ΤΝ απεικονίζει τις αντίθετες προσεγγίσεις για την αντιμετώπιση των κινδύνων και των πλεονεκτημάτων της ΤΝ [175].

Ένα σημαντικό ζήτημα που έχει αναδυθεί είναι το φαινόμενο του "ethics-washing" και του "ethics-shopping", όπου η επίκληση της δεοντολογίας λειτουργεί περισσότερο ως μορφή αυτορρύθμισης παρά ως πραγματική δέσμευση για δεοντολογικές συμπεριφορές [39], [40]. Αυτή η κατάσταση απαιτεί τη θέσπιση ρητών κριτηρίων για την αξιολόγηση της ισχύος των δεσμεύσεων ηθικής και ανθρωπίνων δικαιωμάτων, ιδιαίτερα μπροστά στην αυξανόμενη αντίσταση στη ρύθμιση [39].

Η διαχείριση κινδύνων απαιτεί μια εξελιγμένη προσέγγιση που συνδυάζει την εκτίμηση των κινδύνων και τα συμφέροντα πολλών ενδιαφερομένων μερών [37]. Καθώς η αποδοχή της τεχνητής νοημοσύνης βρίσκεται ακόμη σε πρώιμο στάδιο, είναι ζωτικής σημασίας οι κανονισμοί να τονώνουν την ανάπτυξή της ενώ ταυτόχρονα αντιμετωπίζουν εύλογες ανησυχίες σχετικά με την ασφάλεια, την προστασία της ιδιωτικής ζωής και τη δεοντολογική χρήση [175].

Μια καινοτόμος προσέγγιση στην αντιμετώπιση αυτών των προκλήσεων είναι η έννοια των ρυθμιστικών sandboxes, που παρέχουν ένα ελεγχόμενο περιβάλλον για τη δοκιμή πρωτοποριακών εφαρμογών ΤΝ υπό ρυθμιστικό έλεγχο [39], [40]. Αυτή η προσέγγιση επιτρέπει την προληπτική αυτορρύθμιση και τη διακυβέρνηση βάσει αρχών, παρέχοντας έναν πρακτικό τρόπο για την εξισορρόπηση καινοτομίας και ηθικών περιορισμών [39].

Οι υπερβολικά περιοριστικοί κανόνες που βασίζονται στην αρχή της προφύλαξης μπορεί να περιορίσουν την ικανότητα της ΤΝ να βελτιώσει τις εμπειρίες των καταναλωτών και να βοηθήσει σε ζωτικές δραστηριότητες [175]. Συνεπώς, συνιστάται στους υπεύθυνους χάραξης πολιτικής να επικεντρωθούν σε νόμους που αφορούν συγκεκριμένους τομείς και όχι σε μια προσέγγιση που ταιριάζει σε όλους, ώστε να διασφαλιστεί ότι οι κανόνες προσαρμόζονται στις ξεχωριστές δυσκολίες και ευκαιρίες που θέτει η ΤΝ σε διάφορους τομείς [175].

4.1.5.4 Συνεργασία και συντονισμός φορέων

Η αποτελεσματική ηθική καθοδήγηση στην Τεχνητή Νοημοσύνη απαιτεί την ενεργό συμμετοχή και συνεργασία όλων των ενδιαφερόμενων φορέων. Η συνεργασία μεταξύ της κυβέρνησης και του ιδιωτικού τομέα είναι ζωτικής σημασίας για την οικοδόμηση ενός πλαισίου για τη νομοθεσία ΤΝ, αναγνωρίζοντας τα αλληλένδετα συμφέροντα των κοινωνικών, πολιτικών, εμπορικών και ερευνητικών φορέων [22].

Η δυνατότητα της ΤΝ να συμβάλει στο κοινωνικό καλό απαιτεί έναν ενδεδειγμένο κοινωνικοπολιτικό σχεδιασμό, λαμβάνοντας υπόψη τον μεταβαλλόμενο χαρακτήρα των κοινωνιών της πληροφορίας και την αυξανόμενη εξάρτησή τους από την τεχνολογία ΤΝ [22]. Αυτή η εξάρτηση έχει ουσιαστικές συνέπειες για τις κοινές αξίες, καθιστώντας απαραίτητη τη συμμετοχή όλων των κοινωνικών εταίρων στη διαμόρφωση των σχετικών πολιτικών [22].

Οι νομικές και τεχνολογικές πτυχές της τεχνητής νοημοσύνης μελετώνται τόσο από ακαδημαϊκούς όσο και από τη βιομηχανία, υπογραμμίζοντας την ανάγκη για συντονισμένη διακυβέρνηση, πρότυπα και ηθικά πλαίσια [176], [177]. Εξαιρετικά σημαντική είναι η διασφάλιση ότι τα συστήματα ΤΝ, ειδικά εκείνα με στοιχεία μηχανικής μάθησης όπως τα βαθιά νευρωνικά δίκτυα, εξετάζονται ολιστικά για την ποιότητά τους, συνδυάζοντας τη γνώση και την εμπειρία συγκεκριμένων τομέων [176], [177].

Τα μέτρα της Ευρωπαϊκής Επιτροπής, συμπεριλαμβανομένης της αυξημένης χρηματοδότησης της ΤΝ και της ίδρυσης της Ευρωπαϊκής Συμμαχίας για την ΤΝ, υπογραμμίζουν τη δέσμευση για την οικοδόμηση ενός ηθικού και νομικού πλαισίου για την ΤΝ μέσω διεθνούς συνεργασίας [134]. Αυτή η συνεργατική προσέγγιση είναι απαραίτητη για την αντιμετώπιση των παγκόσμιων προκλήσεων που θέτει η ΤΝ, καθώς καμία μεμονωμένη χώρα ή οργανισμός δεν μπορεί να τις αντιμετωπίσει αποτελεσματικά μόνος του.

Επιπλέον, η συμμετοχή της κοινωνίας των πολιτών είναι κρίσιμη για τη διασφάλιση ότι η ανάπτυξη και η εφαρμογή της ΤΝ παραμένουν προσανατολισμένες στο κοινωνικό όφελος [36], [137], [139]. Οι νομικοί ενθαρρύνονται να θεωρούν την ΤΝ ως εργαλείο που μπορεί να ενδυναμώσει την κοινωνία των πολιτών, να αυξήσει την ευαισθητοποίηση και να εκδημοκρατίσει την ΤΝ μέσω προπαθειών από κάτω προς τα πάνω [178], [179].

4.2 Ευρήματα Περιόδου 2024

Η περίοδος 2024 χαρακτηρίζεται από την έντονη επίδραση της Γενετικής ΤΝ στο πεδίο της ηθικής της τεχνητής νοημοσύνης. Η ευρεία διάδοση αυτών των τεχνολογιών έχει φέρει στο προσκήνιο νέες προκλήσεις και έχει οδηγήσει στην ανάπτυξη πιο εξειδικευμένων προσεγγίσεων για την ηθική διακυβέρνηση της ΤΝ, ενώ παράλληλα έχει αναδείξει την ανάγκη για επικαιροποίηση των υφιστάμενων ρυθμιστικών πλαισίων και την ανάπτυξη νέων μηχανισμών προστασίας.

4.2.1 Εξελίξεις στην ηθική χρήση της ΤΝ

Στο πλαίσιο των ραγδαίων εξελίξεων στον τομέα της ΤΝ κατά την περίοδο 2024, η διασφάλιση της ηθικής χρήσης των συστημάτων ΤΝ έχει αποκτήσει νέες διαστάσεις. Η αυξανόμενη πολυπλοκότητα και αυτονομία των προηγμένων συστημάτων επεξεργασίας γλώσσας, όπως το ChatGPT και το Claude, απαιτεί πιο εξειδικευμένες προσεγγίσεις στην αλγοριθμική δικαιοσύνη, την ανθρώπινη

εποπτεία, και την προστασία των δικαιωμάτων των χρηστών, ενώ παράλληλα αναδεικνύει την ανάγκη για πιο ολοκληρωμένα ρυθμιστικά πλαίσια και μηχανισμούς ελέγχου.

4.2.1.1 Αλγοριθμική δικαιοσύνη και λογοδοσία στα σύγχρονα συστήματα TN

Η εξέλιξη των μηχανισμών αλγοριθμικής δικαιοσύνης για την αντιμετώπιση των προκλήσεων που θέτουν τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) και η GenAI έχει σημειώσει σημαντική πρόοδο. Σύμφωνα με πρόσφατες έρευνες, έχουν αναπτυχθεί ολοκληρωμένες ταξινομίες για την αξιολόγηση και τον μετριασμό των προκαταλήψεων σε διάφορα επίπεδα των μοντέλων, συμπεριλαμβανομένων των ενσωματώσεων, των πιθανοτήτων και του παραγόμενου κειμένου [180]. Παράλληλα, έχουν εισαχθεί εργαλεία όπως το FairTransLing που ενισχύουν την ερμηνευσιμότητα και διασφαλίζουν τη δικαιοσύνη στις διαδικασίες λήψης αποφάσεων της TN [181].

Ιδιαίτερη έμφαση έχει δοθεί στην ανάπτυξη νέων μεθόδων λογοδοσίας και διαφάνειας για συστήματα TN που βασίζονται σε μεγάλα γλωσσικά μοντέλα. Η μέθοδος TokenSHAP, που χρησιμοποιεί εκτίμηση της τιμής Monte Carlo Shapley, προσφέρει ένα αυστηρό πλαίσιο για την κατανόηση των αποκρίσεων του μοντέλου, βελτιώνοντας τη διαφάνεια και την ερμηνευσιμότητα [182]. Επιπλέον, το πλαίσιο SCENE παράγει κατάλληλες επεξηγήσεις μέσω ήπιων αντιπαραδειγμάτων, παρέχοντας πληροφορίες για τα πλεονεκτήματα και τους περιορισμούς διαφόρων τεχνικών ΧΑΙ [183].

Η αντιμετώπιση της προκατάληψης στα σύγχρονα συστήματα Γενετικής TN αποτελεί κρίσιμο ζήτημα που απαιτεί πολύπλευρη προσέγγιση. Έχουν αναπτυχθεί πλαίσια που ενσωματώνουν στρατηγικές ανίχνευσης και διόρθωσης προκαταλήψεων, χρησιμοποιώντας προηγμένες μετρικές και αλγορίθμους για την παραγωγή πιο ισορροπημένων αποτελεσμάτων [184]. Η κατασκευή συνόλων δεδομένων με κοινωνική κλίση μπορεί να βοηθήσει στη διόρθωση ανισορροπιών και στην εκπαίδευση μοντέλων για την παραγωγή κοινωνικά συμπεριληπτικού περιεχομένου [185].

Σημαντική πρόοδος έχει σημειωθεί στην ανάπτυξη νέων προτύπων και μετρικών για την αξιολόγηση της δικαιοσύνης στα συστήματα TN. Το πλαίσιο FAIR-Frame αξιολογεί τη δικαιοσύνη σε πολλαπλά προστατευόμενα χαρακτηριστικά, εστιάζοντας τόσο στη αναπαραστατική όσο και στην κατανομητική βλάβη [186]. Παράλληλα, το πλαίσιο EARN διευκολύνει τις συλλογικές αποφάσεις μεταξύ των ενδιαφερόμενων μερών σχετικά με τις μετρικές δικαιοσύνης, χωρίς να απαιτείται εξειδίκευση στην TN [187]. Νέες μετρικές όπως το Normalized Fairness Gain (NFG) Score και το Harmonic Fairness (HF) Score έχουν προταθεί για τη μέτρηση της ατομικής δικαιοσύνης και τη διευκόλυνση συγκρίσεων μεταξύ μοντέλων μετριασμού της αδικίας [188], [189].

Η σχέση μεταξύ ανθρώπινης εποπτείας και αυτόνομης λήψης αποφάσεων στα σύγχρονα συστήματα TN έχει εξελιχθεί σημαντικά, με έμφαση στην ισορροπία μεταξύ των δυνατοτήτων της TN και της ανθρώπινης εποπτείας. Η Ευρωπαϊκή Πράξη για την TN και άλλα ρυθμιστικά πλαίσια τονίζουν τη σημασία της ανθρώπινης επίβλεψης για την πρόληψη της υπερβολικής εξάρτησης από τα συστήματα TN [190]. Τα υβριδικά συστήματα και η συνεργασία ανθρώπου-TN έχουν αναδειχθεί ως βέλτιστη πρακτική, επιτρέποντας στην TN να χειρίζεται συνήθεις εργασίες ενώ οι άνθρωποι παρεμβαίνουν σε πολύπλοκα ή κρίσιμα σενάρια [191], [192].

4.2.1.2 Τεχνικές προσεγγίσεις για ηθική TN

Η ανάπτυξη τεχνικών προσεγγίσεων για τη διασφάλιση της ηθικής συμπεριφοράς στα σύγχρονα συστήματα TN έχει σημειώσει σημαντική πρόοδο. Το πλαίσιο DIKE ενισχύει την ικανότητα των LLMs να εσωτερικεύουν παγκόσμιες ανθρώπινες αξίες και να προσαρμόζονται σε πολιτισμικά πλαίσια, χρησιμοποιώντας ανταγωνιστικές αναθεωρήσεις και self-supervised μάθηση [193]. Παράλληλα, το Lang-BiTe προσφέρει μια λύση καθοδηγούμενη από μοντέλα που καθορίζει ηθικές απαιτήσεις και αυτοματοποιεί τον έλεγχο προκαταλήψεων στα LLMs, παρέχοντας πλήρη ιχνηλασιμότητα μεταξύ των ηθικών απαιτήσεων και της αξιολόγησής τους [194].

Οι τεχνικές διασφαλίσεις στα σύγχρονα συστήματα TN έχουν εξελιχθεί σημαντικά, με την εισαγωγή καινοτόμων μηχανισμών ασφαλείας. Οι "διακόπτες κυκλώματος" (circuit breakers), εμπνευσμένοι από τη μηχανική αναπαράστασης, διακόπτουν τα μοντέλα TN όταν επιχειρούν να παράγουν επιβλαβή αποτελέσματα, παρέχοντας ισχυρή άμυνα έναντι κακόβουλων επιθέσεων χωρίς να διακυβεύεται η χρησιμότητα [195]. Επιπλέον, έχουν αναπτυχθεί αρχιτεκτονικές ασφαλείας που περιλαμβάνουν φίλτρα εισόδου-εξόδου, ενσωματωμένους παράγοντες ασφαλείας και ιεραρχικά συστήματα ανάθεσης με ενσωματωμένους ελέγχους ασφαλείας [196].

Οι προκλήσεις και οι περιορισμοί των τρεχουσών τεχνικών λύσεων για την ηθική TN παραμένουν σημαντικοί, ιδιαίτερα σε πολύπλοκα συστήματα όπως τα LLMs και η Γενετική TN. Τα ζητήματα προκατάληψης και δικαιოსύνης εξακολουθούν να αποτελούν σημαντική πρόκληση, καθώς τα LLMs συχνά αντανakλούν προκαταλήψεις που υπάρχουν στα δεδομένα εκπαίδευσής τους [197]. Επιπλέον, η παραγωγή ανακριβών ή παραπλανητικών πληροφοριών (hallucinations) θέτει σημαντικούς κινδύνους σε εφαρμογές που απαιτούν υψηλή αξιοπιστία, όπως οι χρηματοοικονομικές συμβουλές ή η υγειονομική περίθαλψη [198].

Οι τρέχουσες τεχνικές μέθοδοι για τον εντοπισμό και την πρόληψη ανήθικης συμπεριφοράς στα συστήματα TN έχουν αποδειχθεί ιδιαίτερα αποτελεσματικές στον τομέα της ανίχνευσης απάτης. Τεχνικές μηχανικής μάθησης και βαθιάς μάθησης, συμπεριλαμβανομένων των νευρωνικών δικτύων, των δέντρων αποφάσεων και της ομαδοποίησης, χρησιμοποιούνται εκτενώς για τον εντοπισμό δόλιων δραστηριοτήτων μέσω της ανάλυσης ιστορικών δεδομένων και της αναγνώρισης ανωμαλιών [199]. Επιπλέον, η επεξεργασία φυσικής γλώσσας (NLP) χρησιμοποιείται για την ανάλυση κειμενικών δεδομένων, όπως emails και περιγραφές συναλλαγών, ενισχύοντας την ανίχνευση ύποπτων μοτίβων [200].

Ο ρόλος των τεχνικών αρχιτεκτονικών και του σχεδιασμού μοντέλων στην προώθηση της ηθικής συμπεριφοράς της TN είναι θεμελιώδης. Μια προσέγγιση κύκλου ζωής στην ηθική της TN περιλαμβάνει την ενσωμάτωση ηθικών πρακτικών όπως η αξιοπιστία, η ασφάλεια και η εμπιστοσύνη από το σχεδιασμό έως την αξιολόγηση [201]. Ιδιαίτερη έμφαση δίνεται στον σχεδιασμό με επίκεντρο τον χρήστη, διασφαλίζοντας ότι τα συστήματα TN αναπτύσσονται με γνώμονα τις ανάγκες των χρηστών και τις ηθικές παραμέτρους, προωθώντας τη διαφάνεια, τη δικαιοσύνη και την προσβασιμότητα [202].

4.2.1.3 Ρυθμιστικό και κανονιστικό πλαίσιο

Το ρυθμιστικό πλαίσιο για τα προηγμένα συστήματα TN εξελίσσεται σημαντικά, με χαρακτηριστικό παράδειγμα τον Ευρωπαϊκό Κανονισμό για την TN (AI Act), ο οποίος αποτελεί

πρωτοποριακή προσπάθεια ρύθμισης των frontier μοντέλων [203]. Η προσέγγιση που υιοθετείται συνδυάζει τόσο αρχές (principle-based) όσο και κανόνες (rule-based), με τάση μετάβασης προς πιο συγκεκριμένους κανόνες καθώς βελτιώνεται η κατανόηση των κινδύνων [204]. Παράλληλα, προτείνεται η δημιουργία εθνικών μητρώων για τα μοντέλα TN [205] και η εφαρμογή ενός συστήματος έγκρισης παρόμοιου με αυτό του FDA, που θα εστιάζει στην πειραματική επικύρωση της ασφάλειας πριν την ανάπτυξη των συστημάτων [206], [207].

Η αποτελεσματική διακυβέρνηση της TN απαιτεί τη συνεργασία πολλαπλών φορέων και την ενσωμάτωση καινοτόμων μηχανισμών εποπτείας. Ιδιαίτερη έμφαση δίνεται στην ενσωμάτωση του ανθρώπινου παράγοντα στον έλεγχο (Human-in-the-loop) σε συνδυασμό με την τεχνολογία blockchain για τη διασφάλιση της διαφάνειας και της λογοδοσίας [208]. Επιπλέον, προτείνεται ένα πολυεπίπεδο πλαίσιο διακυβέρνησης που χαρτογραφεί τους εμπλεκόμενους φορείς σε έξι επίπεδα λήψης αποφάσεων, από το διεθνές έως το οργανωτικό, επιτρέποντας την εξέλιξη υφιστάμενων εργαλείων ή τη δημιουργία νέων για την καθοδήγηση της συμπεριφοράς των εμπλεκόμενων μερών [209]. Η προσέγγιση αυτή συμπληρώνεται από την προσαρμογή παραδοσιακών ρυθμιστικών μοντέλων στα μοναδικά χαρακτηριστικά της TN, συμπεριλαμβανομένων των καθεστώτων αδειοδότησης και της τυπικής επαλήθευσης της συμπεριφοράς των συστημάτων TN [210].

4.2.1.4 Προκλήσεις και κίνδυνοι

Τα προηγμένα συστήματα TN και τα LLMs παρουσιάζουν σημαντικές προκλήσεις και κινδύνους σε διάφορα επίπεδα. Κεντρικά ζητήματα αποτελούν η προστασία της ιδιωτικότητας και η ασφάλεια των δεδομένων, καθώς τα LLMs μπορούν να απομνημονεύσουν και πιθανώς να αποκαλύψουν ευαίσθητες πληροφορίες [211]. Παράλληλα, η αδιαφάνεια στη λήψη αποφάσεων των συστημάτων TN υπονομεύει την εμπιστοσύνη και τη λογοδοσία [212], ενώ τα hallucinations μπορεί να έχουν σοβαρές επιπτώσεις, ιδιαίτερα σε τομείς όπως η υγεία και η εκπαίδευση [213]. Επιπρόσθετα, τα ζητήματα μεροληψίας και δικαιοσύνης παραμένουν κρίσιμα, καθώς τα μοντέλα μπορούν να διαιωνίσουν κοινωνικές προκαταλήψεις λόγω των δεδομένων εκπαίδευσής τους [214].

Η αντιμετώπιση αυτών των προκλήσεων απαιτεί μια ολοκληρωμένη προσέγγιση που περιλαμβάνει τεχνολογικές, ηθικές και ρυθμιστικές στρατηγικές. Η ανάπτυξη προηγμένων αλγορίθμων αναγνώρισης κινδύνων σε συνδυασμό με μηχανισμούς ηθικής εποπτείας είναι απαραίτητη για την προστασία από τεχνολογικές αστοχίες και ηθικές παραβάσεις [215]. Η συνεχής παρακολούθηση και οι δυναμικές στρατηγικές μετριασμού των κινδύνων, προσαρμοσμένες στις μοναδικές ευπάθειες των μοντέλων GenAI, αποτελούν επίσης κρίσιμα στοιχεία [216]. Επιπλέον, η εμπλοκή των ενδιαφερόμενων μερών στην ανάπτυξη, παρακολούθηση και βελτίωση των προτύπων και κανονισμών TN είναι ζωτικής σημασίας [217], ενώ η κυβερνητική παρέμβαση σε ομοσπονδιακό επίπεδο κρίνεται απαραίτητη για την αντιμετώπιση κακόβουλων χρήσεων όπως τα deepfakes και οι προκαταλήψεις στα μοντέλα TN [218].

4.2.1.5 Ανθρωποκεντρική προσέγγιση

Η διασφάλιση της ανθρωποκεντρικής προσέγγισης στην ανάπτυξη και χρήση των προηγμένων συστημάτων TN απαιτεί την ενσωμάτωση ισχυρών ηθικών θεμελίων και μηχανισμών λογοδοσίας. Η έρευνα αναδεικνύει τη σημασία της ενσωμάτωσης αρχών όπως η δικαιοσύνη, η διαφάνεια και η μη διάκριση [219], ενώ παράλληλα τονίζεται η ανάγκη για επεξηγήσιμα συστήματα

TN που επιτρέπουν στους χρήστες να κατανοούν το σκεπτικό πίσω από τις αποφάσεις τους [220]. Ιδιαίτερη έμφαση δίνεται στη συνεργασία ανθρώπου και TN, με στόχο την ενίσχυση των διαδικασιών λήψης αποφάσεων και τη συμπληρωματική λειτουργία των συστημάτων TN με τις ανθρώπινες ικανότητες [221].

Αναδύονται νέα πλαίσια και πρακτικές που προωθούν την ανθρωποκεντρική ανάπτυξη της TN, με έμφαση στον ανθρωποκεντρικό σχεδιασμό και την αυτοματοποίηση με επίκεντρο τον άνθρωπο (Human-Centered Automation). Αυτές οι προσεγγίσεις προωθούν διαισθητικές διεπαφές και λύσεις ανοικτού κώδικα για τον εκδημοκρατισμό της πρόσβασης στις τεχνολογίες TN [222]. Επιπλέον, η δημόσια συμμετοχή και εκπαίδευση διαδραματίζουν καθοριστικό ρόλο στην ευθυγράμμιση των συστημάτων TN με τις ανθρώπινες αξίες, όπως αποδεικνύεται από πρωτοβουλίες όπως το AI in Public Health Toolkit, που ενδυναμώνει τα άτομα να αξιολογούν κριτικά τα συστήματα TN και να συμμετέχουν ενεργά στη διαμόρφωσή τους [223], [224].

4.2.2 Νέες προκλήσεις ασφάλειας

Η εμφάνιση προηγμένων transformers έχει δημιουργήσει ένα νέο τοπίο προκλήσεων ασφαλείας που απαιτεί εξειδικευμένες λύσεις και μηχανισμούς προστασίας. Η αυξημένη ικανότητα αυτών των συστημάτων να επεξεργάζονται και να παράγουν περιεχόμενο, σε συνδυασμό με την ευρεία χρήση τους σε κρίσιμους τομείς, έχει αναδείξει νέα ζητήματα που αφορούν την ασφάλεια των δεδομένων, την προστασία από κακόβουλη χρήση και την αξιοπιστία των αποτελεσμάτων.

4.2.2.1 Θεμελιώδεις αρχές υπεύθυνης TN

Οι θεμελιώδεις αρχές που πρέπει να διέπουν την ανάπτυξη και χρήση των προηγμένων συστημάτων TN εστιάζουν στη διαφάνεια, τη λογοδοσία, τη δικαιοσύνη, την ιδιωτικότητα και την ασφάλεια. Η διαφάνεια των αλγορίθμων και των διαδικασιών λήψης αποφάσεων είναι κρίσιμη για την οικοδόμηση εμπιστοσύνης και τον έλεγχο των συστημάτων [225], ενώ παράλληλα απαιτούνται σαφείς μηχανισμοί λογοδοσίας για την απόδοση ευθυνών σε περιπτώσεις βλάβης ή αδικίας [226]. Ιδιαίτερη έμφαση δίνεται στην αντιμετώπιση των προκλήσεων που θέτουν τα ισχυρά γλωσσικά μοντέλα, όπως η διαχείριση των κοινωνικών προκαταλήψεων και η πιθανή κακή χρήση τους, μέσω της ενσωμάτωσης ηθικής εποπτείας και μηχανισμών λογοδοσίας του κλάδου [227].

Τα αναδυόμενα πλαίσια για την εφαρμογή των αρχών υπεύθυνης TN στην πράξη επικεντρώνονται στην ενσωμάτωση ηθικών παραμέτρων σε όλο τον κύκλο ζωής της ανάπτυξης TN. Ιδιαίτερη έμφαση δίνεται στην ενσωμάτωση των αρχών υπεύθυνης TN στις διάφορες φάσεις του Κύκλου Ζωής Ανάπτυξης Λογισμικού (SDLC), ειδικά κατά τη φάση Εκπαίδευσης Απαιτήσεων [228]. Παράλληλα, προτείνονται καινοτόμες προσεγγίσεις όπως η αξιοποίηση του κβαντικού υπολογισμού για την αντιμετώπιση προκλήσεων όπως η μεροληψία και η συμπερίληψη [229], ενώ αναπτύσσονται εξειδικευμένα πλαίσια για συγκεκριμένους τομείς, όπως η ανώτατη εκπαίδευση, που δίνουν έμφαση στις αρχές της διαφάνειας, της δικαιοσύνης και της λογοδοσίας [230].

4.2.2.2 Μηχανισμοί ασφάλειας και προστασίας

Η ασφάλεια των προηγμένων συστημάτων TN και η προστασία των δεδομένων τους απαιτούν την ανάπτυξη πολυεπίπεδων μηχανισμών ασφαλείας. Κεντρικό ρόλο διαδραματίζει η ενσωμάτωση τεχνικών τεχνητής νοημοσύνης στην κυβερνοασφάλεια, όπως η μηχανική μάθηση και

η ανίχνευση ανωμαλιών, που επιτρέπουν την ανάλυση μεγάλων συνόλων δεδομένων σε πραγματικό χρόνο για τον εντοπισμό πιθανών απειλών [231]. Παράλληλα, υιοθετείται μια προσέγγιση άμυνας σε βάθος (Defense-in-Depth) που ενσωματώνει λειτουργικές προσεγγίσεις, προσεγγίσεις κύκλου ζωής και προσεγγίσεις βασισμένες στις απειλές για τη διαχείριση των κινδύνων που σχετίζονται με την TN, αξιοποιώντας πλαίσια όπως το NIST Cybersecurity Framework [232].

Για την ενίσχυση της ανθεκτικότητας και της αξιοπιστίας των προηγμένων μοντέλων TN, εφαρμόζονται τεχνικές όπως η ανταγωνιστική εκπαίδευση (adversarial training) και η επαύξηση δεδομένων για την προστασία από κακόβουλες επιθέσεις [233]. Ιδιαίτερη έμφαση δίνεται στην προστασία των LLMs μέσω της ενσωμάτωσης προηγμένων τεχνολογιών όπως το blockchain για τη δημιουργία αυτόνομων λύσεων ασφαλείας [234], ενώ παράλληλα εφαρμόζονται ηθικές δικλίδες ασφαλείας όπως το φιλτράρισμα ευαίσθητου λεξιλογίου και η ανίχνευση ανάληψης ρόλων για την πρόληψη παραβιάσεων της ιδιωτικότητας [235].

4.2.2.3 Διαφάνεια και επεξηγησιμότητα συστημάτων

Οι νέες μεθοδολογίες για την ενίσχυση της διαφάνειας και επεξηγησιμότητας των συστημάτων TN εστιάζουν στο επίπεδο της αρχιτεκτονικής των μοντέλων. Το παράδειγμα Integrated Invariance Interpretation (I3) περιλαμβάνει αλγορίθμους όπως η Ανάλυση Επίδρασης Χαρακτηριστικών (FIA) και η Ανάλυση Κύριων Συνιστωσών (PIA) για τη βελτίωση της ερμηνευσιμότητας των νευρωνικών δικτύων [236]. Παράλληλα, η προσέγγιση της μηχανιστικής ερμηνευσιμότητας επιχειρεί την αντίστροφη μηχανική των νευρωνικών δικτύων για την κατανόηση των υπολογιστικών μηχανισμών τους, κάτι που είναι κρίσιμο για την ασφάλεια και την ευθυγράμμιση των συστημάτων TN [237].

Η εξέλιξη των τεχνικών οπτικοποίησης παίζει επίσης σημαντικό ρόλο στην κατανόηση των διαδικασιών λήψης αποφάσεων της TN. Τεχνικές μείωσης διαστάσεων όπως η PCA και t-SNE χρησιμοποιούνται για τη μετατροπή πολύπλοκων δεδομένων υψηλής διάστασης σε πιο ερμηνεύσιμες μορφές [238]. Επιπλέον, αναπτύσσονται προηγμένες πλατφόρμες οπτικοποίησης που υποστηρίζονται από TN, οι οποίες παρέχουν λεπτομερείς τρισδιάστατες απεικονίσεις για την υποστήριξη της λήψης αποφάσεων ακριβείας, βελτιώνοντας την επικοινωνία μεταξύ των ενδιαφερόμενων μερών και την κατανόηση πολύπλοκων δεδομένων [239].

4.2.2.4 Προετοιμασία για το μέλλον της TN

Η προετοιμασία της κοινωνίας και του εργατικού δυναμικού για την ευρεία υιοθέτηση των συστημάτων TN απαιτεί ολοκληρωμένες στρατηγικές ανάπτυξης δεξιοτήτων και οργανωσιακού μετασχηματισμού. Ιδιαίτερη έμφαση δίνεται στην προσαρμοστική ανάπτυξη δεξιοτήτων και στην υλοποίηση συνεχών προγραμμάτων κατάρτισης για την προετοιμασία των εργαζομένων στο νέο περιβάλλον [240]. Οι οργανισμοί καλούνται να αναπτύξουν τεχνολογικούς οδικούς χάρτες 3-5 ετών που περιγράφουν τα χρονοδιαγράμματα υιοθέτησης της TN και τον επιχειρησιακό μετασχηματισμό [241], ενώ παράλληλα προωθείται η συνεργασία μεταξύ ακαδημαϊκών ιδρυμάτων, βιομηχανίας και κυβέρνησης για την υπεύθυνη ανάπτυξη των δυνατοτήτων της TN [242].

Τα εκπαιδευτικά συστήματα εξελίσσονται για να αντιμετωπίσουν τις απαιτήσεις του μέλλοντος της TN, ενσωματώνοντας εξατομικευμένη και προσαρμοστική μάθηση που ανταποκρίνεται στις ατομικές ανάγκες των μαθητών [243]. Δίνεται έμφαση στην ανάπτυξη

δεξιοτήτων του 21ου αιώνα, όπως η κριτική σκέψη, η δημιουργικότητα και η ψηφιακή ευχέρεια [244], ενώ παράλληλα αναπτύσσονται ολοκληρωμένα προγράμματα κατάρτισης για τους εκπαιδευτικούς ώστε να μπορούν να ενσωματώσουν αποτελεσματικά την ΤΝ στη μαθησιακή διαδικασία [245], [246].

4.2.3 Σύγχρονες προσεγγίσεις υπευθυνότητας

Οι σύγχρονες προσεγγίσεις υπευθυνότητας στην τεχνητή νοημοσύνη έχουν εξελιχθεί σημαντικά για να ανταποκριθούν στις προκλήσεις που θέτουν τα σύγχρονα συστήματα διαλόγου και παραγωγής περιεχομένου. Η αυξημένη αυτονομία και πολυπλοκότητα αυτών των συστημάτων απαιτεί νέα πλαίσια καταμερισμού ευθυνών, επαναπροσδιορισμό των ορίων ανθρώπινης εποπτείας και ανάπτυξη καινοτόμων μηχανισμών λογοδοσίας που εξασφαλίζουν την υπεύθυνη χρήση και ανάπτυξη της τεχνολογίας.

4.2.3.1 Ορισμός και πλαίσιο υπευθυνότητας ΤΝ

Η έννοια της υπευθυνότητας στην ΤΝ έχει επαναπροσδιοριστεί σημαντικά μετά την ευρεία υιοθέτηση των LLMs και της παραγωγικής ΤΝ. Η ενσωμάτωση των LLMs σε διάφορους τομείς, όπως η τηλεϊατρική και η εκπαίδευση, έχει αναδείξει την ανάγκη για υπεύθυνη χρήση, με έμφαση στην ηθική της ΤΝ για τη διασφάλιση αξιόπιστων εφαρμογών [247]. Οι βασικοί τομείς που απαιτούν ιδιαίτερη προσοχή περιλαμβάνουν την ιδιωτικότητα, την ασφάλεια, τη διαφάνεια και την ανθρώπινη αυτονομία, στοιχεία που συχνά δεν διερευνώνται επαρκώς αλλά είναι κρίσιμα για την υπεύθυνη ανάπτυξη της ΤΝ [248].

Η κατανομή της ευθύνης μεταξύ των διαφόρων εμπλεκόμενων μερών αποτελεί κεντρικό ζήτημα στο σύγχρονο πλαίσιο υπευθυνότητας της ΤΝ. Οι προγραμματιστές είναι υπεύθυνοι για την ενσωμάτωση ηθικών παραμέτρων στα συστήματα ΤΝ, εστιάζοντας στη διαφάνεια, τη λογοδοσία και τη δικαιοσύνη [249], ενώ οι εταιρείες καλούνται να εφαρμόσουν δομές διακυβέρνησης που δίνουν προτεραιότητα στις ηθικές ανησυχίες και την εμπιστοσύνη στην ΤΝ [250]. Παράλληλα, οι χρήστες αποτελούν αναπόσπαστο μέρος του πλαισίου διακυβέρνησης, τονίζοντας τη σημασία των ανθρωποκεντρικών αξιών και της δικαιοσύνης στην ανάπτυξη και χρήση των συστημάτων ΤΝ [251].

4.2.3.2 Δεοντολογικές παράμετροι και πληροφοριακή ηθική

Η εξέλιξη των θεμελιωδών ηθικών αρχών για την ΤΝ έχει καθοδηγηθεί από την ανάγκη αντιμετώπισης των προκλήσεων που θέτουν τα συστήματα παραγωγικής ΤΝ και τα αυτόνομα συστήματα. Οι βασικές προκλήσεις περιλαμβάνουν την αντιμετώπιση της μεροληψίας και της δικαιοσύνης, με έμφαση στη χρήση αλγορίθμων προεπεξεργασίας δεδομένων για τη διασφάλιση της ισότιμης μεταχείρισης διαφορετικών πληθυσμιακών ομάδων [252]. Παράλληλα, η άνοδος των τεχνολογιών παραγωγικής ΤΝ έχει εντείνει τις ανησυχίες σχετικά με την προστασία της ιδιωτικότητας και των προσωπικών δεδομένων, καθώς τα συστήματα αυτά μπορούν να δημιουργήσουν συνθετικά μέσα που ενδέχεται να παραβιάζουν τα δικαιώματα προσωπικών δεδομένων [253], [254].

Η διαχείριση της πληροφορίας στο πλαίσιο των σύγχρονων συστημάτων ΤΝ απαιτεί μια πολύπλευρη προσέγγιση που εστιάζει στην προστασία της ιδιωτικότητας, τη συναίνεση και το

διαμοιρασμό πληροφοριών. Οι τεχνικές διαφορικής ιδιωτικότητας και κρυπτογράφησης αναδεικνύονται ως απαραίτητες για τη διατήρηση της εμπιστευτικότητας, ιδιαίτερα σε ευαίσθητους τομείς όπως η υγειονομική περίθαλψη [255]. Η εξασφάλιση της συνειδητής συναίνεσης αποτελεί κρίσιμο ηθικό ζήτημα, ειδικά στον τομέα της υγείας και των χρηματοοικονομικών υπηρεσιών, όπου τα προσωπικά δεδομένα χρησιμοποιούνται εκτενώς, απαιτώντας διαφανείς πρακτικές διαχείρισης δεδομένων και ρητή συγκατάθεση για την οικοδόμηση εμπιστοσύνης [256], [257].

4.2.3.3 Αυτοματοποιημένη λήψη αποφάσεων και εποπτεία

Η εξέλιξη της αυτοματοποιημένης λήψης αποφάσεων έχει επηρεαστεί σημαντικά από την πρόοδο των LLMs και της παραγωγικής ΤΝ. Η ενσωμάτωση των LLMs με εξελικτικούς αλγόριθμους έχει βελτιώσει τη διαδικασία λήψης αποφάσεων σε πολύπλοκες εργασίες βελτιστοποίησης, επιτρέποντας πιο λεπτομερείς εξηγήσεις και ευθυγράμμιση με την εμπειρογνομοσύνη διαφόρων ενδιαφερόμενων μερών [258]. Παράλληλα, οι νευρο-συμβολικές προσεγγίσεις, όπως η πλατφόρμα Cora, συνδυάζουν τα LLMs με συμβολικό συλλογισμό για την αντιμετώπιση των απαιτήσεων ακρίβειας και λογικής συνέπειας σε τομείς υψηλού ρίσκου όπως η ιατρική, τα νομικά και τα χρηματοοικονομικά [259].

Η διαφάνεια και η επεξηγησιμότητα των συστημάτων λήψης αποφάσεων ΤΝ παραμένουν κρίσιμες προκλήσεις, ιδιαίτερα στους ρυθμιζόμενους κλάδους. Πολλά μοντέλα ΤΝ λειτουργούν ως "μαύρα κουτιά", καθιστώντας δύσκολη την κατανόηση των διαδικασιών λήψης αποφάσεων, γεγονός που μπορεί να οδηγήσει σε ζητήματα εμπιστοσύνης και ηθικές ανησυχίες [260], [261]. Για την αντιμετώπιση αυτών των προκλήσεων, προτείνεται η εφαρμογή τεχνικών ΧΑΙ και υβριδικών συστημάτων που βασίζονται σε κανόνες, τα οποία μπορούν να βελτιώσουν τη διαφάνεια χωρίς να θυσιάζουν την απόδοση [262]. Επιπλέον, η ανάπτυξη ολοκληρωμένων πλαισίων διακυβέρνησης της ΤΝ που δίνουν έμφαση στη διαφάνεια, τη δικαιοσύνη και τη λογοδοσία είναι κρίσιμη, συμπεριλαμβανομένης της επικύρωσης, πιστοποίησης και συνεχούς παρακολούθησης των συστημάτων ΤΝ [263].

4.2.3.4 Αυτονομία και όρια υπευθυνότητας

Ο καθορισμός της ευθύνης για ζημιές που προκαλούνται από αυτόνομα συστήματα ΤΝ, ιδιαίτερα εκείνα που χρησιμοποιούν μεγάλα γλωσσικά μοντέλα, αποτελεί ένα πολύπλοκο ζήτημα που απαιτεί ολοκληρωμένα νομικά και ηθικά πλαίσια. Η Ευρωπαϊκή Ένωση έχει αναπτύξει σημαντικές πρωτοβουλίες, όπως ο Νόμος για την ΤΝ και η Οδηγία περί Ευθύνης ΤΝ (AILD), που επικεντρώνονται στην εκ των προτέρων αξιολόγηση κινδύνων και την εναρμόνιση του δικαίου περί αδιοπραξιών για την αντιμετώπιση ζημιών που σχετίζονται με την ΤΝ [264]. Τα πλαίσια αυτά προτείνουν τη διευκόλυνση του βάρους απόδειξης για τα θύματα και την εξέταση της αντικειμενικής ευθύνης λόγω της απρόβλεπτης φύσης των συστημάτων ΤΝ [265].

Η εξισορρόπηση της αυτονομίας της ΤΝ με την ανθρώπινη εποπτεία απαιτεί πολύπλευρη προσέγγιση από τους οργανισμούς. Η συνεργασία μεταξύ ΤΝ και ανθρώπινης εποπτείας είναι θεμελιώδης, καθώς η ΤΝ ενισχύει τις ανθρώπινες δυνατότητες αλλά απαιτεί ανθρώπινη παρέμβαση για τη διασφάλιση ηθικών και ακριβών αποτελεσμάτων, ιδιαίτερα σε τομείς υψηλού ρίσκου όπως η υγειονομική περίθαλψη [266]. Η καθιέρωση σαφών πλαισίων διακυβέρνησης της ΤΝ είναι ζωτικής σημασίας, συμπεριλαμβανομένων διαφανών διαδικασιών, τακτικών ελέγχων και συνεχούς

κατάρτισης των εργαζομένων για τον μετριασμό των κινδύνων και τη διασφάλιση της συμμόρφωσης με τα ηθικά πρότυπα [267], [268].

4.2.3.5 Κοινωνικές και ηθικές επιπτώσεις

Η ευρεία υιοθέτηση της παραγωγικής ΤΝ έχει σημαντικές κοινωνικές και ηθικές επιπτώσεις, ιδιαίτερα στους τομείς της απασχόλησης, της εκπαίδευσης και των κοινωνικών αλληλεπιδράσεων. Στον τομέα της απασχόλησης, η ΤΝ μπορεί να οδηγήσει σε μετατόπιση θέσεων εργασίας λόγω αυτοματοποίησης, αλλά παράλληλα δημιουργεί νέους ρόλους στη διαχείριση και ανάπτυξη της ΤΝ [253], [269]. Στην εκπαίδευση, η ΤΝ έχει τη δυνατότητα να φέρει επανάσταση στις εκπαιδευτικές μεθόδους προσφέροντας εξατομικευμένες μαθησιακές εμπειρίες, αλλά απαιτεί ηθικές κατευθυντήριες γραμμές για τη διασφάλιση της υπεύθυνης χρήσης και της ακαδημαϊκής ακεραιότητας [270], [271].

Η διασφάλιση της ισότιμης πρόσβασης στα οφέλη της ΤΝ, παράλληλα με τον μετριασμό των κινδύνων του ψηφιακού χάσματος και των αλγοριθμικών διακρίσεων, απαιτεί μια πολύπλευρη προσέγγιση. Τα μεγάλα γλωσσικά μοντέλα συχνά εμφανίζουν προκαταλήψεις προς την αγγλική γλώσσα και το αμερικανοκεντρικό πλαίσιο, καθιστώντας απαραίτητη τη βελτιστοποίησή τους με διαφοροποιημένα σύνολα δεδομένων για τη βελτίωση της απόδοσής τους σε διαφορετικές γλώσσες και πολιτισμούς [272], [273]. Παράλληλα, η εφαρμογή πλαισίων δικαιοσύνης και τεχνικών μετριασμού προκαταλήψεων, όπως η αντιπαραθετική μάθηση, μπορεί να βοηθήσει στη μείωση των προκαταλήψεων στα συστήματα ΤΝ, ιδιαίτερα σε εκπαιδευτικά πλαίσια [274], [275].

4.2.4 Προστασία προσωπικών δεδομένων

Η προστασία των προσωπικών δεδομένων έχει αποκτήσει νέες διαστάσεις με την εμφάνιση των σύγχρονων συστημάτων τεχνητής νοημοσύνης. Τα ζητήματα ιδιωτικότητας δεν περιορίζονται πλέον μόνο στην ασφαλή αποθήκευση και διαχείριση των δεδομένων, αλλά επεκτείνονται στην προστασία από ακούσια αποκάλυψη πληροφοριών κατά την αλληλεπίδραση με τα συστήματα, τη διασφάλιση της ανωνυμίας στα δεδομένα εκπαίδευσης, και την πρόληψη της μη εξουσιοδοτημένης χρήσης προσωπικών πληροφοριών στο παραγόμενο περιεχόμενο.

4.2.4.1 Προστασία δεδομένων και ιδιωτικότητας στην εποχή των γλωσσικών μοντέλων και του Generative AI

Η προστασία δεδομένων και ιδιωτικότητας στην εποχή των LLMs και των συστημάτων generative AI παρουσιάζει νέες σημαντικές προκλήσεις. Σύμφωνα με τους [276], τα συστήματα αυτά είναι ευάλωτα τόσο σε παθητική διαρροή ευαίσθητων πληροφοριών όσο και σε ενεργές επιθέσεις ιδιωτικότητας, όπως η δηλητηρίαση δεδομένων (data poisoning) και η διαρροή Προσωπικά Αναγνωρίσιμων Πληροφοριών (PII). Επιπλέον, οι [278] επισημαίνουν ότι τα μοντέλα αυτά είναι ευάλωτα σε επιθέσεις παραβίασης (jailbreaking) και επιθέσεις backdoor, ενώ οι [264] τονίζουν τα κενά στη νομοθεσία που πρέπει να αντιμετωπιστούν για τη διαχείριση των μοναδικών προκλήσεων που θέτουν αυτές οι τεχνολογίες.

Για την αντιμετώπιση αυτών των προκλήσεων, έχουν αναπτυχθεί διάφορες τεχνικές και προσεγγίσεις προστασίας. Οι [279], [280] προτείνουν τη χρήση διαφορικής ιδιωτικότητας (differential privacy) σε συνδυασμό με άλλες μεθόδους για την ενίσχυση της ακρίβειας των

συστάσεων και της εμπειρίας χρήστη, διατηρώντας παράλληλα την προστασία της ιδιωτικότητας. Οι [281] παρουσιάζουν τεχνικές κωδικοποίησης αντικατάστασης (Replacement Encoding) για την απόκρυψη ευαίσθητων πληροφοριών κατά την εκπαίδευση μοντέλων AI, ενώ οι [282] προτείνουν μεθόδους ανωνυμοποίησης με χρήση τεχνητής νοημοσύνης για την προστασία των προσωπικών δεδομένων διατηρώντας παράλληλα τη χρησιμότητά τους για έρευνα και ανάλυση.

4.2.4.2 Ρυθμιστικό πλαίσιο και μηχανισμοί προστασίας στην εποχή του Generative AI

Τα υπάρχοντα ρυθμιστικά πλαίσια και οι πολιτικές προστασίας δεδομένων αντιμετωπίζουν σημαντικές προκλήσεις στην εποχή των μεγάλων γλωσσικών μοντέλων και των συστημάτων Generative AI. Οι [283] και [284] επισημαίνουν ότι οι κίνδυνοι είναι ιδιαίτερα έντονοι σε ευαίσθητους τομείς όπως η υγειονομική περίθαλψη, όπου οι παραβιάσεις μπορεί να έχουν σοβαρές συνέπειες. Οι [264] και [285] τονίζουν ότι τα υφιστάμενα νομικά πλαίσια, συμπεριλαμβανομένου του προτεινόμενου Κανονισμού για την TN της ΕΕ, δεν είναι πλήρως εξοπλισμένα για την αντιμετώπιση των μοναδικών προκλήσεων που θέτουν αυτά τα συστήματα.

Για την αντιμετώπιση αυτών των προκλήσεων, το 2024 σημειώθηκαν σημαντικές εξελίξεις στους μηχανισμούς προστασίας. Οι [286] και [287] υπογραμμίζουν τη σημασία των υποχρεώσεων "by-design" στον Κανονισμό TN της ΕΕ, απαιτώντας τα συστήματα TN να σχεδιάζονται εξ αρχής με γνώμονα την προστασία της ιδιωτικότητας. Οι [288] τονίζουν την έμφαση που δίνεται στην επεξηγησιμότητα των συστημάτων TN, ιδιαίτερα για συστήματα υψηλού κινδύνου, ενώ οι [289] προτείνουν ένα εννοιολογικό πλαίσιο για την ανάπτυξη συστημάτων TN με επίγνωση της ιδιωτικότητας και της ασφάλειας.

4.2.4.3 Συγκρότηση πολιτικών και τεχνικών λύσεων στην εποχή του Generative AI

Το 2024 χαρακτηρίζεται από την ανάπτυξη εξειδικευμένων τεχνικών λύσεων για την ενίσχυση της προστασίας της ιδιωτικότητας στα LLMs. Μεταξύ των σημαντικότερων εξελίξεων είναι η εφαρμογή μηχανισμών φιλτραρίσματος ευαίσθητων δεδομένων και ανίχνευσης σεναρίων role-playing για την πρόληψη παραβιάσεων της ιδιωτικότητας [290], καθώς και η ενσωμάτωση τεχνικών διαφορετικής ιδιωτικότητας κατά την εκπαίδευση ή την εξαγωγή συμπερασμάτων για την ελαχιστοποίηση του κινδύνου αποκάλυψης ευαίσθητων πληροφοριών [291]. Επιπλέον, η εφαρμογή Zero-Knowledge Proofs (ZKPs) επιτρέπει την ανεξάρτητη επικύρωση του περιεχομένου που παράγεται από την TN, προωθώντας την εμπιστοσύνη και την προστασία της ιδιωτικότητας [292].

Παράλληλα, αναπτύσσονται προηγμένες τεχνικές ανωνυμοποίησης που εξισορροπούν την προστασία της ιδιωτικότητας με τη χρησιμότητα των δεδομένων. Η ομαδοποίηση βάσει ομοιότητας και διαφορετικότητας, σε συνδυασμό με τη δυναμική επιλογή χαρακτηριστικών με χρήση TN, έχει αποδειχθεί αποτελεσματική στη μείωση των κινδύνων ιδιωτικότητας διατηρώντας παράλληλα υψηλή χρησιμότητα δεδομένων [293], [294]. Επιπρόσθετα, η χρήση υβριδικών σχημάτων ανωνυμοποίησης που συνδυάζουν διαφορετική ιδιωτικότητα και k-ανωνυμία συμβάλλει στη διατήρηση της δομής και των τιμών των δεδομένων, μειώνοντας τους κινδύνους για την ιδιωτικότητα [295].

4.2.4.4 Προκλήσεις και μελλοντικές κατευθύνσεις στην εποχή των Γλωσσικών Μοντέλων

Οι προκλήσεις προστασίας της ιδιωτικότητας στα συστήματα TN έχουν σημαντικές επιπτώσεις τόσο σε κοινωνικό όσο και σε ατομικό επίπεδο, ιδιαίτερα σε ευαίσθητους τομείς όπως η υγειονομική περίθαλψη, τα χρηματοοικονομικά και η εκπαίδευση. Η εκτεταμένη χρήση μεγάλων συνόλων δεδομένων εγείρει ζητήματα συναίνεσης, ιδιοκτησίας δεδομένων και ατομικής αυτονομίας [296], ενώ τα συστήματα TN μπορούν να διαιωνίσουν και να ενισχύσουν υπάρχουσες κοινωνικές προκαταλήψεις [253]. Στον τομέα της υγείας, η ανάγκη εξισορρόπησης μεταξύ της χρησιμότητας των δεδομένων και της προστασίας της ιδιωτικότητας απαιτεί τη χρήση προηγμένων μεθόδων όπως η διαφορική ιδιωτικότητα και η κρυπτογράφηση [255], [297], ενώ στον χρηματοοικονομικό τομέα, τα ιδρύματα πρέπει να προστατεύουν τα προσωπικά και οικονομικά δεδομένα από παραβιάσεις που μπορεί να έχουν σοβαρές οικονομικές και φήμης συνέπειες [298].

Τα διδάγματα από το 2024 έχουν αναδείξει τη σημασία της ενσωμάτωσης αρχών Privacy by Design (PbD) και της ανάπτυξης προηγμένων μηχανισμών ανίχνευσης παραβιάσεων [299]. Η ανάγκη για αποκεντρωμένες πλατφόρμες TN και ισχυρά νομικά πλαίσια έχει επίσης τονιστεί ως κρίσιμη για την εξισορρόπηση των οφελών της TN με τα δικαιώματα ιδιωτικότητας [300]. Επιπλέον, η εμπειρία με συστήματα όπως το OpenAI Codex και το GitHub Copilot έχει αποκαλύψει τους κινδύνους ακούσιας δημιουργίας κώδικα που περιέχει ευαίσθητες πληροφορίες [301].

4.2.5 Σύγχρονα πλαίσια ηθικής καθοδήγησης

Καθώς η τεχνητή νοημοσύνη κατακτά ολοένα και πιο κεντρικό ρόλο στη λήψη αποφάσεων και στην παραγωγή περιεχομένου, τα σύγχρονα πλαίσια ηθικής καθοδήγησης έχουν εξελιχθεί σημαντικά ώστε να ανταποκρίνονται στις νέες προκλήσεις. Η ανάγκη για συστηματική ηθική καθοδήγηση έχει μετατοπιστεί από το πεδίο της θεωρητικής συζήτησης στο επίκεντρο της πρακτικής ανάπτυξης συστημάτων, με έμφαση στην ενσωμάτωση ηθικών αρχών στον σχεδιασμό, την εκπαίδευση και την εφαρμογή των συστημάτων τεχνητής νοημοσύνης.

4.2.5.1 Εκπαίδευση και ηθική κατάρτιση στην εποχή του Generative AI

Η εκπαίδευση και επαγγελματική κατάρτιση στον τομέα της ηθικής της TN έχει εξελιχθεί σημαντικά το 2024, με έμφαση στην αντιμετώπιση των προκλήσεων που θέτουν τα μοντέλα γενετικής TN και τα μεγάλα γλωσσικά μοντέλα. Τα εκπαιδευτικά ιδρύματα και οι οργανισμοί αναπτύσσουν ολοκληρωμένα πλαίσια ηθικής που εστιάζουν στην ανθρωποκεντρική προσέγγιση και την υπεύθυνη χρήση της TN [270], [303]. Παράλληλα, δίνεται ιδιαίτερη έμφαση στην ευαισθητοποίηση των εκπαιδευτικών σχετικά με τις ηθικές και παιδαγωγικές επιπτώσεις της χρήσης εργαλείων γενετικής TN στην εκπαίδευση [304].

Η συνεργασία μεταξύ εκπαιδευτικών ιδρυμάτων και επιχειρήσεων έχει οδηγήσει στην ανάπτυξη καινοτόμων προγραμμάτων κατάρτισης που ενσωματώνουν τεχνικές όπως η παιχνιδοποίηση και η διαδραστική μάθηση [305]. Τα προγράμματα αυτά στοχεύουν στην αντιμετώπιση ζητημάτων όπως η λογοκλοπή, η προστασία προσωπικών δεδομένων και η μεροληψία στο περιεχόμενο που παράγεται από την TN, ενώ παράλληλα προωθούν τη διαφάνεια και τη λογοδοσία [306], [307]. Οι εκπαιδευτικοί οργανισμοί ενθαρρύνονται να συνεργάζονται στον καθορισμό πολιτικών ηθικής χρήσης της TN στις τάξεις, προωθώντας έναν διάλογο για την υπεύθυνη χρήση της [308].

4.2.5.2 Θεσμικά πλαίσια και πρότυπα στην εποχή της Generative AI

Το 2024 χαρακτηρίζεται από την ανάδυση νέων θεσμικών πλαισίων και προτύπων για τη διακυβέρνηση των συστημάτων γενετικής ΤΝ, με έμφαση στην εναρμόνιση των νομικών πλαισίων σε διεθνές επίπεδο. Η Ευρωπαϊκή Ένωση πρωτοστατεί με την προσαρμογή του Artificial Intelligence Act (AIA) στις νέες προκλήσεις, εστιάζοντας σε ζητήματα ευθύνης, προστασίας προσωπικών δεδομένων και κυβερνοασφάλειας [264]. Παράλληλα, τα πανεπιστήμια αναπτύσσουν πλαίσια για την υπεύθυνη χρήση της γενετικής ΤΝ στην έρευνα, συμβάλλοντας στη διαμόρφωση ηθικών ερευνητικών πρακτικών [309].

Ιδιαίτερη έμφαση δίνεται στη συν-ρύθμιση και τη θέσπιση τεχνικών προτύπων για τον προσδιορισμό και την αντιμετώπιση των κινδύνων της ΤΝ [217]. Η διαφάνεια και η λογοδοσία αποτελούν κεντρικούς πυλώνες των νέων ρυθμιστικών πλαισίων, με την εφαρμογή νομικών και κανονιστικών μηχανισμών που προωθούν τη δημοκρατική νομιμότητα και τη συμμετοχή [310]. Η συνεργασία μεταξύ υπευθύνων χάραξης πολιτικής, ηγετών της βιομηχανίας και ηθικών φιλοσόφων αναδεικνύεται ως καθοριστικός παράγοντας για τη δημιουργία συνεκτικών ρυθμιστικών περιβαλλόντων [311].

4.2.5.3 Πρακτικές προκλήσεις στην εφαρμογή ηθικών αρχών στη Generative AI

Οι πρακτικές προκλήσεις στην εφαρμογή ηθικών αρχών για τα συστήματα παραγωγικής ΤΝ, σήμερα, επικεντρώνονται κυρίως στην προστασία της ιδιωτικότητας, την αντιμετώπιση προκαταλήψεων και την καταπολέμηση της παραπληροφόρησης. Για την αντιμετώπιση αυτών των προκλήσεων, οι οργανισμοί υιοθετούν ολοκληρωμένα πλαίσια διακυβέρνησης και συμμόρφωσης, όπως το μοντέλο Three Lines of Defense (TLoD), που κατανέμει τις ευθύνες μεταξύ διαχείρισης, εποπτείας κινδύνων και εσωτερικού ελέγχου [312]. Η διαφάνεια και η λογοδοσία ενισχύονται μέσω της εφαρμογής ΧΑΙ και ανοικτής κοινοποίησης δεδομένων, ενώ τακτικοί έλεγχοι και ανεξάρτητες εξωτερικές αξιολογήσεις διασφαλίζουν τη συνεχή συμμόρφωση με τις ηθικές αρχές [313], [314]. Η εκπαίδευση των ενδιαφερόμενων μερών και η συμμετοχή τους στο σχεδιασμό των συστημάτων ΤΝ αναδεικνύονται ως καθοριστικοί παράγοντες για την ευθυγράμμιση της τεχνολογικής ανάπτυξης με τις κοινωνικές αξίες [315].

4.2.5.4 Συνεργασία και συντονισμός φορέων στην μετα-LLM εποχή

Το 2024 χαρακτηρίζεται από την ανάδυση νέων συνεργατικών προσεγγίσεων μεταξύ δημόσιου και ιδιωτικού τομέα για την προώθηση της ηθικής ανάπτυξης και εφαρμογής της ΤΝ. Ιδιαίτερη έμφαση δίνεται στη δημιουργία συμπράξεων για την αξιοποίηση της ΤΝ στην επίτευξη των Στόχων Βιώσιμης Ανάπτυξης των Ηνωμένων Εθνών, με έμφαση σε τομείς όπως η υγειονομική περίθαλψη, η εκπαίδευση και η περιβαλλοντική διαχείριση [316], [317]. Πρωτοβουλίες όπως το πλαίσιο Responsible AI for Service Excellence (RAISE) προωθούν τη μετασχηματιστική συνεργασία μεταξύ οργανισμών παροχής υπηρεσιών, υπευθύνων χάραξης πολιτικής και ερευνητών [318].

Παράλληλα, αναπτύσσονται καινοτόμοι μηχανισμοί και πλατφόρμες για τη διευκόλυνση της συνεργασίας μεταξύ διαφορετικών φορέων. Το παράδειγμα της AI-Enhanced Collaborative Theory (AIECT) προωθεί τη δυναμική συνεργασία μεταξύ συστημάτων ΤΝ και ανθρώπων ερευνητών, δίνοντας έμφαση στη συν-δημιουργία και την ανατροφοδότηση [277]. Αυτές οι προσεγγίσεις

συμπληρώνονται από πρωτοβουλίες για την εδραίωση ηθικά τεκμηριωμένων θεμελίων στην ακαδημαϊκή κοινότητα [302].

4.3 Συγκριτική Ανάλυση Ευρημάτων

Η μετάβαση από τα παραδοσιακά συστήματα ΤΝ στα σύγχρονα συστήματα διαλόγου και παραγωγής περιεχομένου έχει επιφέρει ουσιαστικές αλλαγές στην προσέγγιση των ηθικών ζητημάτων, των ρυθμιστικών πλαισίων και των μηχανισμών προστασίας, αντανακλώντας την ωρίμανση του πεδίου και την αυξανόμενη κατανόηση των προκλήσεων που θέτει η εξέλιξη της τεχνολογίας.

4.3.1 Διαχρονική εξέλιξη των ηθικών ζητημάτων

Η συγκριτική ανάλυση των ευρημάτων μεταξύ των περιόδων 2015-2019 και 2024 αναδεικνύει μια σημαντική εξέλιξη στην προσέγγιση των ηθικών ζητημάτων της ΤΝ. Ενώ την περίοδο 2015-2019 η έμφαση δινόταν κυρίως στην αντιμετώπιση της αλγοριθμικής δικαιοσύνης ως τεχνικού ζητήματος και στην ανάπτυξη μεθόδων για την αύξηση της διαφάνειας και ερμηνευσιμότητας των αλγορίθμων, η περίοδος 2024 χαρακτηρίζεται από μια πιο ολιστική προσέγγιση. Η συζήτηση έχει διευρυνθεί για να συμπεριλάβει τις ηθικές διαστάσεις των μεγάλων γλωσσικών μοντέλων και της γενετικής ΤΝ, με ιδιαίτερη έμφαση στην ανάπτυξη εξειδικευμένων ταξινομιών για την αξιολόγηση και τον μετριασμό των προκαταλήψεων [253], [313].

4.3.2 Αλλαγές στο ρυθμιστικό πλαίσιο

Το ρυθμιστικό και κανονιστικό πλαίσιο έχει επίσης εξελιχθεί σημαντικά. Ενώ την περίοδο 2015-2019 το GDPR αποτελούσε το κύριο ρυθμιστικό εργαλείο με έμφαση στη διαφάνεια και τη λογοδοσία, η περίοδος 2024 χαρακτηρίζεται από την ανάδυση πιο εξειδικευμένων ρυθμιστικών πλαισίων, με τον Ευρωπαϊκό Κανονισμό για την ΤΝ (AI Act) να αποτελεί σημείο αναφοράς. Παράλληλα, η έμφαση στην ανθρωποκεντρική προσέγγιση έχει μετατοπιστεί από την απλή διαφάνεια στην αλληλεπίδραση ανθρώπου-αλγορίθμου σε πιο ολοκληρωμένες πρακτικές που προωθούν την ενεργή συμμετοχή των ενδιαφερομένων μερών και την εναρμόνιση των συστημάτων ΤΝ με τις ανθρώπινες αξίες [217], [315].

4.3.3 Νέες τάσεις και προκλήσεις

Η ανάλυση των νέων τάσεων στην ηθική της ΤΝ αναδεικνύει την εξέλιξη των προκλήσεων στην προστασία της ιδιωτικότητας και την ασφάλεια των δεδομένων. Ενώ παλαιότερα η έμφαση δινόταν στη συμμόρφωση με το GDPR, πλέον οι εταιρείες και οι ερευνητές επικεντρώνονται στην ανάπτυξη πιο αποτελεσματικών μεθόδων για τη διαχείριση της συναίνεσης και την προστασία των δεδομένων. Παράλληλα, η συζήτηση για την ανθρώπινη εποπτεία έχει εξελιχθεί σημαντικά, με αυξανόμενο προβληματισμό σχετικά με την υπόθεση ότι τα συστήματα ΤΝ μπορούν να είναι περισσότερο αμερόληπτα από τους ανθρώπους.

Στο πεδίο του *soft law*, οι προκλήσεις παραμένουν σημαντικές, με συνεχιζόμενες ελλείψεις στην παροχή σαφούς καθοδήγησης για τη σύνταξη και εφαρμογή πολιτικών ΤΝ. Η κατάσταση περιπλέκεται περαιτέρω από την έλλειψη σταθερής νομοθεσίας σχετικά με την απόδοση ευθυνών

σε τομείς που χρησιμοποιούν υπεραυτόνομα συστήματα ΤΝ. Αυτό έχει οδηγήσει στην ανάδυση νέων προσεγγίσεων διακυβέρνησης, όπως το μοντέλο TLoD, που στοχεύει στην καλύτερη κατανομή των ευθυνών μεταξύ διαφορετικών επιπέδων εποπτείας.

4.3.4 Μελλοντικές κατευθύνσεις

Η έρευνα στο πεδίο της ηθικής της ΤΝ αναμένεται να εξελιχθεί προς πιο ολοκληρωμένες και διεπιστημονικές προσεγγίσεις, καθώς τα συστήματα γενετικής ΤΝ και τα μεγάλα γλωσσικά μοντέλα γίνονται όλο και πιο διαδεδομένα στην καθημερινή ζωή. Οι μελλοντικές κατευθύνσεις της έρευνας αναμένεται να επικεντρωθούν στην ανάπτυξη ενσωματωμένων μηχανισμών ηθικής στην αρχιτεκτονική των συστημάτων ΤΝ, όπως υποδεικνύουν οι πρόσφατες εξελίξεις στην ΧΑΙ και τα πλαίσια ηθικής διακυβέρνησης. Παράλληλα, η αυξανόμενη συνειδητοποίηση των κοινωνικών επιπτώσεων της ΤΝ πιθανόν να οδηγήσει σε εντατικοποίηση της έρευνας για την ανάπτυξη πιο εύρωστων μεθόδων αξιολόγησης των ηθικών επιπτώσεων, με έμφαση στη συμμετοχική σχεδίαση και τη διαβούλευση με τα ενδιαφερόμενα μέρη. Επιπλέον, η εναρμόνιση των νομικών και ρυθμιστικών πλαισίων σε παγκόσμιο επίπεδο αναμένεται να αποτελέσει κεντρικό άξονα της μελλοντικής έρευνας, καθώς η ανάγκη για διεθνή συνεργασία και συντονισμό γίνεται ολοένα και πιο επιτακτική.

5

Συμπεράσματα

Στο παρόν κεφάλαιο παρουσιάζονται τα βασικά συμπεράσματα της διευρυμένης συστηματικής βιβλιογραφικής ανασκόπησης για τα ηθικά ζητήματα της Τεχνητής Νοημοσύνης, καλύπτοντας δύο διακριτές χρονικές περιόδους: 2015-2019 και 2024. Η αρχική έρευνα, που ξεχώρισε 179 μελέτες πάνω από 4.000 επιστημονικές δημοσιεύσεις της περιόδου 2015-2019, συμπληρώνεται από την ανάλυση πρόσφατων ερευνητικών εργασιών της περιόδου 2024, επιτρέποντας μια συγκριτική θεώρηση της εξέλιξης του πεδίου. Ιδιαίτερη έμφαση δίνεται στις επιπτώσεις της ανάπτυξης της Generative AI στα ζητήματα ηθικής, καθώς και στην ωρίμανση της ακαδημαϊκής και επιστημονικής κοινότητας στην αντιμετώπιση των σχετικών προκλήσεων. Το κεφάλαιο διαρθρώνεται σε τέσσερις ενότητες: αρχικά παρουσιάζονται τα συμπεράσματα για κάθε χρονική περίοδο ξεχωριστά, ακολουθεί η συγκριτική θεώρηση και εξέλιξη του πεδίου, και τέλος διατυπώνονται προτάσεις για μελλοντική έρευνα στον τομέα της ηθικής της Τεχνητής Νοημοσύνης.

5.1 Συμπεράσματα Περιόδου 2015-2019

Η ανάλυση της περιόδου 2015-2019 ανέδειξε θεμελιώδη ζητήματα στην ηθική της Τεχνητής Νοημοσύνης, εστιάζοντας στη διασφάλιση της ηθικής χρήσης, τις προϋποθέσεις ασφαλούς εφαρμογής, τα ζητήματα υπευθυνότητας και ηθικού σχεδιασμού, την προστασία της ιδιωτικότητας, και την ανάγκη ηθικής καθοδήγησης, θέτοντας τις βάσεις για την περαιτέρω εξέλιξη του πεδίου.

5.1.1 Σύνοψη ευρημάτων

Η συστηματική βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε ανέδειξε μια σειρά από σημαντικά ευρήματα σχετικά με τα ηθικά ζητήματα της Τεχνητής Νοημοσύνης και τη σχέση τους με το κανονιστικό πλαίσιο προστασίας δεδομένων.

Αναφορικά με τη διασφάλιση της ηθικής χρήσης της ΤΝ, η έρευνα κατέδειξε την κρίσιμη σημασία της αλγοριθμικής δικαιοσύνης και λογοδοσίας. Η αδιαφάνεια των αλγορίθμων αναδείχθηκε ως μείζον ζήτημα, με την ανάγκη για συγκεκριμένους περιορισμούς και πλαίσια ελέγχου να είναι επιτακτική. Ιδιαίτερη έμφαση δόθηκε στην έννοια της περιβάλλουσας (envelopment) ως εναλλακτική προσέγγιση στο ζήτημα της διαφάνειας, καθώς και στη σημασία των συστημάτων ευφυσούς βιντεοεπιτήρησης.

Σχετικά με τις προϋποθέσεις ασφαλούς χρήσης της ΤΝ, τα ευρήματα υπογράμμισαν την ανάγκη για ισχυρά θεσμικά πλαίσια και πρότυπα. Η έρευνα ανέδειξε τρεις θεμελιώδεις πυλώνες: την υπευθυνότητα, τη διαφάνεια, και την ενσωμάτωση δικαιοσύνης και ηθικής. Η δυνατότητα επαλήθευσης και επικύρωσης των συστημάτων ΤΝ αναδείχθηκε ως κρίσιμο στοιχείο, ιδιαίτερα για εφαρμογές που αφορούν την ασφάλεια.

Στο ζήτημα της υπευθυνότητας και του ηθικού σχεδιασμού, η έρευνα κατέδειξε την ανάγκη για σαφή καθορισμό των ορίων ευθύνης και την αντιμετώπιση των προκλήσεων που θέτει η αυτονομία των συστημάτων ΤΝ. Η αυτοματοποιημένη λήψη αποφάσεων και η εποπτεία αναδείχθηκαν ως κρίσιμα ζητήματα που απαιτούν προσεκτική διαχείριση και ρύθμιση.

Όσον αφορά την προστασία της ιδιωτικότητας, τα ευρήματα υπογράμμισαν τον κεντρικό ρόλο των μεγάλων τεχνολογικών εταιρειών στη διαμόρφωση πολιτικών προστασίας δεδομένων. Η έρευνα ανέδειξε τις προκλήσεις που αντιμετωπίζουν οι αναπτυσσόμενες χώρες στην προστασία των δεδομένων των πολιτών τους, καθώς και τη σημασία της διεθνούς συνεργασίας στη διαμόρφωση προτύπων προστασίας της ιδιωτικότητας.

Τέλος, σχετικά με την ανάγκη ηθικής καθοδήγησης, η έρευνα τόνισε τη σημασία της ενσωμάτωσης ηθικών αρχών στη διδασκαλία της μηχανικής μάθησης και την ανάγκη για δεοντολογικά πρότυπα στην οικοδόμηση αξιόπιστων συστημάτων. Ιδιαίτερη έμφαση δόθηκε στο ρόλο των εκπαιδευτικών ιδρυμάτων και των επιχειρήσεων στην προώθηση της ηθικής συνείδησης, καθώς και στην ανάγκη για συνεργασία μεταξύ κυβέρνησης και ιδιωτικού τομέα στη διαμόρφωση κατάλληλων ρυθμιστικών πλαισίων.

Τα παραπάνω ευρήματα συνθέτουν ένα πολύπλοκο πλέγμα αλληλεξαρτώμενων ζητημάτων που απαιτούν συντονισμένη δράση από όλους τους εμπλεκόμενους φορείς για την επίτευξη μιας ηθικά υπεύθυνης και ασφαλούς ανάπτυξης και εφαρμογής της Τεχνητής Νοημοσύνης.

5.1.2 Απαντήσεις στα ερευνητικά ερωτήματα

Με βάση την ανάλυση της βιβλιογραφίας, μπορούμε να παρέχουμε συγκεκριμένες απαντήσεις στα πέντε ερευνητικά ερωτήματα που τέθηκαν στην αρχή της έρευνας:

1. Τι πρέπει να γίνει για να διασφαλιστεί η ηθική χρήση της Τεχνητής Νοημοσύνης;

Η διασφάλιση της ηθικής χρήσης της ΤΝ απαιτεί μια πολυεπίπεδη προσέγγιση που περιλαμβάνει:

- Την εφαρμογή συγκεκριμένων περιορισμών και πλαισίων ελέγχου για την αντιμετώπιση της αλγοριθμικής αδιαφάνειας
- Την υιοθέτηση της προσέγγισης της περιβάλλουσας (envelopment) για τη διαχείριση αδιαφανών αλγορίθμων
- Τη θέσπιση σαφών μηχανισμών λογοδοσίας και ευθύνης
- Την ανάπτυξη ισχυρών κανονιστικών πλαισίων, συμπεριλαμβανομένου του μη δεσμευτικού δικαίου (soft law)

2. Ποιες είναι οι προϋποθέσεις για την ασφαλή χρήση της Τεχνητής Νοημοσύνης;

Οι βασικές προϋποθέσεις περιλαμβάνουν:

- Την εφαρμογή των τριών θεμελιωδών πυλώνων: υπευθυνότητα, διαφάνεια και ενσωμάτωση δικαιοσύνης
- Την ανάπτυξη αποτελεσματικών μηχανισμών ασφάλειας και προστασίας
- Τη διασφάλιση της διαφάνειας και επεξηγησιμότητας των συστημάτων
- Την προετοιμασία του εργατικού δυναμικού για τις προκλήσεις της ΤΝ
- Τη διεθνή συνεργασία για τη θέσπιση κοινών προτύπων ασφαλείας

3. Μπορεί να θεωρηθεί η Τεχνητή Νοημοσύνη υπεύθυνη (ηθικός σχεδιασμός);

Η έρευνα καταδεικνύει ότι:

- Η υπευθυνότητα της ΤΝ είναι εφικτή μέσω κατάλληλου σχεδιασμού και ελέγχων
- Απαιτείται σαφής καθορισμός των ορίων ευθύνης και αυτονομίας
- Η αυτοματοποιημένη λήψη αποφάσεων πρέπει να συνοδεύεται από επαρκή εποπτεία
- Είναι απαραίτητη η ενσωμάτωση ηθικών αρχών στο σχεδιασμό των συστημάτων ΤΝ

4. Πώς αντιμετωπίζει η Τεχνητή Νοημοσύνη ζητήματα ιδιωτικότητας;

Η προστασία της ιδιωτικότητας διασφαλίζεται μέσω:

- Της συμμόρφωσης με το ρυθμιστικό πλαίσιο προστασίας δεδομένων (GDPR)
- Της ανάπτυξης μηχανισμών προστασίας από τις μεγάλες τεχνολογικές εταιρείες
- Της διεθνούς συνεργασίας για τη θέσπιση προτύπων προστασίας
- Της ιδιαίτερης μέριμνας για τις αναπτυσσόμενες χώρες

5. Κρίνεται αναγκαία η ηθική και δεοντολογική καθοδήγηση σε ζητήματα Τεχνητής Νοημοσύνης;

Τα ευρήματα καταδεικνύουν ότι:

- Η ηθική καθοδήγηση είναι απαραίτητη για την υπεύθυνη ανάπτυξη της ΤΝ
- Απαιτείται η ενσωμάτωση ηθικών αρχών στην εκπαίδευση των μηχανικών
- Είναι αναγκαία η συνεργασία μεταξύ κυβέρνησης και ιδιωτικού τομέα
- Χρειάζονται δεοντολογικά πρότυπα για την οικοδόμηση αξιόπιστων συστημάτων

Οι απαντήσεις αυτές αναδεικνύουν την πολυπλοκότητα και αλληλεξάρτηση των ζητημάτων που σχετίζονται με την ηθική διάσταση της ΤΝ, καθώς και την ανάγκη για συντονισμένη δράση σε πολλαπλά επίπεδα.

5.1.3 Κύρια συμπεράσματα και διαπιστώσεις

Η διασφάλιση της ηθικής και υπεύθυνης χρήσης της ΤΝ βρέθηκε στο επίκεντρο του ενδιαφέροντος, με έμφαση σε θέματα όπως η αλγοριθμική δικαιοσύνη, η λογοδοσία και η διαφάνεια των συστημάτων ΤΝ. Παράλληλα, αναδείχθηκε η ανάγκη για ενίσχυση του ρυθμιστικού και κανονιστικού πλαισίου, καθώς και για υιοθέτηση μιας ανθρωποκεντρικής προσέγγισης στην ανάπτυξη και χρήση της ΤΝ.

Σημαντικό ζήτημα αποτέλεσε επίσης η διαμόρφωση των προϋποθέσεων για την ασφαλή χρήση της ΤΝ, μέσα από τη θέσπιση θεμελιωδών αρχών, την ανάπτυξη μηχανισμών ασφάλειας και προστασίας, αλλά και τη διασφάλιση της διαφάνειας και επεξηγησιμότητας των συστημάτων ΤΝ.

Ταυτόχρονα, τονίστηκε η σημασία της υπευθυνότητας και του ηθικού σχεδιασμού, με έμφαση στον ορισμό σαφών δεοντολογικών παραμέτρων, τον καθορισμό ορίων στην αυτονομία των συστημάτων ΤΝ και την εποπτεία της αυτοματοποιημένης λήψης αποφάσεων.

Τέλος, αναδείχθηκε η κομβική σημασία της προστασίας της ιδιωτικότητας και των προσωπικών δεδομένων στην εποχή της ΤΝ, μέσα από την ανάπτυξη κατάλληλων ρυθμιστικών πλαισίων, πολιτικών και τεχνικών λύσεων. Παράλληλα, επισημάνθηκε η ανάγκη για ηθική καθοδήγηση μέσω της εκπαίδευσης, της θέσπισης προτύπων και της συνεργασίας μεταξύ φορέων.

5.2 Συμπεράσματα Περιόδου 2024

Η ανάλυση της περιόδου 2024 φέρνει στο προσκήνιο νέες προκλήσεις και εξελίξεις στην ηθική της Τεχνητής Νοημοσύνης αναδεικνύοντας την ανάγκη για προσαρμογή των υφιστάμενων πλαισίων και την ανάπτυξη νέων προσεγγίσεων στην αντιμετώπιση των αναδυόμενων ηθικών ζητημάτων.

5.2.1 Σύνοψη νέων ευρημάτων

Η ανάλυση της βιβλιογραφίας για την περίοδο 2024 ανέδειξε σημαντικές εξελίξεις και νέες τάσεις στον τομέα της ηθικής χρήσης της Τεχνητής Νοημοσύνης. Στο πεδίο της αλγοριθμικής δικαιοσύνης, παρατηρήθηκε η ανάπτυξη ολοκληρωμένων ταξινομιών για την αξιολόγηση και τον μετριασμό των προκαταλήψεων στα LLMs και τη GenAI, καθώς και η εισαγωγή εργαλείων όπως το FairTransLing που ενισχύουν την ερμηνευσιμότητα και τη δικαιοσύνη στις διαδικασίες λήψης αποφάσεων.

Παράλληλα, δόθηκε έμφαση στην ανθρώπινη επίβλεψη και στα υβριδικά συστήματα συνεργασίας ανθρώπου-ΤΝ για τη βέλτιστη λειτουργία, ενώ πλαίσια όπως το DIKE συνέβαλαν στην ενίσχυση της ικανότητας των LLMs να εσωτερικεύουν ανθρώπινες αξίες και να προσαρμόζονται πολιτισμικά. Σε επίπεδο ρυθμιστικού πλαισίου, ο AI Act καθιέρωσε προσπάθειες για τη ρύθμιση των μοντέλων αιχμής, ενώ προτάθηκαν εθνικά μητρώα για τα μοντέλα ΤΝ και συστήματα έγκρισης παρόμοια με αυτά του FDA.

Ωστόσο, εκφράστηκαν και ανησυχίες σχετικά με τη διαφάνεια, την ιδιωτικότητα, την ασφάλεια και τη διατήρηση της ανθρώπινης αυτονομίας. Για την αντιμετώπισή τους, δόθηκε έμφαση στη συνεργασία δημόσιου και ιδιωτικού τομέα, στην ανάπτυξη διεθνών συνεργασιών για τη δημιουργία κοινών προτύπων, καθώς και στην αναβάθμιση δεξιοτήτων του υπάρχοντος εργατικού δυναμικού μέσω προγραμμάτων εκπαίδευσης. Τέλος, στο πεδίο της προστασίας δεδομένων και ιδιωτικότητας, παρατηρήθηκε η εφαρμογή τεχνικών όπως η διαφορική ιδιωτικότητα, η κωδικοποίηση αντικατάστασης και τα Zero-Knowledge Proofs για ισχυρότερη προστασία.

5.2.2 Απαντήσεις στα ερευνητικά ερωτήματα υπό το πρίσμα των νέων δεδομένων

Με βάση τα νέα ευρήματα της περιόδου 2024, τα ερευνητικά ερωτήματα που τέθηκαν στη διπλωματική μπορούν να απαντηθούν ως εξής:

1. Τι πρέπει να γίνει για να διασφαλιστεί η ηθική χρήση της Τεχνητής Νοημοσύνης;

Για τη διασφάλιση της ηθικής χρήσης της ΤΝ, απαιτούνται:

- Ισχυροί μηχανισμοί αλγοριθμικής δικαιοσύνης και λογοδοσίας, ιδιαίτερα για τα μεγάλα γλωσσικά μοντέλα και τα συστήματα γενετικής ΤΝ.
- Ενσωμάτωση τεχνικών διασφαλίσεων και ηθικών παραμέτρων στην αρχιτεκτονική των μοντέλων
- Ύπαρξη κατάλληλων ρυθμιστικών πλαισίων, είναι επίσης κρίσιμης σημασίας.

2. Ποιες είναι οι προϋποθέσεις για την ασφαλή χρήση της Τεχνητής Νοημοσύνης;

Οι προϋποθέσεις για την ασφαλή χρήση της ΤΝ περιλαμβάνουν:

- Τήρηση θεμελιωδών αρχών όπως η διαφάνεια, η λογοδοσία και η προστασία της ιδιωτικότητας.
- Τεχνικοί μηχανισμοί ασφαλείας, όπως η ανταγωνιστική εκπαίδευση και η επαύξηση δεδομένων, σε συνδυασμό με
- Ολοκληρωμένη ενημέρωση και εκπαίδευση της κοινωνίας

3. Μπορεί να θεωρηθεί η Τεχνητή Νοημοσύνη υπεύθυνη (ηθικός σχεδιασμός);

Οι σύγχρονες προσεγγίσεις υπευθυνότητας της ΤΝ δίνουν έμφαση:

- Στην επίτευξη ισορροπίας μεταξύ της αυτονομίας των συστημάτων και της ανθρώπινης εποπτείας.
- Στον καθορισμό ξεκάθαρων ορίων υπευθυνότητας και
- Στην αντιμετώπιση των ευρύτερων κοινωνικοηθικών επιπτώσεων που αποτελούν κρίσιμες προκλήσεις.

4. Πώς αντιμετωπίζει η Τεχνητή Νοημοσύνη ζητήματα ιδιωτικότητας;

Η προστασία των προσωπικών δεδομένων στην εποχή των μεγάλων γλωσσικών μοντέλων απαιτεί:

- Νέες, εξειδικευμένες τεχνικές, όπως η διαφορική ιδιωτικότητα, η ανωνυμοποίηση και η κρυπτογράφηση.
- Κανονιστικά πλαίσια όπως ο Κανονισμός ΤΝ της ΕΕ

Τα παραπάνω χρειάζεται να προσαρμοστούν στις μοναδικές προκλήσεις αυτών των συστημάτων.

5. Κρίνεται αναγκαία η ηθική και δεοντολογική καθοδήγηση σε ζητήματα Τεχνητής Νοημοσύνης;

Τα σύγχρονα πλαίσια ηθικής καθοδήγησης για την ΤΝ, περιλαμβάνουν:

- Ανάπτυξη ολοκληρωμένων εκπαιδευτικών προγραμμάτων

- Διαμόρφωση νέων θεσμικών πλαισίων και τυποποιήσεων
- Ενίσχυση της συνεργασίας μεταξύ διαφορετικών φορέων του δημόσιου και του ιδιωτικού τομέα.

Η ηθική ευθυγράμμιση της τεχνολογικής προόδου με τις κοινωνικές αξίες αναδεικνύεται ως κεντρικό ζήτημα.

Συνοπτικά, η εμπειρία του έτους 2024 καταδεικνύει ότι η διασφάλιση μιας ηθικής και υπεύθυνης πορείας για την ΤΝ απαιτεί τη συντονισμένη δράση σε τεχνικό, ρυθμιστικό και κοινωνικό επίπεδο, με ιδιαίτερη έμφαση στην ανάπτυξη ανθεκτικών μοντέλων διακυβέρνησης και εκπαιδευτικών πλαισίων που να ανταποκρίνονται στις προκλήσεις των μεγάλων γλωσσικών μοντέλων και των συστημάτων παραγωγικής ΤΝ.

5.3 Συγκριτική Θεώρηση και Εξέλιξη του Πεδίου

Η συγκριτική ανάλυση των ευρημάτων των δύο περιόδων αναδεικνύει σημαντικές μεταβολές στην προσέγγιση των ηθικών ζητημάτων της Τεχνητής Νοημοσύνης, με την εμφάνιση του Generative AI να αποτελεί καταλύτη για την ωρίμανση του πεδίου και την ανάπτυξη πιο εξειδικευμένων και ολοκληρωμένων προσεγγίσεων στην αντιμετώπιση των ηθικών προκλήσεων.

5.3.1 Αλλαγές στην προσέγγιση της ηθικής διακυβέρνησης

Η συγκριτική ανάλυση των ευρημάτων για τις δύο υπό εξέταση περιόδους αναδεικνύει σημαντικές διαφοροποιήσεις στην προσέγγιση των ηθικών ζητημάτων που σχετίζονται με την Τεχνητή Νοημοσύνη.

Αρχικά, παρατηρείται μια μετάβαση από τις γενικότερες συζητήσεις περί ηθικής και διαφάνειας που χαρακτήριζαν την περίοδο 2015-2019, σε πιο εξειδικευμένα πλαίσια κατά την περίοδο 2024. Χαρακτηριστικό παράδειγμα αποτελεί ο Κανονισμός της ΕΕ για την Τεχνητή Νοημοσύνη (AI Act), ο οποίος θέτει συγκεκριμένους κανόνες και προδιαγραφές, με ιδιαίτερη έμφαση στην προστασία των δεδομένων και την αποφυγή προκαταλήψεων.

Επιπλέον, ενώ η περίοδος 2015-2019 χαρακτηριζόταν από την έμφαση στην αυτορρύθμιση και το soft law, η περίοδος 2024 σηματοδοτεί τη μετάβαση σε πιο υποχρεωτικά πλαίσια, με αυξημένη κυβερνητική παρέμβαση και τη δημιουργία νομικών προτύπων και εθνικών μητρώων, ιδίως στο πλαίσιο της Ευρωπαϊκής Ένωσης.

Παρά τις διαφοροποιήσεις, και οι δύο περίοδοι αναγνωρίζουν τη σημασία της εκπαίδευσης στα ζητήματα ηθικής της ΤΝ. Ωστόσο, η περίοδος 2024 δίνει ιδιαίτερη έμφαση στην ενσωμάτωση της ηθικής στην ίδια την αρχιτεκτονική των μοντέλων ΤΝ, καθώς και στην ανάγκη ενίσχυσης της συνεργασίας μεταξύ του δημόσιου και του ιδιωτικού τομέα για την αντιμετώπιση των προκλήσεων.

Τέλος, η εξέλιξη των LLMs κατά την περίοδο 2024 έφερε στο προσκήνιο νέους κινδύνους, όπως οι επιθέσεις ιδιωτικότητας, η χειραγωγή και η διασφάλιση της ακεραιότητας των πληροφοριών. Αυτά τα ζητήματα δεν είχαν τονιστεί τόσο έντονα κατά την περίοδο 2015-2019, αναδεικνύοντας την ανάγκη για συνεχή επαγρύπνηση και προσαρμογή των ηθικών πλαισίων στις ραγδαίες τεχνολογικές εξελίξεις.

5.3.2 Επίδραση του Generative AI στο πεδίο

Η ανάδυση και ταχεία εξέλιξη του GenAI κατά την περίοδο 2024 είχε σημαντικό αντίκτυπο στον τομέα της ηθικής της Τεχνητής Νοημοσύνης, εισάγοντας νέες προκλήσεις και μετατοπίζοντας τις προτεραιότητες σε σύγκριση με την περίοδο 2015-2019.

Αρχικά, το GenAI οδήγησε σε μια γενική αύξηση της πολυπλοκότητας των ηθικών ζητημάτων. Ενώ η περίοδος 2015-2019 επικεντρώθηκε σε θεμελιώδη θέματα όπως η αλγοριθμική δικαιοσύνη, η διαφάνεια και η λογοδοσία, η εμφάνιση του GenAI έφερε στο προσκήνιο νέες προκλήσεις. Αυτές περιλαμβάνουν τη διαχείριση των ψευδαισθήσεων που μπορεί να παράγουν τα μοντέλα, την αντιμετώπιση των κινδύνων ασφάλειας που απορρέουν από την ευκολότερη δημιουργία πειστικών αλλά ψευδών δεδομένων, και την ανάγκη για ενσωμάτωση ηθικών φραγμών απευθείας στην αρχιτεκτονική των μοντέλων.

Παράλληλα, παρατηρήθηκε μια μετατόπιση της έμφασης από τη ρύθμιση στην εποπτεία. Ενώ στην αρχική περίοδο η συζήτηση αφορούσε κυρίως τη δημιουργία ρυθμιστικών πλαισίων και κανόνων, με το GenAI η έμφαση μετατοπίστηκε στην ανάγκη για εποπτεία και λογοδοσία σε πραγματικό χρόνο, καθώς οι επιπτώσεις και οι κίνδυνοι έγιναν πιο δυναμικοί και λιγότερο προβλέψιμοι.

Η περίοδος 2024 χαρακτηρίστηκε επίσης από μια έκρηξη στην ανάπτυξη εξειδικευμένων τεχνικών για την αντιμετώπιση των προκλήσεων που θέτει το GenAI. Αυτές περιλαμβάνουν μεθόδους για την ενίσχυση της ερμηνευσιμότητας (π.χ., TokenSHAP), πλαίσια για την ανίχνευση και διόρθωση προκαταλήψεων (π.χ., FAIR-Frame), καθώς και νέες μετρικές για την αξιολόγηση της δικαιοσύνης (π.χ., NFG Score).

Επιπλέον, το GenAI δημιούργησε νέες ανησυχίες για την ιδιωτικότητα, καθώς τα μοντέλα μπορούν να παράγουν ρεαλιστικά δεδομένα που μπορεί να χρησιμοποιηθούν για την ταυτοποίηση ατόμων ή την αποκάλυψη ευαίσθητων πληροφοριών. Ταυτόχρονα, η εκπαίδευση των μοντέλων σε τεράστιες ποσότητες δεδομένων που μπορεί να περιέχουν προκαταλήψεις, οδήγησε στην ένταση των προκαταλήψεων και στην ανάγκη για περαιτέρω προσαρμογή της αλγοριθμικής δικαιοσύνης στα νέα δεδομένα.

Τέλος, η εποχή του GenAI ανέδειξε την επιτακτική ανάγκη για πολυεπίπεδη συνεργασία μεταξύ κυβερνήσεων, βιομηχανίας και ακαδημαϊκών ιδρυμάτων, προκειμένου να αναπτυχθούν πρότυπα, κανονισμοί και πρακτικές που θα αντιμετωπίζουν αποτελεσματικά τις μοναδικές προκλήσεις που θέτουν τα νέα αυτά μοντέλα.

5.3.3 Ωρίμανση της ακαδημαϊκής και επιστημονικής κοινότητας

Η περίοδος 2024 χαρακτηρίζεται από μια αυξημένη εστίαση στην εφαρμοσμένη έρευνα, σε αντίθεση με την περίοδο 2015-2019 που επικεντρώθηκε σε μεγάλο βαθμό σε θεωρητικές συζητήσεις και στον καθορισμό βασικών εννοιών. Η έμφαση πλέον δίνεται σε πρακτικές προκλήσεις και συγκεκριμένες λύσεις για την ενσωμάτωση ηθικών αρχών στην ανάπτυξη και χρήση συστημάτων ΤΝ.

Παράλληλα, οι τεχνικές προσεγγίσεις έχουν εξελιχθεί από γενικές, επικεντρωμένες κυρίως στην αλγοριθμική δικαιοσύνη και την επεξηγησιμότητα, σε πιο εξειδικευμένες τεχνικές, όπως τα

ZKPs, η διαφορική ιδιωτικότητα και η ενσωμάτωση ασφαλών αρχιτεκτονικών, με στόχο την αντιμετώπιση συγκεκριμένων απειλών και προκλήσεων.

Επιπλέον, η έρευνα της περιόδου 2024 εξετάζει πιο αναλυτικά το ρυθμιστικό πλαίσιο, συμπεριλαμβανομένου του EU AI Act, ενώ δίνεται έμφαση στην εναρμόνιση των νομικών πλαισίων σε διεθνές επίπεδο. Σε αντίθεση, η περίοδος 2015-2019 περιορίστηκε σε αναφορές στο GDPR και σε επίπεδο ανάλυσης αρχών και απαιτήσεων.

Η ακαδημαϊκή κοινότητα αντιμετωπίζει πλέον πιο σύνθετες προκλήσεις που σχετίζονται με τα μεγάλα γλωσσικά μοντέλα (LLMs) και την παραγωγική ΤΝ, όπως η αντιμετώπιση ψευδαισθήσεων, η διαχείριση ευαίσθητων δεδομένων και η παραβίαση ιδιωτικότητας, σε αντίθεση με την περίοδο 2015-2019 που επικεντρώθηκε σε γενικές προκλήσεις όπως η αδιαφάνεια, η μεροληψία και η ιδιωτικότητα.

Τέλος, ενώ και στις δύο περιόδους αναγνωρίζεται η σημασία της συνεργασίας μεταξύ κυβέρνησης και ιδιωτικού τομέα, η περίοδος 2024 δίνει ιδιαίτερη έμφαση στην ενίσχυση της συνεργασίας μεταξύ ακαδημαϊκών ιδρυμάτων, βιομηχανίας και κυβερνητικών φορέων για την υπεύθυνη ανάπτυξη της ΤΝ.

5.4 Προτάσεις για Μελλοντική Έρευνα

Η παρούσα εργασία έδωσε χώρο σε νέα ερευνητικά ερωτήματα και κατευθύνσεις για μελλοντική έρευνα, με έμφαση στην αντιμετώπιση των προκλήσεων που θέτουν τα μεγάλα γλωσσικά μοντέλα, καθώς και στην ανάγκη για διεπιστημονικές προσεγγίσεις και συνεργασίες για την ανάπτυξη αποτελεσματικών λύσεων.

5.4.1 Αναδύομενα ερευνητικά ερωτήματα

Με βάση τα ευρήματα της παρούσας διπλωματικής εργασίας και τη διερεύνηση των ερευνητικών ερωτημάτων σχετικά με την ηθική χρήση της Τεχνητής Νοημοσύνης, προκύπτουν νέα ερωτήματα που μπορούν να αποτελέσουν αντικείμενο περαιτέρω έρευνας:

1. Πώς μπορεί να ενισχυθεί η διαφάνεια και η επεξηγησιμότητα των συστημάτων Τεχνητής Νοημοσύνης, ώστε να διασφαλιστεί η εμπιστοσύνη των χρηστών και να ελαχιστοποιηθούν οι κίνδυνοι από την αδιαφανή λήψη αποφάσεων;
2. Ποιες είναι οι βέλτιστες πρακτικές για την ενσωμάτωση ηθικών αρχών και αξιών στον σχεδιασμό και την ανάπτυξη συστημάτων Τεχνητής Νοημοσύνης, και πώς μπορούν να προσαρμοστούν σε διαφορετικά πεδία εφαρμογής;
3. Πώς μπορούν να αντιμετωπιστούν τα ζητήματα μεροληψίας και διακρίσεων που ενδέχεται να προκύψουν από τη χρήση αλγοριθμικών συστημάτων λήψης αποφάσεων, ιδιαίτερα σε ευαίσθητους τομείς όπως η υγεία, η εκπαίδευση και η απασχόληση;
4. Ποιος είναι ο ρόλος της κυβερνητικής πολιτικής και των ρυθμιστικών παρεμβάσεων στη διαμόρφωση ενός ηθικού και υπεύθυνου πλαισίου για την ανάπτυξη και χρήση της Τεχνητής Νοημοσύνης, και πώς μπορεί να επιτευχθεί η ισορροπία μεταξύ καινοτομίας και προστασίας του δημοσίου συμφέροντος;

5. Ποιες είναι οι προϋποθέσεις για την ανάπτυξη μιας κουλτούρας ηθικής και δεοντολογικής καθοδήγησης στον τομέα της Τεχνητής Νοημοσύνης, και πώς μπορεί να προωθηθεί η συνεργασία μεταξύ ερευνητών, επαγγελματιών, φορέων χάραξης πολιτικής και της κοινωνίας των πολιτών;

Τα παραπάνω ερωτήματα αντανακλούν τις προκλήσεις και τις ευκαιρίες που αναδύονται από την ταχεία εξέλιξη της Τεχνητής Νοημοσύνης και την ανάγκη για μια ολιστική προσέγγιση στην αντιμετώπιση των ηθικών διλημάτων που προκύπτουν. Η διερεύνησή τους μπορεί να συμβάλει στην ανάπτυξη ενός ισχυρού θεωρητικού και πρακτικού πλαισίου για την υπεύθυνη και ηθική χρήση της Τεχνητής Νοημοσύνης προς όφελος της κοινωνίας.

5.4.2 Προτεινόμενες κατευθύνσεις

Σε συνέχεια των αναδυόμενων ερωτημάτων που διατυπώθηκαν στην προηγούμενη ενότητα, προτείνονται οι ακόλουθες κατευθύνσεις για μελλοντική έρευνα στο πεδίο της ηθικής της Τεχνητής Νοημοσύνης:

1. Ανάπτυξη καινοτόμων τεχνικών και μεθοδολογιών για την ενίσχυση της διαφάνειας και της επεξηγησιμότητας των συστημάτων Τεχνητής Νοημοσύνης, όπως η χρήση τεχνικών οπτικοποίησης, η ανάπτυξη διεπαφών φυσικής γλώσσας και η δημιουργία εργαλείων ελέγχου και επαλήθευσης.
2. Εμπειρικές μελέτες για την αξιολόγηση της αποτελεσματικότητας διαφορετικών προσεγγίσεων στην αντιμετώπιση ζητημάτων μεροληψίας και διακρίσεων σε αλγοριθμικά συστήματα λήψης αποφάσεων, καθώς και για τον εντοπισμό βέλτιστων πρακτικών.
3. Ανάλυση του ρόλου της κυβερνητικής πολιτικής και των ρυθμιστικών παρεμβάσεων στη διαμόρφωση ενός ηθικού πλαισίου για την Τεχνητή Νοημοσύνη, μέσω της διερεύνησης διαφορετικών νομικών και θεσμικών μοντέλων σε διεθνές επίπεδο.
4. Έρευνα για την ανάπτυξη προηγμένων τεχνολογιών προστασίας της ιδιωτικότητας και των προσωπικών δεδομένων, όπως η κρυπτογράφηση, η ανωνυμοποίηση και οι τεχνικές διαφορικής ιδιωτικότητας, με έμφαση στην εφαρμογή τους σε συστήματα Τεχνητής Νοημοσύνης.
5. Διερεύνηση των παραγόντων που συμβάλλουν στην ανάπτυξη μιας κουλτούρας ηθικής και δεοντολογικής καθοδήγησης στον τομέα της Τεχνητής Νοημοσύνης, μέσω της μελέτης επιτυχημένων πρωτοβουλιών και της ανάλυσης των αναγκών και των προσδοκιών των εμπλεκόμενων φορέων.

5.4.3 Μεθοδολογικές προτάσεις

Αρχικά, η διεξαγωγή συστηματικών ανασκοπήσεων και μετα-αναλύσεων της υπάρχουσας βιβλιογραφίας μπορεί να συμβάλει στη σύνθεση των ευρημάτων και στον εντοπισμό κενών και ευκαιριών για περαιτέρω έρευνα. Παράλληλα, η ανάπτυξη θεωρητικών και εννοιολογικών πλαισίων που ενσωματώνουν έννοιες και αρχές από διαφορετικούς επιστημονικούς κλάδους, όπως η ηθική φιλοσοφία, η πληροφορική, η κοινωνιολογία και οι νομικές επιστήμες, μπορεί να παρέχει

ένα ολοκληρωμένο υπόβαθρο για τη διερεύνηση των ηθικών διαστάσεων της Τεχνητής Νοημοσύνης.

Επιπλέον, η διεξαγωγή εμπειρικών μελετών, συμπεριλαμβανομένων ερευνών, πειραμάτων και μελετών περίπτωσης, μπορεί να παρέχει πολύτιμα πρωτογενή δεδομένα σχετικά με τις αντιλήψεις, τις στάσεις και τις εμπειρίες των εμπλεκόμενων φορέων. Η χρήση προσομοιώσεων και μοντέλων μπορεί επίσης να συμβάλει στη διερεύνηση των επιπτώσεων διαφορετικών σεναρίων και στρατηγικών σχετικά με την ηθική διακυβέρνηση της Τεχνητής Νοημοσύνης.

Ιδιαίτερη έμφαση θα πρέπει να δοθεί στην ανάπτυξη και αξιολόγηση καινοτόμων τεχνολογικών λύσεων και εργαλείων που ενσωματώνουν ηθικές αρχές και αξίες στον σχεδιασμό και τη λειτουργία συστημάτων Τεχνητής Νοημοσύνης. Η προώθηση της διεπιστημονικής συνεργασίας και της ανταλλαγής γνώσεων μεταξύ ερευνητών, επαγγελματιών, φορέων χάραξης πολιτικής και εκπροσώπων της κοινωνίας των πολιτών αποτελεί επίσης κρίσιμο παράγοντα για την επίτευξη προόδου σε αυτόν τον τομέα.

Τέλος, η ενσωμάτωση της έρευνας σε πραγματικές συνθήκες και η υιοθέτηση συμμετοχικών προσεγγίσεων, με την ενεργό εμπλοκή των ενδιαφερόμενων μερών στη διαδικασία σχεδιασμού, υλοποίησης και αξιολόγησης των ερευνητικών έργων, μπορεί να ενισχύσει τη συνάφεια και τον αντίκτυπο της έρευνας στην πράξη.

Συμπερασματικά, η παρούσα διπλωματική εργασία ανέδειξε τη σημασία της διερεύνησης των ηθικών ζητημάτων που σχετίζονται με την Τεχνητή Νοημοσύνη και την ανάγκη για μια ολοκληρωμένη και διεπιστημονική προσέγγιση στην αντιμετώπιση των προκλήσεων που ανακύπτουν. Η υιοθέτηση ενός συνδυασμού καινοτόμων μεθοδολογικών προσεγγίσεων και η συνεργασία όλων των εμπλεκόμενων φορέων αποτελούν απαραίτητες προϋποθέσεις για την ανάπτυξη ενός ισχυρού ηθικού πλαισίου που θα διασφαλίζει την υπεύθυνη και ανθρωποκεντρική εξέλιξη της Τεχνητής Νοημοσύνης προς όφελος της κοινωνίας. Η έρευνα σε αυτό το πεδίο έχει τη δυνατότητα να διαμορφώσει ένα μέλλον στο οποίο η Τεχνητή Νοημοσύνη θα αποτελεί ένα ισχυρό εργαλείο για την αντιμετώπιση των μεγάλων προκλήσεων της εποχής μας, με γνώμονα πάντα τον σεβασμό των ανθρώπινων αξιών και δικαιωμάτων.

6

Βιβλιογρ

αφία

6.1 Έρευνα 2015-2019

- [1] H. Roberts, J. Cows, J. Morley, M. Taddeo, V. Wang, and L. Floridi, ‘The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation’, *AI Soc*, vol. 36, no. 1, pp. 59–77, Mar. 2021, doi: 10.1007/s00146-020-00992-2.
- [2] A. Barredo Arrieta *et al.*, ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/J.INFFUS.2019.12.012.
- [3] V. Bogina, A. Hartman, T. Kuflik, and A. Shulner-Tal, ‘Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics’, *Int J Artif Intell Educ*, vol. 32, no. 3, pp. 808–833, Sep. 2022, doi: 10.1007/s40593-021-00248-0.
- [4] J. M. Logg, J. A. Minson, and D. A. Moore, ‘Algorithm appreciation: People prefer algorithmic to human judgment’, *Organ Behav Hum Decis Process*, vol. 151, pp. 90–103, Mar. 2019, doi: 10.1016/J.OBHDP.2018.12.005.
- [5] M. Goddard, ‘The EU General Data Protection Regulation (GDPR): European regulation that has a global impact’, *International Journal of Market Research*, vol. 59, no. 6, pp. 703–705, 2017.
- [6] F. Rossi, ‘Building trust in artificial intelligence’, *J Int Aff*, vol. 72, no. 1, pp. 127–134, 2018.
- [7] M. Perez Garcia and S. Saffon Lopez, ‘Building trust between users and telecommunications data driven virtual assistants’, in *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 2018, pp. 628–637.
- [8] S. Venkatasubramanian, ‘Algorithmic Fairness’, in *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, New York, NY, USA: ACM, Jun. 2019, pp. 481–481. doi: 10.1145/3294052.3322192.
- [9] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, ‘Re-imagining Algorithmic Fairness in India and Beyond’, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Mar. 2021, pp. 315–328. doi: 10.1145/3442188.3445896.
- [10] A. Amirteimoori, T. Allahviranloo, M. Zadmirzaei, and F. Hasanzadeh, ‘On the environmental performance analysis: A combined fuzzy data envelopment analysis and artificial intelligence algorithms’, *Expert Syst Appl*, vol. 224, p. 119953, Aug. 2023, doi: 10.1016/j.eswa.2023.119953.

- [11] A. Asatiani, P. Malo, P. R. Nagbøl, E. Penttinen, T. Rinta-Kahila, and A. Salovaara, ‘Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems’, *J Assoc Inf Syst*, vol. 22, no. 2, pp. 325–352, 2021, doi: 10.17705/1jais.00664.
- [12] C. Held, J. Krumm, P. Markel, and R. P. Schenke, ‘Intelligent Video Surveillance’, *Computer (Long Beach Calif)*, vol. 45, no. 3, pp. 83–84, Mar. 2012, doi: 10.1109/MC.2012.97.
- [13] O. J. Erdélyi and J. Goldsmith, ‘Regulating Artificial Intelligence’, in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA: ACM, Dec. 2018, pp. 95–101. doi: 10.1145/3278721.3278731.
- [14] K. Siau and W. Wang, ‘Artificial Intelligence (AI) Ethics’, *Journal of Database Management*, vol. 31, no. 2, pp. 74–87, Apr. 2020, doi: 10.4018/JDM.2020040105.
- [15] P.-A. Balland, R. Boschma, J. Crespo, and D. L. Rigby, ‘Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification’, *Reg Stud*, vol. 53, no. 9, pp. 1252–1268, Sep. 2019, doi: 10.1080/00343404.2018.1437900.
- [16] T. A. Börzel, A. Dimitrova, and F. Schimmelfennig, ‘European Union enlargement and integration capacity: concepts, findings, and policy implications’, *J Eur Public Policy*, vol. 24, no. 2, pp. 157–176, Feb. 2017, doi: 10.1080/13501763.2016.1265576.
- [17] L. Floridi *et al.*, ‘AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’, *Minds Mach (Dordr)*, vol. 28, no. 4, pp. 689–707, Dec. 2018, doi: 10.1007/s11023-018-9482-5.
- [18] A. Adadi and M. Berrada, ‘Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)’, *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
- [19] B. P. Green, ‘Ethical Reflections on Artificial Intelligence’, *Scientia et Fides*, vol. 6, no. 2, p. 9, Oct. 2018, doi: 10.12775/SetF.2018.015.
- [20] A. F. T. Winfield and M. Jirotko, ‘Ethical governance is essential to building trust in robotics and artificial intelligence systems’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180085, Nov. 2018, doi: 10.1098/rsta.2018.0085.
- [21] J.-A. Santos, A. Fernández, M. Moreno-Rebato, H. Billhardt, J.-A. Rodríguez-García, and S. Ossowski, ‘Legal and ethical implications of applications based on agreement technologies: the case of auction-based road intersections’, *Artif Intell Law (Dordr)*, vol. 28, no. 4, pp. 385–414, Dec. 2020, doi: 10.1007/s10506-019-09259-8.
- [22] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, ‘Artificial Intelligence and the “Good Society”: the US, EU, and UK approach’, *Sci Eng Ethics*, Mar. 2017, doi: 10.1007/s11948-017-9901-7.
- [23] J. Singh, J. Cobbe, and C. Norval, ‘Decision Provenance: Harnessing Data Flow for Accountable Systems’, *IEEE Access*, vol. 7, pp. 6562–6574, 2019, doi: 10.1109/ACCESS.2018.2887201.
- [24] W. Samek and K.-R. Müller, ‘Towards Explainable Artificial Intelligence’, 2019, pp. 5–22. doi: 10.1007/978-3-030-28954-6_1.
- [25] C. Reed, ‘How should we regulate artificial intelligence?’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, p. 20170360, Sep. 2018, doi: 10.1098/rsta.2017.0360.

- [26] T. Kirchberger, ‘European Union Policy-Making on Robotics and Artificial Intelligence: Selected Issues’, *Croatian Yearbook of European Law and Policy*, vol. 13, no. 13, Dec. 2017, doi: 10.3935/cyelp.13.2017.272.
- [27] W. Hall and J. Pesenti, ‘Growing the artificial intelligence industry in the UK’, *Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. Part of the Industrial Strategy UK and the Commonwealth*, 2017.
- [28] D. Vrontis, M. Christofi, V. Pereira, S. Tarba, A. Makrides, and E. Trichina, ‘Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review’, *The International Journal of Human Resource Management*, vol. 33, no. 6, pp. 1237–1266, Mar. 2022, doi: 10.1080/09585192.2020.1871398.
- [29] M. Almada, ‘Contesting Automated Decisions: Limits to the Right to Human Intervention in Automated Decision-Making’, *SSRN Electronic Journal*, 2018, doi: 10.2139/ssrn.3264189.
- [30] K. Ishii, ‘Comparative legal study on privacy and personal data protection for robots equipped with artificial intelligence: looking at functional and technological aspects’, *AI Soc*, vol. 34, no. 3, pp. 509–533, Sep. 2019, doi: 10.1007/s00146-017-0758-8.
- [31] M. Butterworth, ‘The ICO and artificial intelligence: The role of fairness in the GDPR framework’, *Computer Law & Security Review*, vol. 34, no. 2, pp. 257–268, Apr. 2018, doi: 10.1016/J.CLSR.2018.01.004.
- [32] J. Berscheid and F. Roewer-Despres, ‘Beyond transparency’, *AI Matters*, vol. 5, no. 2, pp. 13–22, Aug. 2019, doi: 10.1145/3340470.3340476.
- [33] M. N. Asghar, N. Kanwal, B. Lee, M. Fleury, M. Herbst, and Y. Qiao, ‘Visual Surveillance Within the EU General Data Protection Regulation: A Technology Perspective’, *IEEE Access*, vol. 7, pp. 111709–111726, 2019, doi: 10.1109/ACCESS.2019.2934226.
- [34] M. Veale and L. Edwards, ‘Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling’, *Computer Law & Security Review*, vol. 34, no. 2, pp. 398–404, Apr. 2018, doi: 10.1016/J.CLSR.2017.12.002.
- [35] F. Thiel and J. Wetzlich, ‘The European Metrology Cloud: Impact of European Regulations on Data Protection and the Free Flow of Non-Personal Data’, in *19th International Congress of Metrology (CIM2019)*, S. Gazal, Ed., Les Ulis, France: EDP Sciences, Sep. 2019, p. 01001. doi: 10.1051/metrology/201901001.
- [36] J. Saltz *et al.*, ‘Integrating Ethics within Machine Learning Courses’, *ACM Transactions on Computing Education*, vol. 19, no. 4, pp. 1–26, Dec. 2019, doi: 10.1145/3341164.
- [37] M. Abrams, J. Abrams, P. Cullen, and L. Goldstein, ‘Artificial Intelligence, Ethics, and Enhanced Data Stewardship’, *IEEE Secur Priv*, vol. 17, no. 2, pp. 17–30, Mar. 2019, doi: 10.1109/MSEC.2018.2888778.
- [38] M. B. Hoeschl, T. C. D. Bueno, and H. C. Hoeschl, ‘Fourth Industrial Revolution and the future of Engineering: Could Robots Replace Human Jobs? How Ethical Recommendations can Help Engineers Rule on Artificial Intelligence’, in *2017 7th World Engineering Education Forum (WEEF)*, IEEE, Nov. 2017, pp. 21–26. doi: 10.1109/WEEF.2017.8466973.
- [39] B. Wagner, ‘Ethics As An Escape From Regulation. From “Ethics-Washing” To Ethics-Shopping?’, in *BEING PROFILED*, Amsterdam University Press, 2019, pp. 84–89. doi: 10.1515/9789048550180-016.

- [40] K. Y. Lee, H. Y. Kwon, and J. I. Lim, ‘Legal Consideration on the Use of Artificial Intelligence Technology and Self-regulation in Financial Sector: Focused on Robo-Advisors’, 2018, pp. 323–335. doi: 10.1007/978-3-319-93563-8_27.
- [41] A. Jobin, M. Ienca, and E. Vayena, ‘The global landscape of AI ethics guidelines’, *Nat Mach Intell*, vol. 1, no. 9, pp. 389–399, 2019.
- [42] V. Dignum, *Responsible artificial intelligence: how to develop and use AI in a responsible way*, vol. 2156. Springer, 2019.
- [43] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, ‘The ethics of algorithms: Mapping the debate’, *Big Data Soc*, vol. 3, no. 2, Dec. 2016, doi: 10.1177/2053951716679679.
- [44] M. A. Boden, *AI: Its Nature and Future*. Oxford University Press, 2016. [Online]. Available: <https://books.google.gr/books?id=yDQTDAAAQBAJ>
- [45] M. Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*. in Pelican Books. Penguin Books Limited, 2019. [Online]. Available: <https://books.google.gr/books?id=VfSIDwAAQBAJ>
- [46] J. McCarthy, M. Minsky, N. Rochester, and C. Shannon, ‘A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955.’, *AI Mag*, vol. 27, pp. 12–14, Jan. 2006.
- [47] N. J. Nilsson, *The quest for artificial intelligence*. Cambridge University Press, 2009.
- [48] M. I. Jordan and T. M. Mitchell, ‘Machine learning: Trends, perspectives, and prospects’, *Science (1979)*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.
- [49] A. M. TURING, ‘I.—COMPUTING MACHINERY AND INTELLIGENCE’, *Mind*, vol. LIX, no. 236, pp. 433–460, Oct. 1950, doi: 10.1093/mind/LIX.236.433.
- [50] Y. LeCun, Y. Bengio, and G. Hinton, ‘Deep learning’, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [51] J. R. Searle, ‘Minds, brains, and programs’, *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–424, Sep. 1980, doi: 10.1017/S0140525X00005756.
- [52] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. [Online]. Available: https://books.google.gr/books?id=7_H8AwAAQBAJ
- [53] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [54] S. Poland, S. Sharad, and G. Wigen-Toccalino, “Powering Innovation: The Comprehensive Impact of Generative AI,” *Climate and Energy*, vol. 40, no. 7, pp. 1–11, Feb. 2024, doi: 10.1002/gas.22385.
- [55] R. K. Gatla, A. Gatla, P. Sridhar, D. G. Kumar, and D. S. N. M. Rao, “Advancements in Generative AI: Exploring Fundamentals and Evolution,” in 2024 International Conference on Electronics, Computing, Communication and Control Technology (ICECCC), IEEE, May 2024, pp. 1–5. doi: 10.1109/ICECCC61767.2024.10594003.
- [56] R. Rathod, “Renaissance on Generative AI,” *Int J Res Appl Sci Eng Technol*, vol. 12, no. 6, pp. 1354–1359, Jun. 2024, doi: 10.22214/ijraset.2024.63324.
- [57] M. B. Garcia, “The Paradox of Artificial Creativity: Challenges and Opportunities of Generative AI Artistry,” *Creat Res J*, pp. 1–14, May 2024, doi: 10.1080/10400419.2024.2354622.
- [58] G. Minni, S. Nagulmeera, B. B. Lakshmi, and N. S. Shaik, “Generative AI: Exploring the Applications of Generative Models in Creative Industries,” *International Journal of Emerging Research in Engineering, Science, and Management*, vol. 3, no. 3, 2024, doi: 10.58482/ijeresm.v3i3.5.

- [59] L. Floridi and J. Cowls, ‘A Unified Framework of Five Principles for AI in Society’, *Harv Data Sci Rev*, Jun. 2019, doi: 10.1162/99608f92.8cd550d1.
- [60] S. Barocas and A. D. Selbst, ‘Big Data’s Disparate Impact’, *SSRN Electronic Journal*, 2016, doi: 10.2139/ssrn.2477899.
- [61] F. Doshi-Velez and B. Kim, ‘Towards a rigorous science of interpretable machine learning’, *arXiv preprint arXiv:1702.08608*, 2017.
- [62] M. Taddeo and L. Floridi, ‘How AI can be a force for good’, *Science (1979)*, vol. 361, no. 6404, pp. 751–752, Aug. 2018, doi: 10.1126/science.aat5991.
- [63] C. Rudin, ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, *Nat Mach Intell*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
- [64] K. Crawford and R. Calo, ‘There is a blind spot in AI research’, *Nature*, vol. 538, no. 7625, pp. 311–313, Oct. 2016, doi: 10.1038/538311a.
- [65] E. Awad *et al.*, ‘The Moral Machine experiment’, *Nature*, vol. 563, no. 7729, pp. 59–64, Nov. 2018, doi: 10.1038/s41586-018-0637-6.
- [66] R. Vinuesa *et al.*, ‘The role of artificial intelligence in achieving the Sustainable Development Goals’, *Nat Commun*, vol. 11, no. 1, p. 233, Jan. 2020, doi: 10.1038/s41467-019-14108-y.
- [67] C. Cath, ‘Governing artificial intelligence: ethical, legal and technical opportunities and challenges’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180080, Nov. 2018, doi: 10.1098/rsta.2018.0080.
- [68] G. Buttarelli, ‘Privacy matters: updating human rights for the digital society’, *Health Technol (Berl)*, vol. 7, no. 4, pp. 325–328, 2017.
- [69] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation (GDPR)*. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-57959-7.
- [70] S. Wachter, B. Mittelstadt, and L. Floridi, ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’, *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, May 2017, doi: 10.1093/idpl/ix005.
- [71] C. Burton, L. De Boel, C. Kuner, A. Pateraki, S. Cadiot, and S. G. Hoffman, ‘The final european union general data protection regulation’, *BNA Privacy & Security Law Report*, vol. 15, p. 153, 2016.
- [72] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, ‘EU General Data Protection Regulation: Changes and implications for personal data collecting companies’, *Computer Law & Security Review*, vol. 34, no. 1, pp. 134–153, Feb. 2018, doi: 10.1016/J.CLSR.2017.05.015.
- [73] L. Edwards and M. Veale, ‘Slave to the algorithm? Why a ‘right to an explanation’ is probably not the remedy you are looking for’, *Duke L. & Tech. Rev.*, vol. 16, p. 18, 2017.
- [74] B. Malle, P. Kieseberg, E. Weippl, and A. Holzinger, ‘The right to be forgotten: Towards Machine Learning on perturbed knowledge bases’, in *Springer Lecture Notes in Computer Science LNCS 9817*, Springer International, Sep. 2016, pp. 251–266.
- [75] C. Tankard, ‘What the GDPR means for businesses’, *Network Security*, vol. 2016, no. 6, pp. 5–8, Jun. 2016, doi: 10.1016/S1353-4858(16)30056-3.

- [76] P. De Hert and V. Papakonstantinou, ‘The new General Data Protection Regulation: Still a sound system for the protection of individuals?’, *Computer Law & Security Review*, vol. 32, no. 2, pp. 179–194, Apr. 2016, doi: 10.1016/J.CLSR.2016.02.006.
- [77] J. P. Albrecht, ‘How the GDPR will change the world’, *Eur. Data Prot. L. Rev.*, vol. 2, p. 287, 2016.
- [78] C. J. Hoofnagle, B. van der Sloot, and F. Z. Borgesius, ‘The European Union general data protection regulation: what it is and what it means’, *Information & Communications Technology Law*, vol. 28, no. 1, pp. 65–98, Jan. 2019, doi: 10.1080/13600834.2019.1573501.
- [79] European Commission, ‘Proposal for a Regulation laying down harmonised rules on artificial intelligence’, Brussels, Apr. 2021. Accessed: Oct. 30, 2024. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>
- [80] H. AI, ‘High-level expert group on artificial intelligence’, *Ethics guidelines for trustworthy AI*, vol. 6, 2019.
- [81] European Commission, ‘White Paper on Artificial Intelligence - A European approach to excellence and trust’, Brussels, Feb. 2020. Accessed: Oct. 30, 2024. [Online]. Available: https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_en?filename=commission-white-paper-artificial-intelligence-feb2020_en.pdf
- [82] M. Veale and F. J. Z. Borgesius, ‘Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach’, *Computer Law Review International*, vol. 22, pp. 97–112, 2021, [Online]. Available: <https://api.semanticscholar.org/CorpusID:235765823>
- [83] N. A. Smuha, ‘From a “Race to AI” to a “Race to AI Regulation” - Regulatory Competition for Artificial Intelligence’, *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3501410.
- [84] L. Floridi, ‘Establishing the rules for building trustworthy AI’, *Ethics, Governance, and Policies in Artificial Intelligence*, pp. 41–45, 2021.
- [85] M. Ebers, ‘Regulating AI and Robotics: Ethical and Legal Challenges’, in *Algorithms and Law*, M. Ebers and S. Navas, Eds., Cambridge: Cambridge University Press, 2020, pp. 37–99. doi: DOI: 10.1017/9781108347846.003.
- [86] P. Cihon, ‘Standards for AI governance: international standards to enable global coordination in AI research & development’, *Future of Humanity Institute. University of Oxford*, vol. 40, no. 3, pp. 340–342, 2019.
- [87] D. Leslie, ‘Understanding artificial intelligence ethics and safety’, *arXiv preprint arXiv:1906.05684*, 2019.
- [88] ISO/IEC JTC 1/SC 42, ‘Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (ISO/IEC 22989:2022)’, Geneva, Switzerland, 2022. Accessed: Oct. 30, 2024. [Online]. Available: <https://www.iso.org/standard/74296.html>
- [89] ISO/IEC JTC 1/SC 42, ‘Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) (ISO/IEC 23053:2022)’, Geneva, Switzerland, 2022. Accessed: Oct. 30, 2024. [Online]. Available: <https://www.iso.org/standard/74438.html>
- [90] ISO/IEC JTC 1/SC 42, ‘Information technology — Artificial intelligence — Management system (ISO/IEC 42001:2023)’, Geneva, Switzerland, 2023. Accessed: Oct. 30, 2024. [Online]. Available: <https://www.iso.org/standard/81230.html?browse=tc>
- [91] G. Adamson, J. C. Havens, and R. Chatila, ‘Designing a value-driven future for ethical autonomous and intelligent systems’, *Proceedings of the IEEE*, vol. 107, no. 3, pp. 518–525, 2019.

- [92] D. Schiff, J. Borenstein, J. Biddle, and K. Laas, 'AI ethics in the public, private, and NGO sectors: A review of a global document collection', *IEEE Transactions on Technology and Society*, vol. 2, no. 1, pp. 31–42, 2021.
- [93] I. D. Raji *et al.*, 'Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing', in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44.
- [94] B. Mittelstadt, 'Principles alone cannot guarantee ethical AI', *Nat Mach Intell*, vol. 1, no. 11, pp. 501–507, Nov. 2019, doi: 10.1038/s42256-019-0114-4.
- [95] S. Larsson and F. Heintz, 'Transparency in artificial intelligence', *Internet Policy Review*, vol. 9, no. 2, May 2020, doi: 10.14763/2020.2.1469.
- [96] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI', *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3518482.
- [97] E. J. Topol, 'High-performance medicine: the convergence of human and artificial intelligence', *Nat Med*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.
- [98] K.-H. Yu, A. L. Beam, and I. S. Kohane, 'Artificial intelligence in healthcare', *Nat Biomed Eng*, vol. 2, no. 10, pp. 719–731, Oct. 2018, doi: 10.1038/s41551-018-0305-z.
- [99] F. S. B. F. S. Board, *Artificial intelligence and machine learning in financial services*. Financial Stability Board, 2017. Accessed: Oct. 30, 2024. [Online]. Available: <https://www.fsb.org/uploads/P011117.pdf>
- [100] B. G. Buchanan, 'Artificial intelligence in finance', 2019.
- [101] B. W. Wirtz, J. C. Weyerer, and C. Geyer, 'Artificial Intelligence and the Public Sector—Applications and Challenges', *International Journal of Public Administration*, vol. 42, no. 7, pp. 596–615, May 2019, doi: 10.1080/01900692.2018.1498103.
- [102] A. Babuta and M. Oswald, 'Data analytics and algorithmic bias in policing', 2019.
- [103] J. Lynch, 'Face off: Law enforcement use of face recognition technology', *Available at SSRN 3909038*, 2020.
- [104] M. Hino, E. Benami, and N. Brooks, 'Machine learning for environmental monitoring', *Nat Sustain*, vol. 1, no. 10, pp. 583–588, Oct. 2018, doi: 10.1038/s41893-018-0142-9.
- [105] D. Rolnick *et al.*, 'Tackling climate change with machine learning', *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–96, 2022.
- [106] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.
- [107] E. Brynjolfsson and A. McAfee, 'The Business of Artificial Intelligence', *Harv Bus Rev*, Jul. 2017, Accessed: Oct. 15, 2024. [Online]. Available: <https://hbr.org/2017/07/the-business-of-artificial-intelligence>
- [108] C. B. Frey and M. A. Osborne, 'The future of employment: How susceptible are jobs to computerisation?', *Technol Forecast Soc Change*, vol. 114, pp. 254–280, Jan. 2017, doi: 10.1016/J.TECHFORE.2016.08.019.
- [109] S. Zuboff, 'The age of surveillance capitalism', in *Social theory re-wired*, Routledge, 2023, pp. 203–213.
- [110] D. J. Solove, 'The myth of the privacy paradox', *Geo. Wash. L. Rev.*, vol. 89, p. 1, 2021.

- [111] D. Helbing *et al.*, ‘Will Democracy Survive Big Data and Artificial Intelligence?’, in *Towards Digital Enlightenment*, Cham: Springer International Publishing, 2019, pp. 73–98. doi: 10.1007/978-3-319-90869-4_7.
- [112] M. Janssen and G. Kuk, ‘The challenges and limits of big data algorithms in technocratic governance’, *Gov Inf Q*, vol. 33, no. 3, pp. 371–377, Jul. 2016, doi: 10.1016/J.GIQ.2016.08.011.
- [113] C. O’neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.
- [114] L. Robinson *et al.*, ‘Digital inequalities and why they matter’, *Inf Commun Soc*, vol. 18, no. 5, pp. 569–582, May 2015, doi: 10.1080/1369118X.2015.1012532.
- [115] J. Van Dijk, *The digital divide*. John Wiley & Sons, 2020.
- [116] N. Bostrom and E. Yudkowsky, ‘The ethics of artificial intelligence’, in *Artificial intelligence safety and security*, Chapman and Hall/CRC, 2018, pp. 57–69.
- [117] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, ‘A Survey on Bias and Fairness in Machine Learning’, *ACM Comput Surv*, vol. 54, no. 6, pp. 1–35, Jul. 2022, doi: 10.1145/3457607.
- [118] K. Morovat and B. Panda, ‘A Survey of Artificial Intelligence in Cybersecurity’, in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, Dec. 2020, pp. 109–115. doi: 10.1109/CSCI51800.2020.00026.
- [119] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, ‘Concrete problems in AI safety’, *arXiv preprint arXiv:1606.06565*, 2016.
- [120] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, ‘Towards the science of security and privacy in machine learning’, *arXiv preprint arXiv:1611.03814*, 2016.
- [121] R. V Yampolskiy, ‘Unpredictability of AI: On the impossibility of accurately predicting all actions of a smarter agent’, *Journal of Artificial Intelligence and Consciousness*, vol. 7, no. 01, pp. 109–118, 2020.
- [122] S. Cave and S. S. ÓhÉigeartaigh, ‘Bridging near- and long-term concerns about AI’, *Nat Mach Intell*, vol. 1, no. 1, pp. 5–6, Jan. 2019, doi: 10.1038/s42256-018-0003-2.
- [123] J. Pfeiffer *et al.*, ‘Algorithmic Fairness in AI’, *Business & Information Systems Engineering*, vol. 65, no. 2, pp. 209–222, Apr. 2023, doi: 10.1007/s12599-023-00787-x.
- [124] J. He, Y. Wan, L. Feng, J. Ai, and Y. Wang, ‘An integrated data envelopment analysis and emergy-based ecological footprint methodology in evaluating sustainable development, a case study of Jiangsu Province, China’, *Ecol Indic*, vol. 70, pp. 23–34, Nov. 2016, doi: 10.1016/J.ECOLIND.2016.05.042.
- [125] G. Sreenu and M. A. Saleem Durai, ‘Intelligent video surveillance: a review through deep learning techniques for crowd analysis’, *J Big Data*, vol. 6, no. 1, p. 48, Dec. 2019, doi: 10.1186/s40537-019-0212-5.
- [126] A. Ben Mabrouk and E. Zagrouba, ‘Abnormal behavior recognition for intelligent video surveillance systems: A review’, *Expert Syst Appl*, vol. 91, pp. 480–491, Jan. 2018, doi: 10.1016/j.eswa.2017.09.029.
- [127] O. Stasse and T. Flayols, ‘An Overview of Humanoid Robots Technologies’, 2019, pp. 281–310. doi: 10.1007/978-3-319-93870-7_13.
- [128] G. E. Marchant and B. Allenby, ‘Soft law: New tools for governing emerging technologies’, *Bulletin of the Atomic Scientists*, vol. 73, no. 2, pp. 108–114, Mar. 2017, doi: 10.1080/00963402.2017.1288447.
- [129] I. Bartoletti, ‘AI in Healthcare: Ethical and Privacy Challenges’, 2019, pp. 7–10. doi: 10.1007/978-3-030-21642-9_2.

- [130] T. Ilkka, *The impact of artificial intelligence on learning, teaching, and education*. European Union, 2018.
- [131] C. Katzenbach and L. Ulbricht, ‘Algorithmic governance’, *Internet Policy Review*, vol. 8, no. 4, Nov. 2019, doi: 10.14763/2019.4.1424.
- [132] C. Fiesler, A. Bruckman, R. E. Kraut, M. Muller, C. Munteanu, and K. Shilton, ‘Research Ethics and Regulation’, in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, New York, NY, USA: ACM, Oct. 2018, pp. 125–128. doi: 10.1145/3272973.3274543.
- [133] F. Muhlenbach and I. Sayn, ‘Artificial Intelligence and Law’, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, New York, NY, USA: ACM, Jun. 2019, pp. 224–228. doi: 10.1145/3322640.3326722.
- [134] A. Lauterbach, ‘Artificial intelligence and policy: quo vadis?’, *Digital Policy, Regulation and Governance*, vol. 21, no. 3, pp. 238–263, Jan. 2019, doi: 10.1108/DPRG-09-2018-0054.
- [135] S. Lumbreras, ‘Getting Ready for the Next Step: Merging Information Ethics and Roboethics—A Project in the Context of Marketing Ethics’, *Information*, vol. 9, no. 8, p. 195, Aug. 2018, doi: 10.3390/info9080195.
- [136] D. Susar and V. Aquaro, ‘Artificial Intelligence’, in *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance*, New York, NY, USA: ACM, Apr. 2019, pp. 418–426. doi: 10.1145/3326365.3326420.
- [137] P. Timmers, ‘Ethics of AI and Cybersecurity When Sovereignty is at Stake’, *Minds Mach (Dordr)*, vol. 29, no. 4, pp. 635–645, Dec. 2019, doi: 10.1007/s11023-019-09508-4.
- [138] R. S. S. Kumar, A. Wicker, and M. Swann, ‘Practical Machine Learning for Cloud Intrusion Detection’, in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, New York, NY, USA: ACM, Nov. 2017, pp. 81–90. doi: 10.1145/3128572.3140445.
- [139] A. Kaplan and M. Haenlein, ‘Rulers of the world, unite! The challenges and opportunities of artificial intelligence’, *Bus Horiz*, vol. 63, no. 1, pp. 37–50, Jan. 2020, doi: 10.1016/J.BUSHOR.2019.09.003.
- [140] J. M. Ptaschunder, ‘The Legal and International Situation of AI, Robotics and Big Data With Attention to Healthcare’, *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3472885.
- [141] G. Malgieri, ‘Automated decision-making in the EU Member States: The right to explanation and other “suitable safeguards” in the national legislations’, *Computer law & security review*, vol. 35, no. 5, p. 105327, 2019.
- [142] M. C. BUITEN, ‘Towards Intelligent Regulation of Artificial Intelligence’, *European Journal of Risk Regulation*, vol. 10, no. 1, pp. 41–59, 2019, doi: 10.1017/err.2019.8.
- [143] S. Basu, A. Omotubora, M. Beeson, and C. Fox, ‘Legal framework for small autonomous agricultural robots’, *AI Soc*, vol. 35, no. 1, pp. 113–134, Mar. 2020, doi: 10.1007/s00146-018-0846-4.
- [144] Y. Wang, M. Xiong, and H. Olya, ‘Toward an Understanding of Responsible Artificial Intelligence Practices’, 2020. doi: 10.24251/HICSS.2020.610.
- [145] C. Reed, ‘How should we regulate artificial intelligence?’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, p. 20170360, Sep. 2018, doi: 10.1098/rsta.2017.0360.
- [146] M. Brkan, ‘AI-supported decision-making under the general data protection regulation’, in *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*, New York, NY, USA: ACM, Jun. 2017, pp. 3–8. doi: 10.1145/3086512.3086513.

- [147] P. Guarda and others, ““ Ok Google, am I sick?”: artificial intelligence, e-health, and data protection regulation’, *BioLaw Journal*, vol. 2019, pp. 359–375, 2019.
- [148] K. H. Hessekiel, E. Kim, J. E. Tierney, J. Y. Yang, and C. T. Bavitz, ‘AGTech Forum Briefing Book: State Attorneys General and Artificial Intelligence’, 2018.
- [149] R. Clarke, ‘Guidelines for the Responsible Business Use of AI Foundational Working Paper’, 2018.
- [150] F. De Rouck, ‘Moral rights & AI environments: the unique bond between intelligent agents and their creations’, *Journal of Intellectual Property Law & Practice*, vol. 14, no. 4, pp. 299–304, Apr. 2019, doi: 10.1093/jiplp/jpz010.
- [151] M. Coeckelbergh, ‘Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability’, *Sci Eng Ethics*, vol. 26, no. 4, pp. 2051–2068, Aug. 2020, doi: 10.1007/s11948-019-00146-8.
- [152] O. Pollicino and G. De Gregorio, ‘A Constitutional-Driven Change of Heart: ISP Liability and Artificial Intelligence in the Digital Single Market’, *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3418995.
- [153] R. J. Howley III, *The Effects of Artificial Intelligence on the Youth*. Utica College, 2019.
- [154] S. Aaronson, ‘DATA MINEFIELD?’, Nov. 2020.
- [155] F. Merz, ‘Europe and the Global AI Race’, Zurich, 2019. doi: 10.3929/ethz-b-000345968.
- [156] E. Magrani, ‘New perspectives on ethics and the laws of artificial intelligence’, *Internet Policy Review*, vol. 8, no. 3, Sep. 2019, doi: 10.14763/2019.3.1420.
- [157] C. Kuner, F. H. Cate, O. Lynskey, C. Millard, N. Ni Loideain, and D. J. B. Svantesson, ‘Expanding the artificial intelligence-data protection debate’, *International Data Privacy Law*, vol. 8, no. 4, pp. 289–292, Nov. 2018, doi: 10.1093/idpl/ipy024.
- [158] S. Li, L. Da Xu, and S. Zhao, ‘The internet of things: a survey’, *Information systems frontiers*, vol. 17, pp. 243–259, 2015.
- [159] N. Poritskiy, F. Oliveira, and F. Almeida, ‘The benefits and challenges of general data protection regulation for the information technology sector’, *Digital Policy, Regulation and Governance*, vol. 21, no. 5, pp. 510–524, Jan. 2019, doi: 10.1108/DPRG-05-2019-0039.
- [160] M. V. d’Alfonso and D. Berkich, ‘On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence’, *Philosophical Studies Series*, 2019, [Online]. Available: <https://api.semanticscholar.org/CorpusID:59307487>
- [161] C. Schaefer and C. Edman, ‘Transparent Logging with Hyperledger Fabric’, in *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, IEEE, May 2019, pp. 65–69. doi: 10.1109/BLOC.2019.8751339.
- [162] T. C. D. Bueno, H. C. Hoeschl, and M. B. Hoeschl, ‘Developing an Ontology-oriented Framework to Enable Peace-building Solutions Based on Fairplay Algorithms’, in *2018 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)*, IEEE, Nov. 2018, pp. 1–6. doi: 10.1109/WEEF-GEDC.2018.8629678.
- [163] G. Fellows, ‘Discord and Disruption - 2019 Global Trends Report’, Waterloo, 2019. Accessed: Oct. 22, 2024. [Online]. Available: <https://balsillieschool.ca/wp-content/uploads/2021/08/Graduate-Fellows-Anthology-2019.pdf>

- [164] P. R. Daugherty, ‘Collaborative Intelligence: Humans and AI Are Joining Forces’, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:53127445>
- [165] A. Mantelero, ‘AI and Big Data: A blueprint for a human rights, social and ethical impact assessment’, *Computer Law & Security Review*, vol. 34, no. 4, pp. 754–772, Aug. 2018, doi: 10.1016/J.CLSR.2018.05.017.
- [166] C. Zednik, ‘Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence’, *Philos Technol*, vol. 34, no. 2, pp. 265–288, Jun. 2021, doi: 10.1007/s13347-019-00382-7.
- [167] S. Wachter, B. Mittelstadt, and L. Floridi, ‘Transparent, explainable, and accountable AI for robotics’, *Sci Robot*, vol. 2, no. 6, May 2017, doi: 10.1126/scirobotics.aan6080.
- [168] G. Contissa *et al.*, ‘Claudette Meets GDPR: Automating the Evaluation of Privacy Policies Using Artificial Intelligence’, *SSRN Electronic Journal*, 2018, doi: 10.2139/ssrn.3208596.
- [169] K. Yeung, A. Howes, and G. Pogrebna, ‘AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing’, *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3435011.
- [170] M. Canellopoulou-Bottis, F. Panagopoulou, A. Michailaki, and M. Nikita, ‘The Right to Human Intervention: Law, Ethics and Artificial Intelligence’, *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3430075.
- [171] P. Aarvik, ‘Artificial Intelligence—a promising anti-corruption tool in development settings’, *U4 Anti-Corruption Resource Centre*, 2019.
- [172] S. Robbins, ‘AI and the path to envelopment: knowledge as a first step towards the responsible regulation and use of AI-powered machines’, *AI Soc*, vol. 35, no. 2, pp. 391–400, Jun. 2020, doi: 10.1007/s00146-019-00891-1.
- [173] E. Fosch Villaronga, ‘“I Love You,” Said the Robot: Boundaries of the Use of Emotions in Human-Robot Interactions’, 2019, pp. 93–110. doi: 10.1007/978-3-319-96722-6_6.
- [174] J. A. Kroll, ‘Data Science Data Governance’, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:203601093>
- [175] D. Castro and M. McLaughlin, ‘Ten Ways the Precautionary Principle Undermines Progress in Artificial Intelligence’, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:159261424>
- [176] P. Treleaven, J. Barnett, and A. Koshiyama, ‘Algorithms: Law and Regulation’, *Computer (Long Beach Calif)*, vol. 52, no. 2, pp. 32–40, Feb. 2019, doi: 10.1109/MC.2018.2888774.
- [177] H. Kuwajima and F. Ishikawa, ‘Adapting square for quality assessment of artificial intelligence systems’, in *2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, 2019, pp. 13–18.
- [178] E. Thelisson, K. Sharma, H. Salam, and V. Dignum, ‘The General Data Protection Regulation’, in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, Apr. 2018, pp. 1–8. doi: 10.1145/3170427.3170632.
- [179] G. Contissa *et al.*, ‘Towards Consumer-Empowering Artificial Intelligence’, in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, pp. 5150–5157. doi: 10.24963/ijcai.2018/714.

6.2 Έρευνα 2024

- [180] I. O. Gallegos *et al.*, “Bias and fairness in large language models: A survey,” *Computational Linguistics*, pp. 1–79, 2024.
- [181] D. Singh, K. I. P. Kumar, G. Nijhawan, C. VEENA, B. Rajalakshmi, and B. T. Geetha, “Ensuring Transparency and Fairness in AI DecisionMaking Processes Influenced by large language Models,” in *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, IEEE, Apr. 2024, pp. 987–994. doi: 10.1109/CSNT60213.2024.10545998.
- [182] R. Goldshmidt and M. Horovicz, “TokenSHAP: Interpreting Large Language Models with Monte Carlo Shapley Value Estimation,” *arXiv preprint arXiv:2407.10114*, 2024.
- [183] H. Zheng and U. Pamuksuz, “SCENE: Evaluating Explainable AI Techniques Using Soft Counterfactuals,” *arXiv preprint arXiv:2408.04575*, 2024.
- [184] N. M Rao, “Solving Real-Time Information Updates and Mitigating Bias in Generative AI Models,” *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 2191–2205, Aug. 2024, doi: 10.38124/ijisrt/IJISRT24JUL1248.
- [185] A. Singh, “Diverse Yet Biased: Towards Mitigating Biases in Generative AI (Student Abstract),” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23653–23654, Mar. 2024, doi: 10.1609/aaai.v38i21.30512.
- [186] J. P. Lalor, A. Abbasi, K. Oketch, Y. Yang, and N. Forsgren, “Should Fairness be a Metric or a Model? A Model-based Framework for Assessing Bias in Machine Learning Pipelines,” *ACM Trans Inf Syst*, vol. 42, no. 4, pp. 1–41, Jul. 2024, doi: 10.1145/3641276.
- [187] L. Luo, Y. Nakao, M. Chollet, H. Inakoshi, and S. Stumpf, “EARN Fairness: Explaining, Asking, Reviewing and Negotiating Artificial Intelligence Fairness Metrics Among Stakeholders,” *arXiv preprint arXiv:2407.11442*, 2024.
- [188] L. Iurada, S. Bucci, T. M. Hospedales, and T. Tommasi, “Fairness Meets Cross-Domain Learning: A Benchmark of Models and Metrics,” *IEEE Access*, vol. 12, pp. 47854–47867, 2024, doi: 10.1109/ACCESS.2024.3383841.
- [189] A. J. D. Mahamadou, L. Goetz, and R. Altman, “Individual Fairness Through Reweighting and Tuning,” *arXiv preprint arXiv:2405.01711*, 2024.
- [190] S. W. S. Fischer, “Questioning AI: Promoting Decision-Making Autonomy Through Reflection,” *arXiv preprint arXiv:2409.10250*, 2024.
- [191] W. Woods, A. Grushin, S. Khan, and A. Velasquez, “Combining AI control systems and human decision support via robustness and criticality,” in *Disruptive Technologies in Information Sciences VIII*, B. T. Wysocki, M. Blowers, and R. Bharadwaj, Eds., SPIE, Jun. 2024, p. 13. doi: 10.1117/12.3016311.
- [192] A. Fuchs, A. Passarella, and M. Conti, “Optimizing Delegation in Collaborative Human-AI Hybrid Teams,” *ACM Transactions on Autonomous and Adaptive Systems*, vol. 19, no. 4, pp. 1–33, Dec. 2024, doi: 10.1145/3687130.
- [193] E. Y. Chang, “Integrating Emotional and Linguistic Models for Ethical Compliance in Large Language Models,” *arXiv preprint arXiv:2405.07076*, 2024.

- [194] S. Morales, R. Clarisó, and J. Cabot, “A DSL for Testing LLMs for Fairness and Bias,” in *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, New York, NY, USA: ACM, Sep. 2024, pp. 203–213. doi: 10.1145/3640310.3674093.
- [195] A. Zou *et al.*, “Improving Alignment and Robustness with Short Circuiting,” *arXiv preprint arXiv:2406.04313*, 2024.
- [196] I. Domkundwar and I. Bhola, “Safeguarding AI Agents: Developing and Analyzing Safety Architectures,” *arXiv preprint arXiv:2409.03793*, 2024.
- [197] M. Bahrami, R. Sonoda, and R. Srinivasan, “LLM Diagnostic Toolkit: Evaluating LLMs for Ethical Issues,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jun. 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10650995.
- [198] A. W. Lo and J. Ross, “Can ChatGPT Plan Your Retirement?: Generative AI and Financial Advice,” *Harv Data Sci Rev*, no. Special Issue 5, May 2024, doi: 10.1162/99608f92.ec74a002.
- [199] Oluwabusayo Adijat Bello and Komolafe Olufemi, “Artificial intelligence in fraud prevention: Exploring techniques and applications challenges and opportunities,” *Computer Science & IT Research Journal*, vol. 5, no. 6, pp. 1505–1520, Jun. 2024, doi: 10.51594/csitrj.v5i6.1252.
- [200] Beatrice Oyinkansola Adelakun, Ebere Ruth Onwubuariri, Gbenga Adeniyi Adeniran, and Afari Ntiakoh, “Enhancing fraud detection in accounting through AI: Techniques and case studies,” *Finance & Accounting Research Journal*, vol. 6, no. 6, pp. 978–999, Jun. 2024, doi: 10.51594/farj.v6i6.1232.
- [201] N. El-Haber *et al.*, “A Lifecycle Approach for Artificial Intelligence Ethics in Energy Systems,” *Energies (Basel)*, vol. 17, no. 14, p. 3572, Jul. 2024, doi: 10.3390/en17143572.
- [202] W. V. Siricharoen, “Elevating User-Centered Design with AI: A Comprehensive Exploration using the AI-UCD Algorithm Framework,” *EAI Endorsed Transactions on Context-aware Systems and Applications*, vol. 10, Mar. 2024, doi: 10.4108/eetcasa.4211.
- [203] G. Bas, C. Salinas, R. Tinoco, and J. Sevilla, “The EU AI Act: A pioneering effort to regulate frontier AI?,” *Inteligencia Artificial*, vol. 27, no. 73, pp. 55–64, Jan. 2024, doi: 10.4114/intartif.vol27iss73pp55-64.
- [204] J. Schuett, M. Anderljung, A. Carlier, L. Koessler, and B. Garfinkel, “From principles to rules: A regulatory approach for frontier AI,” *arXiv preprint arXiv:2407.07300*, 2024.
- [205] E. McKernon, G. Glasser, D. Cheng, and G. Hadfield, “AI Model Registries: A Foundational Tool for AI Governance,” *arXiv preprint arXiv:2410.09645*, 2024.
- [206] D. Carpenter and C. Ezell, “An FDA for AI? Pitfalls and Plausibility of Approval Regulation for Frontier Artificial Intelligence,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2024, pp. 239–254.
- [207] C. Salvador, “Certified Safe: A Schematic for Approval Regulation of Frontier AI,” *arXiv preprint arXiv:2408.06210*, 2024.
- [208] M. V. Krishnamoorthy, “Enhancing Responsible AGI Development: Integrating Human-in-the-loop Approaches with Blockchain-based Smart Contracts,” *Journal of Advances in Mathematics and Computer Science*, vol. 39, no. 9, pp. 14–39, Sep. 2024, doi: 10.9734/jamcs/2024/v39i91924.
- [209] L. D. Pöhler, K. Diepold, and W. Wallach, “A Practical Multilevel Governance Framework for Autonomous and Intelligent Systems,” *arXiv preprint arXiv:2404.13719*, 2024.

- [210] B. Judge, M. Nitzberg, and S. Russell, “When code isn’t law: rethinking regulation for artificial intelligence,” *Policy Soc*, May 2024, doi: 10.1093/polsoc/puae020.
- [211] Y. Qu, M. Ding, N. Sun, K. Thilakarathna, T. Zhu, and D. Niyato, “The Frontier of Data Erasure: Machine Unlearning for Large Language Models,” *arXiv preprint arXiv:2403.15779*, 2024.
- [212] L. Zhui, L. Fenghe, W. Xuehu, F. Qining, and R. Wei, “Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint,” *J Med Internet Res*, vol. 26, p. e60083, Aug. 2024, doi: 10.2196/60083.
- [213] J. Jiao, S. Afroogh, Y. Xu, and C. Phillips, “Navigating llm ethics: Advancements, challenges, and future directions,” *arXiv preprint arXiv:2406.18841*, 2024.
- [214] M. M. Ferdaus, M. Abdelguerfi, E. Ioup, K. N. Niles, K. Pathak, and S. Sloan, “Towards trustworthy ai: A review of ethical and robust large language models,” *arXiv preprint arXiv:2407.13934*, 2024.
- [215] N. I. Qureshi, A. Garg, P. Singh, and N. Retzlaff, “AI and Corporate Risk Management: Identifying and Mitigating Technological and Ethical Risks,” in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, IEEE, Apr. 2024, pp. 1–5. doi: 10.1109/ICKECS61492.2024.10617141.
- [216] A. Srivastava and S. Panda, “A Formal Framework for Assessing and Mitigating Emergent Security Risks in Generative AI Models: Bridging Theory and Dynamic Risk Mitigation,” *arXiv preprint arXiv:2410.13897*, 2024.
- [217] M. Cantero Gamito and C. T. Marsden, “Artificial intelligence co-regulation? The role of standards in the EU AI Act,” *International Journal of Law and Information Technology*, vol. 32, no. 1, Jun. 2024, doi: 10.1093/ijlit/eaee011.
- [218] A. Joshi, M. Chechelnitsky, and M. Chechelnitsky, “How the Government Can Mitigate the Emerging Risks of Artificial Intelligence,” *Journal of Student Research*, vol. 13, no. 1, Feb. 2024, doi: 10.47611/jsrhs.v13i1.6283.
- [219] L. Saraiva, “Moral Agency and Responsibility in AI Systems,” *International Journal of Philosophy*, vol. 3, no. 1, pp. 25–36, May 2024, doi: 10.47941/ijp.1867.
- [220] B. K. Konidena, J. N. A. Malaiyappan, and A. Tadimarri, “Ethical Considerations in the Development and Deployment of AI Systems,” *European Journal of Technology*, vol. 8, no. 2, pp. 41–53, Mar. 2024, doi: 10.47672/ejt.1890.
- [221] S. Holter and M. El-Assady, “Deconstructing Human-AI Collaboration: Agency, Interaction, and Adaptation,” *Computer Graphics Forum*, vol. 43, no. 3, Jun. 2024, doi: 10.1111/cgf.15107.
- [222] C. Toxtli, “Human-Centered Automation,” *arXiv preprint arXiv:2405.15960*, 2024.
- [223] A. J. Wright *et al.*, “The Artificial Intelligence in Public Health Toolkit: A novel resource for stakeholder engagement,” *Wellcome Open Res*, vol. 9, p. 420, Jul. 2024, doi: 10.12688/wellcomeopenres.22480.1.
- [224] K. Crockett, E. Colyer, L. Coulman, C. Nunn, and S. Linn, “PEAs in PODs: Co-Production of Community Based Public Engagement for Data and AI Research,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jun. 2024, pp. 1–10. doi: 10.1109/IJCNN60899.2024.10650039.
- [225] S. Shukla, “Principles Governing Ethical Development and Deployment of AI,” *International Journal of Engineering, Business and Management*, vol. 8, no. 2, pp. 26–46, 2024, doi: 10.22161/ijebm.8.2.5.

- [226] O. A. Lottu, B. S. Jacks, O. A. Ajala, and E. S. Okafo, “Towards a conceptual framework for ethical AI development in IT systems,” *World Journal of Advanced Research and Reviews*, vol. 21, no. 3, pp. 408–415, 2024.
- [227] M. M. Ferdaus, M. Abdelguerfi, E. Ioup, K. N. Niles, K. Pathak, and S. Sloan, “Towards trustworthy ai: A review of ethical and robust large language models,” *arXiv preprint arXiv:2407.13934*, 2024.
- [228] J. Shuford, “An Expedited Examination of Responsible AI Frameworks: Directing Ethical AI Development,” *Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023*, vol. 4, no. 1, pp. 241–251, May 2024, doi: 10.60087/jaigs.v4i1.138.
- [229] Temidayo Olorunsogo, Boma Sonimiteim Jacks, and Olakunle Abayomi Ajala, “LEVERAGING QUANTUM COMPUTING FOR INCLUSIVE AND RESPONSIBLE AI DEVELOPMENT: A CONCEPTUAL AND REVIEW FRAMEWORK,” *Computer Science & IT Research Journal*, vol. 5, no. 3, pp. 671–680, Mar. 2024, doi: 10.51594/csitrj.v5i3.927.
- [230] A. M. Tirado, P. Mulholland, and M. Fernandez, “Towards an Operational Responsible AI Framework for Learning Analytics in Higher Education,” *arXiv preprint arXiv:2410.05827*, 2024.
- [231] Z. Bin Akhtar and A. Tajbiul Rawol, “Enhancing Cybersecurity through AI-Powered Security Mechanisms,” *IT Journal Research and Development*, vol. 9, no. 1, pp. 50–67, Oct. 2024, doi: 10.25299/itjrd.2024.16852.
- [232] S. Ee, J. O’Brien, Z. Williams, A. El-Dakhakhni, M. Aird, and A. Lintz, “Adapting cybersecurity frameworks to manage frontier AI risks: A defense-in-depth approach,” *arXiv preprint arXiv:2408.07933*, 2024.
- [233] S. Tasneem and K. A. Islam, “Improve Adversarial Robustness of AI Models in Remote Sensing via Data-Augmentation and Explainable-AI Methods,” *Remote Sens (Basel)*, vol. 16, no. 17, p. 3210, Aug. 2024, doi: 10.3390/rs16173210.
- [234] T. Nguyen, H. Nguyen, A. Ijaz, S. Sheikhi, A. V Vasilakos, and P. Kostakos, “Large language models in 6G security: challenges and opportunities,” *arXiv preprint arXiv:2403.12239*, 2024.
- [235] Y. He, J. Qiu, W. Zhang, and Z. Yuan, “Fortifying Ethical Boundaries in AI: Advanced Strategies for Enhancing Security in Large Language Models,” *arXiv preprint arXiv:2402.01725*, 2024.
- [236] I. Khan, K. Praveena, Manjunatha, A. Dutt, D. S. Kiran, and T. K. Fadaam, “The Interpretability of Invariant Neural Networks in Deep Learning,” in *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/OTCON60325.2024.10687856.
- [237] L. Bereska and E. Gavves, “Mechanistic Interpretability for AI Safety--A Review,” *arXiv preprint arXiv:2404.14082*, 2024.
- [238] O. Kurasova, A. Budžys, and V. Medvedev, “Exploring Multidimensional Embeddings for Decision Support Using Advanced Visualization Techniques,” *Informatics*, vol. 11, no. 1, p. 11, Feb. 2024, doi: 10.3390/informatics11010011.
- [239] V. Reid, M. Weitz, M. Biancalana, J. A. Cole, and B. Rosen, “Use of an AI-powered visualization platform to support precision clinical decision-making in patients with breast cancer undergoing neoadjuvant chemotherapy,” *Journal of Clinical Oncology*, vol. 42, no. 16_suppl, pp. e12618–e12618, Jun. 2024, doi: 10.1200/JCO.2024.42.16_suppl.e12618.
- [240] G. R. Lokesh, K. S. Harish, V. S. Sangu, S. Prabakar, V. Santhosh Kumar, and M. Vallabhaneni, “AI and the Future of Work: Preparing the Workforce for Technological Shifts and Skill Evolution,” in *2024*

- International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/ICKECS61492.2024.10616486.
- [241] J. Westover, “Future-Proofing Your Organization: Key Principles for Leading Through the AI Transition,” *Human Capital Leadership Review*, vol. 13, no. 4, Oct. 2024, doi: 10.70175/hclreview.2020.13.4.3.
- [242] N. DUBEY, “The Implementation of Artificial Intelligence and its Future Potential,” *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 08, no. 04, pp. 1–5, Apr. 2024, doi: 10.55041/IJSREM31925.
- [243] M. Hamdi, “How AI is Transforming and Shaping the Future of Education,” in *2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES)*, IEEE, Jul. 2024, pp. 000115–000116. doi: 10.1109/INES63318.2024.10629089.
- [244] I. Celik, E. Gedrimiene, S. Siklander, and H. Muukkonen, “The affordances of artificial intelligence-based tools for supporting 21st-century skills:,” *Australasian Journal of Educational Technology*, Jun. 2024, doi: 10.14742/ajet.9069.
- [245] Y. Bekdemir, “The Urgency of AI Integration in Teacher Training: Shaping the Future of Education,” *Journal of Research in Didactical Sciences*, vol. 3, no. 1, p. 3, Dec. 2024, doi: 10.51853/jorids/15485.
- [246] J. Magrill and B. Magrill, “Preparing Educators and Students at Higher Education Institutions for an AI-Driven World,” *Teaching and Learning Inquiry*, vol. 12, Jun. 2024, doi: 10.20343/teachlearningqu.12.16.
- [247] J. Pool, M. Indulska, and S. Sadiq, “Large language models and generative AI in telehealth: a responsible use lens,” *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 2125–2136, Sep. 2024, doi: 10.1093/jamia/ocae035.
- [248] K. Kanont *et al.*, “Generative-AI, a Learning Assistant? Factors Influencing Higher-Ed Students’ Technology Acceptance,” *Electronic Journal of e-Learning*, vol. 22, no. 6, pp. 18–33, Jun. 2024, doi: 10.34190/ejel.22.6.3196.
- [249] G. Morante, C. Vilorio-Núñez, J. Florez-Hamburger, and H. Capdevilla-Molinares, “Proposal of an Ethical and Social Responsibility Framework for Sustainable Value Generation in AI,” in *2024 IEEE Technology and Engineering Management Society (TEMSCON LATAM)*, IEEE, Jul. 2024, pp. 1–6. doi: 10.1109/TEMSCONLATAM61834.2024.10717855.
- [250] P. David, H. Choung, and J. S. Seberger, “Who is responsible? US Public perceptions of AI governance through the lenses of trust and ethics,” *Public Understanding of Science*, vol. 33, no. 5, pp. 654–672, Jul. 2024, doi: 10.1177/09636625231224592.
- [251] A. Habibipour, “Responsible living labs: what can go wrong?,” *Journal of Information, Communication and Ethics in Society*, vol. 22, no. 2, pp. 205–218, Jun. 2024, doi: 10.1108/JICES-11-2023-0137.
- [252] Muhammad Ashraf Faheem, “Ethical AI: Addressing bias, fairness, and accountability in autonomous decision-making systems,” *World Journal of Advanced Research and Reviews*, vol. 23, no. 2, pp. 1703–1711, Aug. 2024, doi: 10.30574/wjarr.2024.23.2.2510.
- [253] M. Al-kfairy, D. Mustafa, N. Kshetri, M. Insiew, and O. Alfandi, “Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective,” *Informatics*, vol. 11, no. 3, p. 58, Aug. 2024, doi: 10.3390/informatics11030058.
- [254] A. Lasker, “EXPLORING ETHICAL CONSIDERATIONS IN GENERATIVE AI,” *Int J Adv Res (Indore)*, vol. 12, no. 04, pp. 531–535, Apr. 2024, doi: 10.21474/IJAR01/18578.

- [255] S. M. Williamson and V. Prybutok, “Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare,” *Applied Sciences*, vol. 14, no. 2, p. 675, Jan. 2024, doi: 10.3390/app14020675.
- [256] N. I. Qureshi, S. S. Choudhuri, Y. Nagamani, R. A. Varma, and R. Shah, “Ethical Considerations of AI in Financial Services: Privacy, Bias, and Algorithmic Transparency,” in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, IEEE, Apr. 2024, pp. 1–6. doi: 10.1109/ICKECS61492.2024.10616483.
- [257] Adekunle Oyeyemi Adeniyi, Jeremiah Olawumi Arowoogun, Chioma Anthonia Okolo, Rawlings Chidi, and Oloruntoba Babawarun, “Ethical considerations in healthcare IT: A review of data privacy and patient consent issues,” *World Journal of Advanced Research and Reviews*, vol. 21, no. 2, pp. 1660–1668, Feb. 2024, doi: 10.30574/wjarr.2024.21.2.0593.
- [258] G. Singh and K. K. Bali, “Enhancing Decision-Making in Optimization through LLM-Assisted Inference: A Neural Networks Perspective,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jun. 2024, pp. 1–7. doi: 10.1109/IJCNN60899.2024.10649965.
- [259] A. Kalyanpur *et al.*, “Multi-step Inference over Unstructured Data,” *arXiv preprint arXiv:2406.17987*, 2024.
- [260] J. R., “Transparency in AI Decision Making: A Survey of Explainable AI Methods and Applications,” *Advances in Robotic Technology*, vol. 2, no. 1, pp. 1–10, Jan. 2024, doi: 10.23880/art-16000110.
- [261] J. Singh, S. Rani, and G. Srilakshmi, “Towards Explainable AI: Interpretable Models for Complex Decision-making,” in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, IEEE, Apr. 2024, pp. 1–5. doi: 10.1109/ICKECS61492.2024.10616500.
- [262] Lakshmi Harika Palivela, “Revolutionizing AI Decision-Making with Hybrid Rule-Based Systems: A Novel Framework for Transparent and Accurate Outcomes,” *Communications on Applied Nonlinear Analysis*, vol. 31, no. 8s, pp. 115–127, Sep. 2024, doi: 10.52783/cana.v31.1460.
- [263] S. Labkoff *et al.*, “Toward a responsible future: recommendations for AI-enabled clinical decision support,” *Journal of the American Medical Informatics Association*, vol. 31, no. 11, pp. 2730–2739, Nov. 2024, doi: 10.1093/jamia/ocae209.
- [264] C. Novelli, F. Casolari, P. Hacker, G. Spedicato, and L. Floridi, “Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity,” *SSRN Electronic Journal*, 2024, doi: 10.2139/ssrn.4694565.
- [265] G. Noto La Diega and L. C. T. Bezerra, “Can there be responsible AI without AI liability? Incentivizing generative AI safety through ex-post tort liability under the EU AI liability directive,” *International Journal of Law and Information Technology*, vol. 32, no. 1, Jun. 2024, doi: 10.1093/ijlit/eaee021.
- [266] P. K. Kopalle, M. Gangwar, and A. Uppal, “Commentary on ‘AI is Changing the World: For Better or for Worse?,’” *Journal of Macromarketing*, vol. 44, no. 4, pp. 886–891, Dec. 2024, doi: 10.1177/02761467241290813.
- [267] Z. Shouran and D. A. Ali, “The Implementation of Artificial Intelligence in Human Resources Management,” *Journal of International Conference Proceedings*, vol. 7, no. 1, pp. 244–258, May 2024, doi: 10.32535/jicp.v7i1.2993.
- [268] S. A. Joseph, T. M. Kolade, O. Obioha-Val, O. O. Adebisi, O. S. Ogungbemi, and O. O. Olaniyi, “AI-Powered Information Governance: Balancing Automation and Human Oversight for Optimal Organization

- Productivity,” *Asian Journal of Research in Computer Science*, vol. 17, no. 10, pp. 110–131, Oct. 2024, doi: 10.9734/ajrcos/2024/v17i10513.
- [269] L. Yorks and M. Y. Jester, “Applying generative AI ethically in HRD practice,” *Human Resource Development International*, vol. 27, no. 3, pp. 410–427, May 2024, doi: 10.1080/13678868.2024.2337963.
- [270] A. Swindell, L. Greeley, A. Farag, and B. Verdone, “Against Artificial Education: Towards an Ethical Framework for Generative Artificial Intelligence (AI) Use in Education,” *Online Learning*, vol. 28, no. 2, Jun. 2024, doi: 10.24059/olj.v28i2.4438.
- [271] R. Raman, S. Mandal, P. Das, T. Kaur, J. P. Sanjanasri, and P. Nedungadi, “Exploring University Students’ Adoption of ChatGPT Using the Diffusion of Innovation Theory and Sentiment Analysis With Gender Dimension,” *Hum Behav Emerg Technol*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/3085910.
- [272] Y. Kwak and Z. A. Pardos, “Bridging large language model disparities: Skill tagging of multilingual educational content,” *British Journal of Educational Technology*, vol. 55, no. 5, pp. 2039–2057, Sep. 2024, doi: 10.1111/bjet.13465.
- [273] O. Viberg, R. F. Kizilcec, A. F. Wise, I. Jivet, and N. Nixon, “Advancing equity and inclusion in educational practices with <scp>AI</scp>-powered educational decision support systems (<scp>AI</scp> - <scp>EDSS</scp>),” *British Journal of Educational Technology*, vol. 55, no. 5, pp. 1974–1981, Sep. 2024, doi: 10.1111/bjet.13507.
- [274] M. Mangal and Z. A. Pardos, “Implementing equitable and intersectionality-aware <scp>ML</scp> in education: A practical guide,” *British Journal of Educational Technology*, vol. 55, no. 5, pp. 2003–2038, Sep. 2024, doi: 10.1111/bjet.13484.
- [275] F. Jourdan, “Advancing Fairness in Natural Language Processing: From Traditional Methods to Explainability,” *arXiv preprint arXiv:2410.12511*, 2024.
- [276] B. Yan *et al.*, “On protecting the data privacy of large language models (llms): A survey,” *arXiv preprint arXiv:2403.05156*, 2024.
- [277] R. E. -, “Revolutionizing Knowledge Synthesis: Introducing AI-Enhanced Collaborative Theory (AIECT),” *International Journal For Multidisciplinary Research*, vol. 6, no. 5, Oct. 2024, doi: 10.36948/ijfmr.2024.v06i05.28203.
- [278] M. Chen *et al.*, “Combating Security and Privacy Issues in the Era of Large Language Models,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 8–18. doi: 10.18653/v1/2024.naacl-tutorials.2.
- [279] J. Cui *et al.*, “ID-SR: Privacy-Preserving Social Recommendation Based on Infinite Divisibility for Trustworthy AI,” *ACM Trans Knowl Discov Data*, vol. 18, no. 7, pp. 1–25, Aug. 2024, doi: 10.1145/3639412.
- [280] M. P. -, A. K. U. -, M. R. -, and J. D. -, “A comparative Study on AI-Driven Anonymization Techniques for Protecting Personal Data,” *International Journal For Multidisciplinary Research*, vol. 6, no. 4, Jul. 2024, doi: 10.36948/ijfmr.2024.v06i04.24770.
- [281] A. H. Farea, O. H. Alhazmi, R. Samet, and M. S. Guzel, “AI-Powered Integrated With Encoding Mechanism Enhancing Privacy, Security, and Performance for IoT Ecosystem,” *IEEE Access*, vol. 12, pp. 121368–121386, 2024, doi: 10.1109/ACCESS.2024.3449630.

- [282] L. Yang, M. Tian, D. Xin, Q. Cheng, and J. Zheng, “AI-Driven Anonymization: Protecting Personal Data Privacy While Leveraging Machine Learning,” *arXiv preprint arXiv:2402.17191*, 2024.
- [283] Y. Wang, C. Liu, K. Zhou, T. Zhu, and X. Han, “Towards regulatory generative AI in ophthalmology healthcare: a security and privacy perspective,” *British Journal of Ophthalmology*, vol. 108, no. 10, pp. 1349–1353, Oct. 2024, doi: 10.1136/bjo-2024-325167.
- [284] Y. Chen and P. Esmailzadeh, “Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges,” *J Med Internet Res*, vol. 26, p. e53008, Mar. 2024, doi: 10.2196/53008.
- [285] A. K. Sharma and R. Sharma, “Generative Artificial Intelligence and Legal Frameworks: Identifying Challenges and Proposing Regulatory Reforms,” *Kutafin Law Review*, vol. 11, no. 3, pp. 415–451, Oct. 2024, doi: 10.17803/2713-0533.2024.3.29.415-451.
- [286] A. M. Horzyk, “Implications of European By-Design Laws on Data-Driven System Design,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jun. 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10649941.
- [287] C. N. Pehlivan, “The EU Artificial Intelligence (AI) Act: An Introduction,” *Global Privacy Law Review*, vol. 5, no. Issue 1, pp. 31–42, Mar. 2024, doi: 10.54648/GPLR2024004.
- [288] M. Nisevic, A. Cuyper, and J. De Bruyne, “Explainable AI: Can the AI Act and the GDPR go out for a date?,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Jun. 2024, pp. 1–8. doi: 10.1109/IJCNN60899.2024.10649994.
- [289] D. Korobenko, A. Nikiforova, and R. Sharma, “Towards a Privacy and Security-Aware Framework for Ethical AI: Guiding the Development and Assessment of AI Systems,” in *Proceedings of the 25th Annual International Conference on Digital Government Research*, New York, NY, USA: ACM, Jun. 2024, pp. 740–753. doi: 10.1145/3657054.3657141.
- [290] Y. He, J. Qiu, W. Zhang, and Z. Yuan, “Fortifying Ethical Boundaries in AI: Advanced Strategies for Enhancing Security in Large Language Models,” *arXiv preprint arXiv:2402.01725*, 2024.
- [291] M. Miranda, E. S. Ruzzetti, A. Santilli, F. M. Zanzotto, S. Bratières, and E. Rodolà, “Preserving privacy in large language models: A survey on current threats and solutions,” *arXiv preprint arXiv:2408.05212*, 2024.
- [292] B.-M. Ganesu and J. Passerat-Palmbach, “Trust the process: Zero-knowledge machine learning to enhance trust in generative ai interactions,” *arXiv preprint arXiv:2402.06414*, 2024.
- [293] A. Majeed, S. Khan, and S. O. Hwang, “Towards Optimization of Privacy-Utility Trade-Off Using Similarity and Diversity Based Clustering,” *IEEE Trans Emerg Top Comput*, vol. 12, no. 1, pp. 368–385, Jan. 2024, doi: 10.1109/TETC.2023.3258528.
- [294] S. Yechuri, “Enhanced Utility-Driven Data Anonymization: Leveraging AI and Machine Learning for Sensitive Data Privacy,” *Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023*, vol. 1, no. 1, pp. 229–232, Jan. 2024, doi: 10.60087/jaigs.v1i1.229.
- [295] A. Majeed and S. O. Hwang, “Differential Privacy and k -Anonymity-Based Privacy Preserving Data Publishing Scheme With Minimal Loss of Statistical Information,” *IEEE Trans Comput Soc Syst*, vol. 11, no. 3, pp. 3753–3765, Jun. 2024, doi: 10.1109/TCSS.2023.3320141.
- [296] M. Y. Nazeer, “Algorithmic Conscience: An In-Depth Inquiry into Ethical Dilemmas in Artificial Intelligence,” *International Journal of Research and Innovation in Social Science*, vol. VIII, no. V, pp. 725–732, 2024, doi: 10.47772/IJRISS.2024.805052.

- [297] K. Yekaterina, “Challenges and Opportunities for AI in Healthcare,” *International Journal of Law and Policy*, vol. 2, no. 7, pp. 11–15, Jul. 2024, doi: 10.59022/ijlp.203.
- [298] S. Mirishli, “Ethical Implications of AI in Data Collection: Balancing Innovation with Privacy,” *ANCIENT LAND*, vol. 6, no. 8, pp. 40–55, Aug. 2024, doi: 10.36719/2706-6185/38/40-55.
- [299] S. U. Okon, O. O. Olateju, O. S. Ogungbemi, S. A. Joseph, A. O. Olisa, and O. O. Olaniyi, “Incorporating Privacy by Design Principles in the Modification of AI Systems in Preventing Breaches across Multiple Environments, Including Public Cloud, Private Cloud, and On-prem,” *Journal of Engineering Research and Reports*, vol. 26, no. 9, pp. 136–158, Sep. 2024, doi: 10.9734/jerr/2024/v26i91269.
- [300] N. Joshi, “Emerging Challenges in Privacy Protection with Advancements in Artificial Intelligence,” *International Journal of Law and Policy*, vol. 2, no. 4, pp. 55–77, Apr. 2024, doi: 10.59022/ijlp.171.
- [301] S. Adhyapak, “Data Privacy and Security Risks in AI-Based Code Understanding,” *Int J Res Appl Sci Eng Technol*, vol. 12, no. 6, pp. 1913–1921, Jun. 2024, doi: 10.22214/ijraset.2024.63423.
- [302] A. Bahr, “Bridges Without Foundation? Why the Use of AI Tools in Academia Needs to Build on Ethics First,” *Filozofia*, vol. 79, no. 5, pp. 486–500, May 2024, doi: 10.31577/filozofia.2024.79.5.2.
- [303] A. Pack and J. Maloney, “Using Artificial Intelligence in TESOL: Some Ethical and Pedagogical Considerations,” *TESOL Quarterly*, vol. 58, no. 2, pp. 1007–1018, Jun. 2024, doi: 10.1002/tesq.3320.
- [304] E. Creely, “Exploring the Role of Generative AI in Enhancing Language Learning: Opportunities and Challenges,” *International Journal of Changes in Education*, 2024.
- [305] Y.-C. Chen and H.-T. Hou, “A Mobile Contextualized Educational Game Framework With ChatGPT Interactive Scaffolding for Employee Ethics Training,” *Journal of Educational Computing Research*, vol. 62, no. 7, pp. 1737–1762, Dec. 2024, doi: 10.1177/07356331241268505.
- [306] L. Peng and B. Zhao, “Navigating the ethical landscape behind ChatGPT,” *Big Data Soc*, vol. 11, no. 1, Mar. 2024, doi: 10.1177/20539517241237488.
- [307] U. A. Bukar, Md. S. Sayeed, S. Fatimah Abdul Razak, S. Yogarayan, and R. Sneesl, “Decision-Making Framework for the Utilization of Generative Artificial Intelligence in Education: A Case Study of ChatGPT,” *IEEE Access*, vol. 12, pp. 95368–95389, 2024, doi: 10.1109/ACCESS.2024.3425172.
- [308] X. (Nancy) Deng and K. D. Joshi, “Promoting Ethical Use of Generative AI in Education,” *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 55, no. 3, pp. 6–11, Aug. 2024, doi: 10.1145/3685235.3685237.
- [309] S. Smith *et al.*, “A University Framework for the Responsible use of Generative AI in Research,” *arXiv preprint arXiv:2404.19244*, 2024.
- [310] B. C. Cheong, “Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making,” *Frontiers in Human Dynamics*, vol. 6, Jul. 2024, doi: 10.3389/fhumd.2024.1421273.
- [311] Edith Ebele Agu, Angela Omozele Abhulimen, Anwuli Nkemchor Obiki-Osafiele, Olajide Soji Osundare, Ibrahim Adedeji Adeniran, and Christianah Pelumi Efunniyi, “Discussing ethical considerations and solutions for ensuring fairness in AI-driven financial services,” *International Journal of Frontline Research in Multidisciplinary Studies*, vol. 3, no. 2, pp. 001–009, Aug. 2024, doi: 10.56355/ijfrms.2024.3.2.0024.

- [312] A. Sattlegger and N. Bharosa, “Beyond principles: Embedding ethical AI risks in public sector risk management practice,” in *Proceedings of the 25th Annual International Conference on Digital Government Research*, New York, NY, USA: ACM, Jun. 2024, pp. 70–80. doi: 10.1145/3657054.3657063.
- [313] Olatunji Akinrinola, Chinwe Chinazo Okoye, Onyeka Chrisanctus Ofodile, and Chinonye Esther Ugochukwu, “Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability,” *GSC Advanced Research and Reviews*, vol. 18, no. 3, pp. 050–058, Mar. 2024, doi: 10.30574/gscarr.2024.18.3.0088.
- [314] A. Jedličková, “Ethical considerations in Risk management of autonomous and intelligent systems,” *Ethics & Bioethics*, vol. 14, no. 1–2, pp. 80–95, Jun. 2024, doi: 10.2478/ebce-2024-0007.
- [315] E. G. Hernández, “Hacia una implementación ética e inclusiva de la Inteligencia Artificial en las organizaciones: un marco multidimensional,” *arXiv preprint arXiv:2405.00225*, 2024.
- [316] B. Lainjo, “The Role of Artificial Intelligence in Achieving the United Nations Sustainable Development Goals,” *J Sustain Dev*, vol. 17, no. 5, p. 30, Aug. 2024, doi: 10.5539/jsd.v17n5p30.
- [317] A. B. Cinar and S. Bilodeau, “Incorporating AI into the Inner Circle of Emotional Intelligence for Sustainability,” *Sustainability*, vol. 16, no. 15, p. 6648, Aug. 2024, doi: 10.3390/su16156648.
- [318] L. Alkire, A. Bilgihan, M. (Myla) Bui, A. J. Buoye, S. Dogan, and S. Kim, “RAISE: leveraging responsible AI for service excellence,” *Journal of Service Management*, vol. 35, no. 4, pp. 490–511, Jul. 2024, doi: 10.1108/JOSM-11-2023-0448.